ICGC PAN-CANCER Abstracts

Compilation as of March 20 2014

Table of Contents

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mutational landscape of *CCNE1* amplified tumors

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

David Bowtell (Peter MacCallum) Cancer Centre
Sean Grimmond (University of Glasgow)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s)& institute(s)of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Dariush Etemadmoghadam (Peter MacCallum Cancer Center) Ann-Marie Patch (Queensland Centre for Medical Genomics)	

Background and preliminary data

Focal amplification of *CCNE1* (Cyclin E1) occurs at a frequency of about 20-25% of high-grade serous ovarian cancer (HGSC) and is also common to other cancer types including bladder, gastric, uterine, lung and breast cancers (Fig 1).

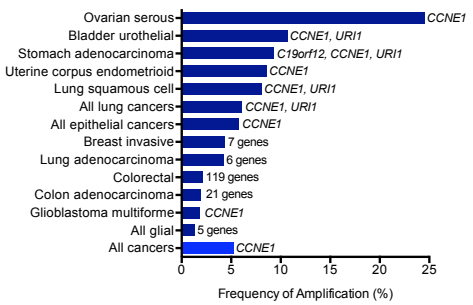


Fig 1. Pan-cancer copy number analysis of 6,547 tumor samples comprising 22 cancer types from TCGA (Etemadmoghadam et al., 2013 *PNAS).*

We have previously shown that Cyclin E1 (*CCNE1*) gene amplification is a key prognostic marker of primary treatment failure (Etemadmoghadam et al., 2009 Clin Can Res) and associated with oncogene addiction in HGSC (Etemadmoghadam et al, 2010 PLoS ONE). Activity of *CCNE1* may be targeted therapeutically via inactivation of its partner kinase CDK2. Whole-genome doubling and polyploidy is a characteristic of *CCNE1* amplified tumors, and is associated with resistance to CDK2 inhibitors *in vitro* (Etemadmoghadam et al., 2013 Clin Can Res). Using a genome-wide shRNA screen, we have also recently demonstrated dependence on HR genes including *BRCA1* and ubiquitin pathway members in *CCNE1* amplified tumors of various tumor types, which may be exploited with the use of proteasome inhibitors (Etemadmoghadam et al., 2013 PNAS). Collectively, our work demonstrates *CCNE1* amplification to be a key prognostic marker and molecular target in HGSC, and these findings may translate to other tumor types.

Investigation of the mutational landscape of *CCNE1* amplified tumors, including co-operating and mutually exclusive genomic changes, may help better understand their biology. Data from TCGA has shown that approximately 30% of HGSC tumors have alterations in the Rb pathway, including amplification of *CCNE1* (~20%), loss of *RB1* (~10%), or gain of *RBBP8* (~4%) (Ciriello et al., 2012 Genome Res). Furthermore, activation of the RB1/CCNE1 pathway is largely exclusive of *BRCA1/2* inactivation (TCGA 2011 Nature and Etemadmoghadam et al., 2013 PNAS). We have also observed that co-amplification of *TPX2*, encoding a microtubule-associated protein, commonly occurs in HGSC and may function with *CCNE1* (Etemadmoghadam et al., 2010 PLoS ONE). Inactivating breakpoints in *TSHZ3*, a homeobox transcription factor gene located telomeric of *CCNE1*, could also play a role in ovarian cancer or simply be collateral to *CCNE1* amplification (McBride, Etemadmoghadam et al., 2012 J Path). Cyclin E1 leads to S-phase entry in the presence of insufficient nucleotide pools, leading to replication-induced damage (Bester et al., 2011 Cancer Cell). Our preliminary analysis of ovarian WGS data generated as part of the ICGC suggests there are specific mutational signatures associated with *CCNE1* status.

We aim to determine if tumors bearing *CCNE1* gene amplification share these and other molecular characteristics across tumor types.

Timelines & resources dedicated to project
The initiation of this exploratory project requires the core variant-calling vcfs and mutational signature analysis data. It is anticipated that the proposed analysis would be complete in 3 months Key intermediate milestones will include stratification of cases into *CCNE1* amplified and non-amplified sub groups and classification of mutational pathogenicity for variants in genes of interest. This project builds on the collaborative strengths of the Australian ICGC ovarian project spanning two established teams with the necessary computer resources as well as the two junior investigators with specialized knowledge and time dedicated for the analysis.
Research proposal
Questions: We propose an exploratory pan-cancer analysis of *CCNE1* amplified tumors to address the following questions: 1) Do the frequency and distribution of structural variation and mutation in *CCNE1* amplified tumors differ to unamplified tumors? 2) Does the previously reported mutual exclusivity of *CCNE1* amplification against HR pathway mutations such as *BRCA1, BRCA2, RB1* and other pathway member mutations, including essential genes identified in our synthetic lethal screen (Etemadmoghadam et al., 2013 PNAS), hold across multiple cancers? 3) Is it possible to identify a mutational footprint or signature associated with *CCNE1* amplification? 4) Aim to better define the structure of the *CCNE1* (19q12) amplicon, identifying genes commonly co-amplified and the effect on neighboring genes (e.g. inactivation *TSHZ3*) **Proposed Analyses:** Levels of *CCNE1* amplification will be determined from copy number data with thresholds to distinguish tumors with amplification from those with intermediate gain or not amplified informed by our previous findings in ICGC current samples. The observed pattern of mutual exclusivity of *CCNE1* amplification and *BRCA1/2* pathway mutations will be tested across the spectrum of cancers. Further investigations of the relationships between genes identified in synthetic lethal screen as essential in *CCNE1* amplified cases may reveal patterns of commonly essential genes. Following this, the ability to associate amplification of *CCNE1* to the mutational signatures analysis could lend independent support of the analyses above. These findings, combined with detailed structural and gene level descriptions of the chr19q12 amplicon that will be generated will provide a key resource for a wide range of cancer researchers and establish the clinical importance of this amplification across all tumor types.
Legacy plans
All software solutions developed for this investigation will be made publically available in an appropriate format.

DAVID BOWTELL

Tel: +61 (03 96561356
E mail: d.bowtell@petermac.org

EDUCATION

1980	Bachelor of Animal Science University of Melbourne
1977 – 81	Bachelor of Veterinary Science with Honors
1982 – 85	PhD University of Melbourne

PROFESSIONAL EMPLOYMENT

1997 –	Group Leader and Head of Cancer Genomics Program, Peter MacCallum Cancer Institute
2002 –	Professorial Fellow Department of Biochemistry & Molecular Biology, University of Melbourne
2000-2009	Director, Research, Peter MacCallum Cancer Centre
1997-1999	Scientific Director, Research, Peter MacCallum Cancer Centre
1993-1997	Howard Hughes Institute International Research Scholar
1987-1991	CJ Martin Research Fellow, Howard Florey Institute, Melbourne and UC Berkeley CA

SUMMARY

Professor Bowtell is the Head of the Cancer Genomics and Genetics Program at Peter MacCallum Cancer Centre and principal investigator for the Australian Ovarian Cancer Study (AOCS). He was Director of Research at Peter Mac for the last decade, returning to fulltime research in 2010. Professor Bowtell is one of Australia's leading ovarian cancer and human molecular genetics researchers. In addition to his appointment at Peter Mac, Professor Bowtell is holds a joint position at Imperial College London, is a Visiting Professor at the Dana Farber Cancer Institute (Boston), and Professor of Cancer Genomics at the University of Melbourne. Professor Bowtell has published over 200 papers (h-score=55, >13,000 citations). He has published high profile articles in *Nature, Oncogene, Nature Structural Biology, EMBO J, PNAS, Cancer Cell, Nature Cell Biology, Nature Reviews Cancer, Cancer Research, Clinical Cancer Research, PLoSONE, New England J Medicine, J Pathology, and J Clin Oncology.*

PUBLICATIONS

Ahmed, A. A., D. Etemadmoghadam, J. Temple, A. G. Lynch, M. Riad, R. Sharma, C. Stewart, S. Fereday, C. Caldas, A. Defazio, D. Bowtell* and J. D. Brenton (2010). "Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary." J Pathol **221**(1): 49-56. [114] *Co-senior author.

Etemadmoghadam, D., A. deFazio, R. Beroukhim, C. Mermel, J. George, G. Getz, R. Tothill, A. Okamoto, M. B. Raeder, P. Harnett, S. Lade, L. A. Akslen, A. V. Tinker, B. Locandro, K. Alsop, Y. E. Chiew, N. Traficante, S. Fereday, D. Johnson, S. Fox, W. Sellers, M. Urashima, H. B. Salvesen, M. Meyerson and D. Bowtell (2009). "Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas." Clin Cancer Res **15**(4): 1417-1427. [65]

TCGA (2011). "Integrated genomic analyses of ovarian carcinoma." Nature **474**(7353): 609-615. [348]

Alsop, K., S. Fereday, C. Meldrum, A. deFazio, C. Emmanuel, J. George, A. Dobrovic, M. J. Birrer, P. M. Webb, C. Stewart, M. Friedlander, S. Fox, D. Bowtell* and G. Mitchell (2012). "BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group." J Clin Oncol **30**(21): 2654-2663. [24] *Corresponding author and co-senior author.

Etemadmoghadam, D., B. A. Weir, G. Au-Yeung, K. Alsop, G. Mitchell, J. George, G. Australian Ovarian Cancer Study, S. Davis, A. D. D'Andrea, K. Simpson, W. C. Hahn and D. D. Bowtell (2013). "Synthetic lethality between CCNE1 amplification and loss of BRCA1." Proc Natl Acad Sci U S A **110**(48): 19489-19494. Featured N&V in *Cancer Discovery.*

International Cancer Genome Consortium

Sean Michael GRIMMOND

Current Poitions: Chair of Medical Genomics,
Director, Howat Canceer Genomics Facility
Wolfson Wohl Cancer Research Centre,
University of Glasgow, Garscube Estate,
Switchback Road, Bearsden,
Glasgow Scotland G61 1BD
United Kingdom.

Education and Degrees:

2011 Founding Fellow (Faculty of Science), Royal College of Pathologists of Australasia.
1994 PhD (University of Queensland
1987 B.Sc. Hons University of New England

Awards & Scholarships

2013 Royal Society Wolfson Merit Award
2012 NH&MRC Principal Research Fellowship
2011 Julian Wells Medal for Transcriptomics
2007 NH&MRC Senior Research Fellowship (SRFA)
2004 Eppendorf Australian Genomics Research Medal
2002 NH&MRC Career Development Fellowship
1997 NH&MRC CJ Martin Travelling Fellow

Selected publications

Andrew V. Biankin, Nicola Waddell, Karin S. Kassahn, Marie-Claude Gingras, Amber L. Johns, David K. Miller, Peter J. Wilson, Ann-Marie Patch, Jianmin Wu, David K. Chang[3], Mark J. Cowley, Brooke B. Gardiner, Sarah Song, Ivon Harliwong, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Shivangi Wani, Milena Gongora, Marina Pajic, Christopher J. Scarlett, Anthony J. Gill, Elizabeth A. Musgrove, Robert L. Sutherland, Andreia V. Pinho, Ilse Rooman, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor Scott Wood, Christina Xu, Katia Nones, J. Lynn Fink, Angelika Christ, Tim Bruxner, Nicole Cloonan, Gabriel Kolle, Felicity Newell, Mark Pinese, Scott Mead, Jeremy L. Humphris, Warren Kaplan, Marc D. Jones, Emily K. Colvin, Adnan M. Nagrial, Emily S. Humphrey Angela Chou, Venessa T. Chin, Lorraine A. Chantrill. Jaswinder S. Samra, James G. Kench, Jessica A. Lovell, Roger J. Daly, Neil D. Merrett, Christopher Toon, Krishna Epari, Nam Q. Nguyen, Andrew Barbour, Nikolajs Zeps, Australian Pancreatic Cancer Genome Initiative, Nipun Kakkar, Fengmei Zhao, Yuan Qing Wu, Min Wang, Donna M. Muzny, William E. Fisher, F. Charles Brunicardi, Sally E. Hodges, Jennifer Drummond, Kyle Chang, Yi Han, Lora L. Lewis, Huyen Dinh, Christian J. Buhay, Lakshmi Muthuswamy, Timothy Beck, Lee Timms, Michelle Sam, Kimberly Begley, Andrew Brown, Deepa Pai, Ami Panchal, Nicholas Buchner, Richard De Borja, Robert E. Denroche, Christina K. Yung, Stefano Serra, Nicole Onetto, Debabrata Mukhopadhyay, Ming-Sound Tsao, Patricia A Shaw, Gloria Petersen, Steven Gallinger, Lincoln D. Stein, Ralph H. Hruban, Anirban Maitra, Christine A. Iacobuzio-Donahue Richard D. Schulick, Christopher L. Wolfgang, Richard A. Morgan, Rita T. Lawlor, Stefania Beghell, Vincenzo Corbo, Maria Scardoni, Claudio Bassi, Margaret A. Tempero, Karen M. Mann, Nancy A. Jenkins, Pedro A. Perez-Mancera, David J. Adams, David A. Largaespada, Lodewyk F. Wessels, Alistair G. Rust, David A. Tuveson, Neal G. Copeland, Thomas J. Hudson, Aldo Scarpa, James R. Eshleman, David A. Wheeler, John V. Pearson, John D. McPherson, Richard A. Gibbs and Sean M. Grimmond (2012) Genomic Analysis Reveals Roles for Chromatin Modification and Axon Guidance in Pancreatic Cancer. ***Nature, epub 24th Oct, 2012.***

Pedro A. Pérez-Mancera, Alistair G. Rust, Louise van der Weyden, Glen Kristiansen, Allen Li, Aaron L. Sarver, Kevin A. T. Silverstein, Robert Grützmann, Daniela Aust, Petra Rümmele, Thomas Knösel, Colin Herd, Derek L. Stemple, Ross Kettleborough, Jacqueline A. Brosnan, Ang Li, Richard Morgan, Spencer Knight, Jun Yu, Shane Stegeman, Lara S. Collier, Jelle J. ten Hoeve, Jeroen de Ridder, Alison P. Klein, Michael Goggins, Ralph H. Hruban, David K. Chang, Andrew V. Biankin, **Sean M. Grimmond**, Lodewyk F. A. Wessels, Stephen A. Wood, Christine A. Iacobuzio-Donahue, Christian Pilarsky, David A. Largaespada, David J. Adams, David A. Tuveson (2012) The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. ***Nature****, 486, 266–270*.

Karen M. Mann, Jerrold M. Ward, Christopher Chin Kuan Yew, Anne Kovochich, David W. Dawson, Michael A. Black, Benjamin T. Brett, Todd E. Sheets, Adam J Dupuy, David K. Chang, Andrew V. Biankin, Nic Waddell, Karin S. Kassahn, **Sean M Grimmond**, Alistair G. Rust, David J. Adams, Nancy A. Jenkins, and Neal G. Copeland (2012) Sleeping Beauty Mutagenesis Reveals Cooperating Mutations and Pathways in Pancreatic Adenocarcinoma ***Proc Natl Acad Science USA*** 2012 Mar 15. 109(16):5934-41.

DARIUSH ETEMADMOGHADAM, PHD
dariush.etemadmoghadam@petermac.org

Summary

Dr Etemadmoghadam is a Senior Research Fellow at the Peter MacCallum Cancer Institute, after first joining the Cancer Genomics and Genetics Laboratory as a PhD student in 2004. He has led the work identifying Cyclin E1 gene amplification as a key prognostic marker of primary treatment resistance and a therapeutic target in high-grade serous ovarian cancer (Etemadmoghadam et al, 2009 *Clin Can Res*; 2010 *PLoS ONE*; 2012 *J Path;* 2013 *Clin Can Res* and 2013 *PNAS*). Dariush is currently involved in the International Cancer Genome Consortium (ICGC), a large-scale sequencing study of tumours from over 50 different cancer types (ICGC, 2010 *Nature*).

Education

2004 – 2008 **PhD:** Peter MacCallum Cancer Centre, *East Melbourne* and the Department of Biochemistry and Molecular Biology, University of Melbourne, *Parkville*

1999-2002 **Bachelor of Biomedical Science (First-class honours):** Faculty of Medicine, Dentistry and Health Sciences and Faculty of Science, University of Melbourne, *Parkville.* Major: Functional, Computational and Applied Genomics

Professional Employment

2013 – **Senior Research Fellow:** Cancer Genomics and Genetics Laboratory, Peter MacCallum Cancer Centre, *East Melbourne* (PI: Prof David Bowtell)

2008 – 2013 **Post-Doctoral Researcher:** Cancer Genomics and Genetics Laboratory, Peter MacCallum Cancer Centre, *East Melbourne* (PI: Prof David Bowtell)

2003 – 2004 **Research Assistant:** Renal Laboratory, Austin Research Institute, *Heidelberg* (PI: A. Professor John Kanellis)

Grant Funding

2013 – 2015 **Chief Investigator (CIA, New Investigator):** The National Health and Medical Research Council (NHMRC) Project Grant

Publications

24 publications in peer-reviewed journals including 5 first-author publications

Selected Publications

1. **Etemadmoghadam**, D., et al., *Synthetic lethality between CCNE1 amplification and loss of BRCA1.*, in *Proc Natl Acad Sci USA*. 2013. p. 19489-19494.
   Received commentary: *Cyclin E1 amplification confers sensitivity to proteasome inhibition.,* in *Cancer Discovery* 2014 Jan;4(1):OF15
2. **Etemadmoghadam, D**., et al., *Resistance to CDK2 inhibitors is associated with selection of polyploid cells in CCNE1 amplified ovarian cancer.*, in *Clin Cancer Res*. 2013. p. 5960-71
3. Karst, A.M., et al., *Cyclin E1 deregulation occurs early in secretory cell transformation to promote formation of fallopian tube derived high-grade serous ovarian cancers.*, in Cancer Research. 2013.
4. Mcbride*, D.J., D., et al., *Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes.*, in *J. Pathol.* 2012. p. 446-455 (*co-first author).
5. **Etemadmoghadam**, et al., *Amplicon-dependent CCNE1 expression is critical for clonogenic survival after cisplatin treatment and is correlated with 20q11 gain in ovarian cancer.*, in *PLoS ONE*. 2010. p. e15498.
6. ICGC, *International network of cancer genome projects.*, in *Nature*. 2010. p. 993-998.
7. **Etemadmoghadam**, D., et al. *Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas.*, in *Clin Cancer Res*. 2009. p. 1417-1427.
8. Tothill, R.W., et al., *Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome.*, in *Clin Cancer Res*. 2008. p. 5198-5208.

Ann-Marie Patch

Tel: +61 (0)7 3346 2088
E mail: a.patch@uq.edu.au

Education

Oct 2001 - Oct 2005	University of Exeter	**PhD Biological Sciences** (Awarded June 2006) A comparative analysis of tandem repeats in the fission yeast and budding yeast genomes.
Oct 1998 - June 2001	University of Exeter	**BSc Hons Biological Sciences - First Class**

Professional Employment

April 2011 - Present **Senior Bioinformatics Officer** - Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland

- Leading analysis for the Australian arm of the International Cancer Genome Consortium Ovarian Cancer project in collaboration with David Bowtell and team at Peter MacCallum Cancer Centre, Melbourne.

Feb 2009 - April 2011 **Bioinformatician** - Molecular Genetics Laboratory - Royal Devon and Exeter NHS Foundation Trust Hospital

- Management of implementation of new laboratory management system (StarLIMS) Supporting diagnostic team with evaluation variants of unknown pathogenicity, evaluation of software solutions and providing training sessions for staff of all levels.

Honorary Research Fellow - Peninsula College of Medicine & Dentistry (as below)

Jan 2007 - Jan 2009 **Associate Research Fellow** - Peninsula College of Medicine & Dentistry

- Developing pipeline for next generation sequencing and homozygosity mapping analysis of neonatal diabetes cases from consanguineous unions for the identification of novel diabetes genes. Also Informatics support of molecular genetics laboratory process redesign.

Jan 2006 - Dec 2006 **Research Genetic Technologist** - Molecular Genetics Laboratory - Royal Devon and Exeter NHS Foundation Trust Hospital

Technical skills

- Proven research and project management skills.
- Experience evaluating and analysing Next Generation massively parallel sequencing results from Illumina HiSeq for human whole genome and exome capture experiments.
- Programming proficiency in Perl
- Solid background in molecular genetics techniques and theory

Publications

Total 29: (19 in last 5 years) including 3 Nature, 1 Science, 1 Nature Genetics, 3 PNAS. 3 first author papers and h-index 15

Nones K, Waddell N, Song S, **Patch AM** et al. Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling. **Int J Cancer** (accepted 16/01/14)

Kassahn KS, Holmes O, Nones K, **Patch AM** et al. Somatic point mutation calling in low cellularity tumors. **PLoS One**. 2013 Nov 8;8(11)

Weedon MN, Cebola I, **Patch AM** et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. **Nat Genet**. 2014 Jan;46(1):61-4

Chou A, Waddell N, Cowley MJ, Gill AJ, Chang DK, **Patch AM** et al. Clinical and molecular characterization of HER2 amplified-pancreatic cancer. **Genome Med**. 2013 Aug 31;5(8):78.

Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, **Patch AM**, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. **Nature.** 2012 Nov 15;491(7424):399-405

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Network-based pan-cancer data integration for tumor stratification and identification of inter- and intra-tumor type mechanisms of action

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Prof. Steven Van Laere, KU Leuven & University of Antwerp, member of the ICGC Breast Cancer Group

Dr Ewan Birney, Associate Director, European Bioinformatics Institute. ICGC Breast Cancer Group

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
	Prof. Jan Fostier Department of Information Technology, Ghent University - iMinds, Belgium **Prof. Kathleen Marchal** Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

Background and preliminary data

Cancer is a heterogeneous disease that can be caused by different molecular events, and thus response to treatments varies among patients. Different tumor subtypes could be reflected at specific molecular levels (genomics, epigenomics or transcriptome), thus an integrative analysis is required to better understand the mechanisms responsible for such heterogeneity. By incorporating pathway interaction data into the analysis, we aim at providing the molecular context in which (epi)genomic alterations influence tumor stratification.
Our research focus is the development of scalable integrated methods, which we have applied to public (TCGA) breast cancer data, with a focus on networks and per-patient analysis. A new data model is being developed that shows great promise for answering questions ranging from tumor subtyping to per-patient sub-network delineation (manuscript in preparation, see project proposal).

During this research, the complementarity of different data sources (mRNA, miRNA, methylation, copy number and mutation) was evident even from unsupervised analysis (very few clinical data were available). Yet, we suspect that the obtained results often reflect only the marginal effects of strongly interacting genetic (somatic and germline) and epigenetic factors. A large multi-omics pan-cancer data set should allow for a better detection of interacting drivers of cancer and could reveal one or more cross-cancer mechanisms in tumor development.

As proof-of-principle, we will focus on patient samples classified as inflammatory-like according to a model described by Van Laere et. al. (Clin Cancer Res, 2013). This model was developed by comparing gene expression profiles of patients with and without inflammatory breast cancer (IBC), an aggressive form of breast cancer with high metastatic potential. In the above referenced paper, it was shown that the model has prognostic value in a large series of patients with non-IBC, which was independent from classical parameters such as tumor grade and hormone receptor status. Obviously, identifying driver events for this molecular phenotype might have significant clinical impact.

Timelines & resources dedicated to project
One full time post-doc (Jimmy Van den Eynden, Ph.D., MD, Ghent University) and one part-time post-doc (Ana Carolina Fierro, Ghent University) will be allocated, as well as one full time Ph.D. student (funded by Ghent University). For this project, the research partners will have access to the Flemish supercomputing facilities including the Top500 Tier 1 supercomputer. Prof. Van Laere will dedicate 10% of his research time to this project and will assist in guiding the project and disseminating the results. Ewan Birney will provide input into the design of the analysis and the interpretation of the results.

Research proposal
Our group is developing methods that allow for the simultaneous analysis of multiple types of 'omics data in the context of cancer. Analysis of inter- and intra-tumor subtype variability is based on a novel network based data representation where nodes can correspond to different types of entities including genes, mutations, copy number aberrations, methylations and patients. Nodes are then connected through edges that represent relations between different or identical entities. This data representation allows for intuitive integration of prior knowledge in the analysis. Population structure can be added through patient-patient links, known gene interactions are added through gene-gene links, etc. Another advantage is that new types of data are easily integrated as long as meaningful relationships with the other entities in the data model can be identified. The data model allows for different types of analysis that all have in common that before anything else, a network-based similarity measure is calculated that relates all entities in the network to all other entities. Crucial to the calculation of the network-based similarities is the application of graph node kernel-based techniques that we previously used in gene prioritization systems (EPSILON: an eQTL prioritization framework using similarity measures derived from local networks, Verbeke et al., 2013, Bioinformatics). On top of these similarities, we built methods for unsupervised tumor subtyping, mechanism of action detection, driver identification, per-patient pathway impact analysis and per-patient sub-network delineation (manuscript in preparation). Because our methods can be used to investigate inter- and intra-subtype variation, we think they are well suited for the analysis of large pan-cancer, multi-omics datasets. Running the analysis on thousands of tumors can help in understanding the common mechanisms underlying cancer. In a first step we want to identify clusters of related tumor samples and determine the corresponding common mechanisms of action, if any. In a second step we will look for common or mutually exclusive genetic drivers of cancer within the previously found clusters. We especially think that a larger number of data points can reveal interactions between somatic and germline mutations that we were previously unable to detect using public single cancer datasets. Also, applying our analysis techniques in combination with RNA-seq data should allow for the identification of particular post-translational modifications and regulation strategies that give tumor cells a growth advantage. By applying our algorithms to 'omics data of cancer patients classified as inflammatory-like, we hope to identify driver events underlying this phenotype. Although the classification model relates to breast cancer, this analysis is relevant for cancer in general, since the gene expression model appears to offer an intriguing explanation for the frequent observation of tumor emboli in IBC. Tumor emboli are considered as the exponent of aggressive cancer cell behavior in this disease. But, tumor emboli are not only observed in this particular type of breast cancer, but also in other, usually aggressive cancer types such as malignant melanoma and clear cell renal cell carcinoma. Therefore, uncovering the biological principals and drivers of this phenomenon might be applicable to cancer in general and might be of clinical interest. We realize that integrating large amounts of data not only raises conceptual challenges but will also require practical engineering solutions. To address these, we have included high performance computing aspects and parallelization from the very beginning, resulting in methods that scale well for multi-patient and multi-omics data sets.
Legacy plans
Any intermediate and final results will be made available and archived according to the principles and guidelines of the ICGC/TCGA WGS pan-cancer project. Any software tools or analysis pipelines that will be developed in the course of this project will be made available, at least as source code, but preferably as an end-user ready analysis tool. Additionally, we expect to construct a pan-cancer model that – once computed – should allow for the analysis of novel tumor samples (from previously unseen patients). The precomputed model represents an integrated knowledge base that will be made available to all interested parties.

CV Steven J Van Laere

A. Affiliation

Prof. Dr. Steven Van Laere (Steven.VanLaere@med.kuleuven.be)
Department of Oncology, KU Leuven, Leuven, Belgium
Faculty of Medicine and Health Sciences, Antwerp University, Antwerp, Belgium
Member of the ICGC Breast Group

B. Education/training

Bachelor degree: 07/2000, Antwerp University, Belgium; Master degree: 07/2002, Antwerp University, Belgium
PhD in medical sciences: 04/2009, Antwerp University, Belgium: "Molecular profiling of inflammatory breast cancer by genome-wide gene expression analysis"

C. Positions and Employment

2002-2009: PhD Student, University of Antwerp, Belgium
2009-2011: Postdoctoral researcher, Translational Cancer Research Unit, GZA Hospitals Sint-Augustinus, Wilrijk, Belgium
2011-present: Associate Professor, Department Oncology, KU Leuven, Leuven, Belgium
2013-present: Associate Professor, Faculty of Medicine and Health Sciences, Antwerp University, Antwerp, Belgium

D. Research Focus:

The research of Prof. Van Laere, primarily focuses on deciphering the molecular biology responsible for the development of metastases in patients with breast cancer. The research program consists of 3 major topics. One of these involves the study of inflammatory breast cancer (IBC), an aggressive and highly metastatic form of locally advanced breast cancer. The research group is currently leading an international effort that aims at (re)defining the molecular profile of IBC. A second research topic relates to the study of circulating tumor cells (CTCs), or tumor cells that reside in the circulatory system of patients with cancer. CTCs are actively metastasizing tumor cells and constitute an interesting cell population, both from biological and clinical perspective. Using Next Generation Sequencing technology, TCRU aims at characterizing these cells at the genomic and transcriptomic level. Relevant research questions address the intra-patient heterogeneity of this cell population, the identification of novel targets for treatment and the acquisition of novel insights into the biology of these tumor cells. A final research topic involves the study of the molecular biology of liver metastases in patients with cancer. Particularly interesting are liver metastases from patients with breast cancer, which are characterized by a very specific growth pattern in which tumor cells replace normal hepatocytes without disturbing the liver architecture. This growth pattern has important clinical implications, because these metastases are difficult to detect using classical imaging technologies. In addition, resistance to anti-angiogenic therapy is often observed in these patients. For all these reasons, new predictive biomarkers for this type of liver metastases is mandatory and a more comprehensive biological blueprint of these tumors is urgently needed.

E. Honors

AACR-Pezcoller Foundation Scholar-in-Training Award for the AACR 101st Annual Meeting 2010

Curriculum Vitae, Ewan Birney

Full Name: John Frederick William Birney
Date of Birth: 12 December 1972
Nationality: UK
Email: birney@ebi.ac.uk

77 Lancaster Road
London N4 4PL

Employment:

2012-Current : Associate Director, European Bioinformatics Institute
2000-2012: Head of Nucleotide data, European Bioinformatics Institute
Current supervisor for 4 PhD students

On a variety of SAB boards (includes Riken Institute, BCGSC, Leipzig MPI, Roslin Institute, IMP, TGAC)

1996-2000: PhD at the Sanger Centre (Supervisor, Richard Durbin)

Other positions held:

- A number of consultancy contracts, both strategic and technical in the biotech and pharmaceutical industry, including funding and finance orientated roles.
- Equity Research in SBC Warburg Pharmaceutical division (summer 1995).
- Freelance journalist (Economist) (1995).
- Research Assistant at Cold Spring Harbor Laboratory and EMBL Heidelberg.

Prizes and Awards

EMBO Member, Elected 2012
Winner of the Overton Award from the International Computational Biology Society, 2005
Winner of the Benjamin Franklin Award from Bioinformatics.org/BioIT in 2005
Winner of the Royal Society's Francis Crick Lecture in 2003

Patents:

US Provisional Patent Application 61/654295, *High-capacity storage of digital information in DNA*, filed 1 June 2012 (co-applicant with Nick Goldman)

Education:

1996-1999: PhD, St John's College Cambridge. Awarded a Scholarship
1992-1996: BA Biochemistry, Balliol College Oxford. 1st Class degree. Awarded a Scholarship

Publications

181 Peer reviewed publications, 23 in Nature (5 first/last author), 9 Science (1 last author). 1 Cell (joint last author). H-index: 83. Avg Citations/Paper 331. (Google Scholar)

CV Jan Fostier

A. Affilliation

Prof. dr. ir. Jan Fostier (jan.fostier@intec.ugent.be)
Department of Information Technology (INTEC), IBCN research group, Ghent University - iMinds

B. Education/training

Ghent University, Belgium, Bachelor (2003) and Master (2005) of Science in Engineering: Applied Physics
Ghent University, Belgium, Bachelor (2004) of Science in Informatics
Ghent University, Belgium, PhD in Engineering Physics (2009): computational physics
Visiting research at the EMBL-EBI, Cambridge, UK with dr. Ewan Birney for three months

C. Positions and Employment

2005-2009	PhD Fellow, Faculty of Engineering, Ghent University
2009-2011	Postdoctoral Fellow, Dept. of Information Technology (INTEC), Ghent University
2011-present	Assistant professor, Dept. of Information Technology (INTEC), Ghent University (100%)

Research focus:

Jan Fostier was recently (2011) appointed tenure track professor at the IBCN research group (computer science) within the faculty of engineering at Ghent University. He has a background in computational methods, algorithm design and high performance computing, originally applied to problems in the domain of computational physics. In 2010, the joined the Multidisciplinary Research Partnership (MRP) "from nucleotides to networks" (N2N – see http://www.fromnucleotides2networks.be), a centre of excellence in bioinformatics.

Within this MRP, he collaborates with biologists, bioinformaticians, statisticians and mathematicians in order to develop novel methods for evolutionary biology (i-ADHoRe 3.0), comparative genomics (BLSSpeller), genomic prediction (DAIRRy-BLUP) and systems biology (EPSILON). His focus is on method development, with an emphasis on computationally intensive and/or big data applications. He is a partner in the 'ExaScience Life Lab' (funded by the IWT), a collaboration between the 5 Flemish universities, Johnson & Johnson and Intel which aims to bring existing expertise in the field of high performance computing to the domain of life sciences.

J. Fostier is currently (co-)supervisor of 5 Ph.D. students. In particular, J. Fostier and K. Marchal are co-supervising Lieven Verbeke in the domain of network-based data integration for cancer research, currently applied to breast tumors.

CV Kathleen Marchal

A. Affilliation

Prof. Dr. Kathleen Marchal (Kathleen.marchal@ugent.be)
Department of Plant Biotechnology and Bioinformatics, Ghent University
Department of Information Technology (INTEC), Ghent University
Department of Microbial and Molecular Systems, KU Leuven

B. Education/training

KU Leuven, Belgium, bachelor 07/1992, master degree Bio-engineering, 07/1995
KU Leuven, Belgium, teaching degree, 09/1998
PhD in Bioscience engineering, KU Leuven, Belgium, 11/1999, microbial molecular biology

C. Positions and Employment

1996-1999	PhD Fellow, Fac of Bioengineering, KU Leuven
1999-2004	Postdoctoral Fellow, FWO, Dep. of electrical engineering, KU Leuven
2004-2008	Lecturer (docent), Dep. Microbial and Molecular Systems, KU Leuven
2008- 2011	Associate professor (hoofddocent), Dep. Microbial and Molecular Systems, KU Leuven
2011-	Associate professor (hoofddocent), Dep. Plant Biotechnology and Bioinformatics, Dep. Plant Systems Biology, Ghent University (100%)

Research focus: Kathleen Marchal (http://bioinformatics.psb.ugent.be/DBN/) started her group in 2004 at the KU Leuven (Dept. of Microbial and Molecular Systems). Since 2011 she moved to Ghent University. Her interdisciplinary research group currently consists of 3 postdocs and 8 PhD students and has more than 100 peer reviewed publications (http://www.researcherid.com/rid/B-5001-2013) (H-index 27). Her research focuses on the development of computational methods for systems biology and dataintegration. The group has developed several methods to infer transcriptional regulatory programs from high throughput omics data (ReMoDiscovery, Distiller), for cross-species clustering (COMODO) and for the detection of transcription factor binding sites (Motif suite, CPModule). Currently the group focuses on the use of interaction networks for genotype-expression phenotype mapping in clonal species.

K. Marchal has coordinated the larger research consortium Bioframe (http://www.bioframe.net/) an algorithmic platform for integrative modelling in systems biology (supported by IWT). It was also member of the Center of Excellence Symbiosys (ends in 2010, http://www.kuleuven.be/symbiosys/; KU Leuven) and recently of the Center of Excellence NATAR (start in 2010, KU Leuven) focusing on genotype-phenotype mapping in microbes and the MRP 'from nucleotides to networks" (UGhent). K. Marchal is coordinator of the scientific research community SYNCELLS (http://bioi.biw.kuleuven.be/syncells/) (at the interface between systems and synthetic biology). She has been supervising more than 22 PhDs students. She is member of several national (e.g., Fund for scientific research Flanders) and international boards (e.g. INSERM, France).

Honors: She obtained the Laureaat DSM prijs voor Chemie en Technologie, 2000 and the 2002 Biannual Siemens award (2002)

Editorial work:

2000 -	Associate Editor, BMC release notes
2011 -	Associate Editor, Journal of Integrative Omics
2012-	Associate Editor, BMC Bioinformatics
2013-	Associate Editor, Plant Journal

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Cancer driver mutations in cellular interaction networks

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gary D. Bader, the Donnelly Centre, University of Toronto

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Jüri Reimand, the Donnelly Centre, U. Toronto Mohamed Helmy, the Donnelly Centre, U. Toronto	

Background and preliminary data

We hypothesize that many cancer driver mutations specifically alter cellular networks. Complex networks of protein-protein interactions, signal transduction and gene regulation give rise to phenotypes and are perturbed in cancer. Molecular interactions are mediated by sites in protein, RNA and DNA sequences that can be affected by mutations, potentially rewiring interactions. Analysis of somatic mutations in the context of biomolecular networks is important in elucidating cellular mechanisms underlying cancer.

We studied cancer driver mutations related to phosphorylation, a post-translational modification (PTM) central to cancer processes and targeted therapies. We devised the Poisson regression model, ActiveDriver, to find frequently mutated hotspots coinciding with specific protein sites (*Reimand and Bader 2013, Mol Sys Biol*). ActiveDriver assumes that recurrent site mutations are unlikely unless the site is important in cancer. We analyzed ~80,000 phosphosites and ~10,000 SNVs in ~800 cancer genomes, revealing ~1,000 phosphorylation-related SNVs (pSNVs) enriched in known cancer drivers and pathways, kinase signaling networks, and pSNV signatures informative of patient survival. Thus somatic mutation of phosphorylation is an important oncogenic mechanism, and interpreting SNVs in signaling context helps decipher mechanisms of known cancer genes and predict new genes. We also studied the TCGA pan-cancer dataset of 3,200 tumors of 12 types and found ~16x more pSNVs occurring in 90% of tumors (*Reimand et al 2013, Nat Sci Rep*). We developed a novel method MIMP, to assess the impact of pSNVs on kinase binding specificity (*Wagih, Reimand, Bader, in prep*). MIMP predicts whether pSNVs disrupt or create new kinase binding sites, potentially leading to oncogenic loss and gain of signaling. We have recently expanded this analysis to consider additional PTM types (phosphorylation, acetylation, ubiquitination) and genomes (1000 Genomes, ESP6500). We will leverage these resources to study network-rewiring cancer drivers in the ICGC project, extending our method to take advantage of whole genome sequence data, including non-coding region analysis.

Timelines & resources dedicated to project

We will analyze data as soon as variants are called and available. Two postdoctoral fellows will work on this project. Local, secure compute capacity will be used for early method development, but we anticipate increasing use of available cloud resources once they are available.

Research proposal
We will study cancer mutations such as single nucleotide variants (SNVs) that potentially rewire cellular interaction networks. We will initially focus on SNVs as these are likely to be the highest confidence and also easiest to interpret in terms of their effect on short binding sites. We will initially focus on protein-coding variants and experimentally derived signaling sites (post-translational modifications, PTMs), in particular kinase-signaling networks for which most comprehensive data are available. We will extend our ActiveDriver model to non-coding sequence to identify statistically significant mutation hotspots in gene regulatory sites, such as splicing motifs in introns, transcription factor binding sites in gene promoters, using validated sites from existing technologies (e.g. ChIP-seq) and high-confidence binding models (position weight matrices) of diverse protein interaction domains in collaboration with relevant experts. We will predict network-rewiring cancer driver mutations based on their mutation frequency and explore the effect of additional variables to mutational heterogeneity such as protein disorder, evolutionary conservation, and GC content in the ActiveDriver model. We will investigate the mutational impact to critical residues in binding sites using the MIMP method and extend this approach to broader protein, DNA and RNA motifs and domains to reveal oncogenic rewiring in diverse cellular networks. We will consider cellular context of molecular interactions using paired RNA-seq and proteomics data (where available) to identify the most confident hypotheses of mutation-induced network rewiring events. We will also apply paired -omics data to computationally validate specific mutations, such as differential expression of target genes in cells with mutated transcription factors, or differential chromatin state of tumors with histone mutations in acetylation sites. We will perform individual cancer type and pan-cancer analyses to find type-specific and general network-rewiring signatures, and will explore unsupervised clustering to predict pan-cancer subtypes with distinct network-mutation signatures. Pathway enrichment analysis will be used to interpret rare network-rewiring mutations. We will correlate mutations with clinical information where available, for example to find survival-associated interaction network modules with mutations, using methods we developed earlier (HyperModules; *Reimand and Bader 2013 Mol Sys Biol*). We will compare pairs of germline/tumor genomes, and pairs of primary and recurrent/metastatic tumors to identify candidate network-rewiring mutations in cancer pre-disposition, and post-treatment tumor evolution, respectively. In summary, the integration of cancer mutations and the context of network interactions will predict functional consequences of mutations with testable mechanistic hypotheses, help distinguish drivers from passengers, and provide pathway context to connect rare mutations into altered cellular systems.

Legacy plans
We will make new and updated tools and statistical models available to the community as free open-source software packages, as we have a great track record of doing. Results of statistical analyses, such as gene lists, mutations and pathways will be published according to consensus guidelines of the project on the Synapse system.

SHORT CURRICULUM VITAE

Name
Gary D. Bader
Position/Title
Associate Professor

Institutional Affiliation
The Donnelly Centre, University of Toronto

WORK EXPERIENCE

Period	*Position*	*Institution*
2011-present	Associate Professor	University of Toronto (The Donnelly Centre, Department of Molecular Genetics, Department of Computer Science)
2006-2011	Assistant Professor	University of Toronto (The Donnelly Centre, Department of Molecular Genetics, Department of Computer Science)
2006-present	Associate Member	Lunenfeld Research Institute, Mount Sinai Hospital, Toronto
2002-2006	Post-doctoral Fellow	Memorial Sloan-Kettering Cancer Center, New York

ACADEMIC AND TRAINING BACKGROUND

Period	*Degree*	*Institution and Location*	*Field of Study*
2003-2006		Memorial Sloan-Kettering Cancer Center	Computational Biology
1998-2002	Ph.D.	University of Toronto, Toronto	Biochemistry
1993-1997	B.Sc.	McGill University, Montreal	Biochemistry

SELECTED PEER-REVIEWED PUBLICATIONS *(from 95 total)*

1. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013 Jan 22;9:637
2. Northcott PA et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature. 2012 Aug 2;488(7409):49-56
3. Witt H, Mack SC, Ryzhova M, Bender S, Sill M, Isserlin R, Benner A, Hielscher T, Milde T, Remke M, Jones DT, Northcott PA, Garzia L, Bertrand KC, Wittmann A, Yao Y, Roberts SS, Massimi L, Van Meter T, Weiss WA, Gupta N, Grajkowska W, Lach B, Cho YJ, von Deimling A, Kulozik AE, Witt O, **Bader GD**, Hawkins CE, Tabori U, Guha A, Rutka JT, Lichter P, Korshunov A, Taylor MD, Pfister SM. Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. Cancer Cell. 2011 Aug 16;20(2):143-57
4. Merico D, Isserlin R, Stueker O, Emili A, **Bader GD**. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010 Nov 15;5(11):e13984. PMCID: PMC2981572
5. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, **Bader GD**, Sander C. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011 Jan;39(Database issue):D685-90. Epub 2010 Nov 10. PMCID: PMC3013659
6. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, **Bader GD**. Cytoscape Web: an interactive web-based network browser. Bioinformatics. 2010 Sep 15;26(18):2347-8. Epub 2010 Jul 23. PMCID: PMC2935447
7. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, **Bader GD**, Morris Q. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010 Jul 1;38 Suppl:W214-20. PMCID: PMC2896186
8. Pinto D et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9. PMCID: PMC3021798
9. International Cancer Genome Consortium. International network of cancer genome projects. Nature. 2010 Apr 15;464(7291):993-8.
10. **Bader GD**, Betel D, Hogue CW BIND-The Biomolecular Interaction Network Database *Nucleic Acids Research* Jan 1, 2003 31(1): 248-250 PMCID: PMC165503

SHORT CURRICULUM VITAE

Name
Jüri Reimand
Position/Title
Postdoctoral Fellow
Institutional Affiliation
The Donnelly Centre, University of Toronto

WORK EXPERIENCE

Period	*Position*	*Institution*
2011-present	Postdoctoral Fellow	University of Toronto (The Donnelly Centre, Department of Molecular Genetics, Department of Computer Science)
2012-present	Postdoctoral Fellow	SickKids Research Institute (Department of Developmental and Stem Cell Biology)
2006-2010	Scientific programmer	University of Tartu (Department of Computer Science)
2007-2008	Marie Curie Fellow	EMBL European Bioinformatics Institute, Cambridge UK

ACADEMIC AND TRAINING BACKGROUND

Period	*Degree*	*Institution and Location*	*Field of Study*
2003-2006		The Donnelly Centre, University of Toronto	Computational Biology
1998-2002	Ph.D.	University of Tartu, Estonia	Computer Science
1993-1997	B.Sc.	University of Tartu, Estonia	Computer Science

SELECTED PEER-REVIEWED PUBLICATIONS *(from 22 total)*

1. **Reimand J**, Wagih O, Bader GD. The mutational landscape of phosphorylation signaling in cancer. *Nat Sci Rep.* 2013 Oct 2;3:2651.
2. **Reimand J**, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol*. 2013 Jan 22;9:637.
3. Northcott PA *et al*. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*. 2012 Aug 2;488(7409):49-56.
4. Uusküla L, Männik J, Rull K, Minajeva A, Kõks S, Vaas P, Teesalu P, **Reimand J***, Laan M*. Mid-gestational gene expression profile in placenta and link to pregnancy complications. *PLoS One* 2012;7(11):e49248.
5. **Reimand J**, Aun A, Vilo J, Vaquerizas JM, Sedman J, Luscombe NM. m:Explorer: multinomial regression models reveal positive and negative regulators of longevity in yeast quiescence. *Genome Biol.* 2012 Jun 21;13(6):R55.
6. Altmäe S*, **Reimand J***, Hovatta O, Zhang P, Kere J, Laisk T, Saare M, Peters M, Vilo J, Stavreus-Evers A, Salumets A. Research resource: interactome of human embryo implantation: identification of gene expression pathways, regulation, and integrated regulatory networks. *Mol Endocrin*. 2012 Jan;26(1):203-17.
7. **Reimand J**, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM. Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Res*. 2010 Aug;38(14):4768-77.
8. Schulz H *et al*. The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS One* 2009 Sep 3;4(9):e6804.
9. **Reimand J***, Tooming L*, Peterson H, Adler P, Vilo J. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.* 2008 Jul 1;36:W452-9.
10. **Reimand J**, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 2007 Jul;35:W193-200.

SHORT CURRICULUM VITAE

Name
Mohamed Helmy
Position/Title
Postdoctoral Researcher

Institutional Affiliation
The Donnelly Centre, University of Toronto

WORK EXPERIENCE

Period	*Position*	*Institution*
2013(9)-present	Postdoc. Researcher	The Donnelly Centre, University of Toronto, Canada
2012-2013	Postdoc. Researcher	G. Sch. of Pharmaceutical Sciences, Kyoto University, Japan
2011-2012	Research Fellow	Japan Society for promotion of Science (JSPS), Japan
2009-2011	Research Assistant	The Global Center of Excellence (G-COE), Keio University
2007-2012	Research Assistant	Institute for Advanced Biosciences, Keio University, Japan

ACADEMIC AND TRAINING BACKGROUND

Period	*Degree*	*Institution and Location*	*Field of Study*
2009-2012	Ph.D.	Institute for Advanced Biosciences, Keio University, Japan	Systems Biology
2007-2009	M.Sc.	Institute for Advanced Biosciences, Keio University, Japan	Systems Biology
2001-2002	Diploma	Information Technology Institute (ITI), Egypt	Software development
1996-2000	B.Sc.	Al-Azhar University, Egypt	Genetics

SELECTED PEER-REVIEWED PUBLICATIONS *(from 12 total)*

1- Elmtwally, S, Hamza, T, Zakarya, M and **Helmy, M*.** Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges, PLOS Computational Biology, (In Press).

2- **Helmy, M.**, Gohda, J., Inoue, J., Tomita, M., Tsuchiya, M. and Selvarajoo, K. (2009) Predicting novel features of Toll-like Receptor 3 signalling in macrophages, PLoS ONE, 4(3): e4661. doi:10.1371/journal.pone.0004661.

3- Selvarajoo, K., Takabe, Y., Gohda, J., **Helmy, M.**, Akira, S., Tomita, M., Tsuchiya, M., Inoue, J. and Matsuo, K. (2008) Signalling Flux Redistribution at Toll-like Receptor Pathway Junctions., PLoS ONE , 3(10) , e3430, doi:10.1371/journal.pone.0003430.

4- **Helmy, M.**, Tomita, M., Ishihama Y. (2011) OryzaPG-DB: Rice Proteome Database based on Shotgun Proteogenomics, BMC Plant Biology, 11,63, 2011. doi:10.1186/1471-2229-11-63.

5- **Helmy, M**, Tomita M, Ishihama Y. (2012) Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics. Genes, Genomes and Genomics, 6(SI1), 76-85.

6- **Helmy, M**, Sugiyama N, Tomita M, Ishihama Y. (2012) The rice proteogenomics database OryzaPGDB: development, expansion and new features. Frontiers of Plant Sciences, 3:65, doi: 10.3389/fpls.2012.00065.

7- **Helmy, M**, Sugiyama N, Tomita M, Ishihama Y. (2012) Mass Spectrum Sequential Subtraction: bioinformatics method facilitates searching large dataset of peptide MS/MS spectra against large nucleotide databases for proteogenomics. Genes to Cells, 17(8), 633-644, doi: 10.1111/j.1365-2443.2012.01615.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Population-based detection of structural variants in normal and aberrant genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Guillaume Bourque, McGill University

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Jean Monlong, PhD student	

Background and preliminary data

While several approaches are currently used to detect Structural Variation from High-Throughput Sequencing in cancer, problematic regions of the genome such as low-mappability or repeat-rich regions are poorly addressed and generally excluded from analysis. We propose to overcome these limitations by using a large number of samples as reference to identify abnormal patterns. By integrating a large number of reference experiments, abnormal variation can be robustly detected, even in low-mappability regions.

Pan-cancer project is particularly appropriate for our approach. First, its enormous sample size will ensure extensive resolution. Secondly, thanks to whole-genome sequencing provided, our approach will be able to examine low-mappability regions to unprecedented resolution. Finally, cell-heterogeneity or temporal cancel evolution inference from our robust SVs calls will benefit from the availability of resequenced samples (primary/metastasis or diagnosis/progression).

On Pan-cancer project and using our population-based approach, we would focus on:

- Copy-Number Variation in low-mappability regions: repeat-rich, centromeric/telomeric regions, satellites expansion.
- Partial Copy-Number signal and cell-heterogeneity inference.

We are currently developing and testing our approach on renal cancer data. Preliminary results show, as expected, superior performance compared to state-of-the-art methods, especially in problematic regions.

Timelines & resources dedicated to project

Our approach requires aligned reads from whole-genome or exome sequencing experiments.

Research proposal

Our population-based approach compares Read-Depth (RD) signal between a sample of interest and a large set of reference samples. In Pan-cancer project, the normal samples will be used as reference and will define “normality”. Abnormality in tumor (or normal) samples is hence detected through divergence from the reference samples. In practice, RD signal is summarized by cutting the genome into non-overlapping windows and counting reads mapped within.

In the first step, a set of homogeneous reference samples is defined. In Pan-cancer project, samples will likely originate from different populations with different ancestry history. Hence, some effort might be necessary to define appropriate reference set(s). Then, RD signal is normalized across samples in order to ensure valid prospective test and remove variation of technical origin. We observed that a general normalization is not sufficient to correct systematic sample-specific variation and developed a flexible and targeted approach yielding solid results. Finally and for each sample, RD in each genomic window is compared to the RD in the reference samples by computing a Z-score. After adjustment through multiple-testing correction, abnormal regions are identified. Copy-number estimate of a region is given by comparing the RD in the tested sample to the mean value across the reference samples, assumed to represent diploidy.

Thanks to paired sample information and extensive sample size, somatic and germline variants of good quality will be retrieved. Additionally, exome-sequencing, RNA-Seq and methylation experiment could be used to confirm relevant candidates. Centromeric-telomeric regions, as well as satellites and transposable elements will be investigated directly from the genome-wide calls. Moreover, targeted test could be specifically designed to interrogate particular regions as active transposons or satellites and achieve optimal resolution. Through incomplete copy-number estimates, cell-heterogeneity of tumoral sample will be assessed. Comparing sensitive measure of variant heterogeneity across samples from the same or different cancers should highlight essential candidates and assist inference of cancer temporal evolution. Furthermore, additional resequencing of tumor at different stages/location will be used to reinforce this inference. Potential sample contamination will also be investigated through study of partial copy number estimates in normal samples or comparison of RD signal between normal/tumor paired samples.

Finally, potential balanced strucutral variation or de novo insertion will be marked using the same population-based approach but using only discordant read pairs. Once thoses regions marked, available approach using intensive computation (e.g. re-mapping, assembly, database mapping) will be used to characterize the event detected. Again, we hope to robustly include problematic regions in this analysis.

Legacy plans

We plan to publish and make our software available open source. Our population-based approach as well as extra analysis functions will be made available, most likely in the form of R script/package.

NAME	POSITION TITLE
Bourque, Guillaume	Associate Professor, Department of Human Genetics, McGill University and Bioinformatics Director, McGill University & Genome Quebec Innovation Center

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Université de Montréal, Montréal, Canada	B.Sc.	1998	CS and Mathematics
University of Southern California, CA	M.A.	2000	Applied Mathematics
University of Southern California, CA	Ph.D.	2002	Applied Mathematics
Université de Montréal, Montréal, Canada	Postdoc	2004	Computational Biology

A. EXPERTISE KEYWORDS

Genomics, Bioinformatics, Genome Analysis, High-performance Computing, Gene Regulation, Comparative Genomics, Human Genetics, Transposable Elements, Epigenetics

B. POSITIONS AND HONORS.

Positions and Employment:

2010 – Bioinformatics Director, McGill University & Genome Quebec Innovation Center, Montréal

2010 – Associate Professor, Department of Human Genetics, McGill University, Montréal

2007 – 2010 Senior Group Leader & Associate Director, Computational & Mathematical Biology, Genome Institute of Singapore

2004 – 2007 Group Leader, Information and Mathematical Sciences, Genome Institute of Singapore

Honors and Awards:

2012 Chercheur-boursier Junior 2, Fonds de recherche Santé Québec (FRSQ)

C. SELECTED RESEARCH SUPPORT.

2012 – 2017 Integrative Epigenomic Data Coordination Centre (EDCC) at McGill**.** (CIHR, CEEHRC Epigenomics Platform, C$ 1,500,000, Role: PI)

2012 – 2017 Multidimensional Epigenomics Mapping Centre (EMC) at McGill. (CIHR, CEEHRC Epigenomics Platform, C$ 5,850,000, PI: Mark Lathrop, Role: Co-Inv)

2011 – 2015 Functional characterization of the endogenous retrovirus HERV-H family in human embryonic stem cells. (CIHR, Operating Grant, C$ 426,482, Role: PI)

D. SELECTED PEER-REVIEWED PUBLICATIONS (out of a total of 51)

1. **Bourque G**, Leong B, Vega VB, Chen X, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, Liu ET. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* 2008 18(11):1752-62.
2. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, **Bourque G**. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet.* 2010 42(7):631-4.
3. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, **Bourque G**, Yandell M, Feschotte C. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013 9(4):e1003470.
4. Jacques PE, Jeyakani J and **Bourque G**. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements*. PLoS Genet.* 2013 9(5):e1003504.

Jean MONLONG
Date of birth: August 3rd 1988 (25 year old)
Nationality: French
4143 Boulevard Saint-Laurent, apt 2
H2W 1Y7 Montreal
Phone: (+1) 514 691 5326
Email: jean.monlong@mail.mcgill.ca
Driver's license

Education

2012 – now McGill University, Montreal, Canada.
PhD thesis in Human Genetics Department in Guillaume Bourque's Lab..

2010 – 2011 Universitat Politècnica de Catalunya, Barcelona, Spain.
Master in Statistic and Operations Research as an exchange student in the Faculty of Mathematics and Statistics.

2008 – 2012 Ensimag, Grenoble, France.
Computer Science and **Mathematics** course with specialization in **Bioinformatics**.

2006 – 2008 Lycée Michel Montaigne, Bordeaux, France.
Preparatory classes for entrance to the Grandes Écoles. Mathematics and Physics.

2003 – 2006 Lycée de Navarre, St Jean-Pied-de-Port, 64, France.
High School student with specialisation in Mathematics.

Professional experiences and projects

2011 – 2012 (1 year) Centre for Genomic Regulation(CRG), Barcelona, Spain.
Graduation project in Roderic Guigó's bioinformatics group. Comparison of splicing activity from RNASeq experiments. Participation in international projects: Geuvadis (Nature 2013) and GTEx (articles in preparation)..

2011 (3 months) Universitat Politècnica de Catalunya, Barcelona, Spain.
Study of the regularization of the generalized canonical correlation analysis.
Poster presentation at the *XIIIth Spanish Biometry Conference and 3rd Ibero-American Biometry Meeting - 2011.*

2010 (3 months) Neomades(Mobile software development), Bidart, 64, France.
Implementation of a Java module of their principal product.

2010 (1 months) Ensimag, Grenoble, France.
Breast cancer modelization and Bayesian estimation of the overdiagnosis rate. Collaboration with La Tronche's hospital(Grenoble). Article published in the *British Medical Journal.*

Skills and languages

◇ Deep knowledge of **statistic** and **mathematical modelization** principles.
◇ Deep knowledge of **computer science** and **algorithmic**.
◇ Knowledge of **biology** for Bioinformatic.
◇ Deep experience of several programming languages and softwares : **C**, **R**, **Java** et **Perl**.
◇ Experience with other languages : C++, Ada, GNU Assembly, Scilab, SQL, HTML/CSS, PHP, LaTeX, Pack Office.

English : fluent.

Spanish : fluent.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Data Standardization: Removing Site-Specific Biases in WGS Data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Dr. Paul C. Boutros, Principal Investigator, Informatics & Biocomputing, Ontario Institute for Cancer Research; Canadian Prostate Cancer Genome Network (CPC-GENE: Canadian Prostate ICGC)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

The ICGC Benchmark studies, benchmarking work done by my own lab, and the preliminary results of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge suggest that the pan-cancer studies will need to control for very substantial biases between sequencing centers. Some of those biases will be removed by the use of a consistent analysis pipeline, but others are going to remain. I propose to attempt to quantitate these and attempt to remove these biases.

ICGC Benchmarking Study 2 has shown that sequencing different aliquots of the same DNA sample at different centers yields markedly different error- and coverage-profiles. These lead to very poor overlap in mutation calls made at different centers, even after the same analysis pipeline is applied to each (which may or may not be appropriate). It will be critical to understand where those biases are and to try to evaluate what their effect will be on each type of downstream analysis.

Timelines & resources dedicated to project

One FTE of biostatistician is in-place and is working on these problems in the context of integrating prostate cancer data from different sequencing centers for a meta-analysis. This analysis would require aligned data, but would start at that point. We would hope to provide to downstream analyses:

a) per-tumour assessment of coverage biases (4-6 weeks after alignments are finished)

b) per-tumour-type assessment of biases with reduced statistical confidence (10-14 weeks after alignments are finished)

c) per pathway-analyses of the effects of coverage biases (4-8 weeks after the prior two analyses are completed)

d) quantify bias for individual non-coding genomic elements (4-8 weeks after priot three analyses)

Research proposal

We propose three separate aims:

Aim 1: Quantify per-base coverage differences
We will look at aligned WGS data only and identify bases that have aberrantly high or low coverages across all tumours or in individual tumours. We will identify associations in coverage by sequencing-site, and will attempt to use tumours of similar types and the ICGC benchmark data in as control data. Our deliverable at this stage will be lists of individual genomic regions where coverage challenges make confident data-use challenging, both per-tumour and across the entire dataset.

Aim 2: Quantify biases in pathway-level analyses
After taking into account the per-site biases, we will try to understand how they will bias pathway-level conclusions. To do this, we will take the pathway-analysis protocols in-use by that working-group and will simulate the effect of bias on them. This will quantify the bias that can be expected, and may indicate the exclusion of specific pathways in specific tumour-types or across the entire study.

Aim 3: Quantify biases for individual non-coding genomic elements
One of the central goals of the pan-cancer analysis will be to identify putatively-functional ncSNVs. However associating such ncSNVs with functional annotation such as from ENCODE or based on TFBS profiles will be confounded with biases in the underlying sequencing data (e.g. GC-content biases). We will again use simulation to try to quantify the effects and to identify specific regions that cannot have robust conclusions made in either specific tumour-types or more globally.

Legacy plans

All coding will be done in Vms in collaboration with the variant-calling subgroups.

Paul C. Boutros

Ontario Institute for Cancer Research
Tel: (416) 673-8564
Paul.Boutros@oicr.on.ca

1. Education

University of Toronto (2004-2008)
PhD, Department of Medical Biophysics
Integrated Molecular Prediction of Patient Prognosis
Supervisors: Dr. Linda Z. Penn & Dr. Igor Jurisica
University of Waterloo (2000-2004)
B.Sc, Chemistry, Honours Co-operative Education

2. Employment (Current)

Assistant Professor (11/2012 to Present)
Department of Pharmacology and Toxicology, University of Toronto
Assistant Professor (11/2011 to Present)
Department of Medical Biophysics, University of Toronto
Principal Investigator (11/2010 to Present)
Biocomputing Platform, Ontario Institute of Cancer Research

3. ICGC Contributions

1. Bioinformatics Team Lead, Canadian Prostate Cancer Genome Network
2. Lead PI, The ICGC-TCGA DREAM Somatic Mutation Calling Challenge
3. Member ICGC Bioinformatics AWG
4. Member Disease Working Group, Esophageal TCGA
5. Member Disease Working Group, Head & Neck TCGA
6. Member Analysis Working Group, Prostate TCGA

4. Selected Submitted ICGC/TCGA Manuscripts

The Cancer Genome Atlas Consortium (2013) "Comprehensive genomic characterization of head and neck squamous cell carcinomas"

Boutros PC *et al.* (2013) "A Molecular Portrait of Gleason 7 Prostate Cancer"

Lalonde *et al.* (2013) "Heterogeneity Analyses Using Genomics and the Tumour Microenvironment Predict Prostate cancer Recurrence"

Govind *et al.* (2013) "ShatterProof: operational detection and quantification of chromothripsis"

Chong *et al.* (2013) "SeqControl: Optimizing DNA Sequencing Experiments By Modelling Library Complexity Using Quality Metrics"

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Sex-Associated Differences in Somatic Mutational Profiles

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Dr. Paul C. Boutros, Principal Investigator, Informatics & Biocomputing, Ontario Institute for Cancer Research; Canadian Prostate Cancer Genome Network (CPC-GENE: Canadian Prostate ICGC)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

Differences in cancer between the sexes are well-known. These range from differences in incidence rates -- even after best efforts to control for confounding-factors such as smoking status and other life-style considerations -- to differences in outcome after treatment. In some cases, these are already known to reflect underlying differences in the somatic mutation profile of the tumour. Male:female incidence rate ratios exceed 1.3 for many common cancers (colorectal, lung, lymphoma, bladder).

Consider, as an example of sex-differences in cancer, non-small cell lung cancer -- the most common subtype. Sex differences in this tumour-type are many. First, histological type is sex-biased, with men most often diagnosed with squamous cell carcinomas and women most commonly diagnosed with adenocarcinomas. Second, women are more prone to constitutive activation of K-Ras. Third, women have substantially better survival than men, including enhanced response to chemotherapy and radiation. Fourth, FDG-PET imaging has shown that sugar metabolism differs between women and men. Sixth, high levels of the most abundant human estrogen, 17-beta-estradiol, are a strong negative prognostic factor in women.

To understand the molecular origins of this phenomenon we compared prognostic mRNA markers (that is, associations of individual genes with overall survival) in 1184 patients (542 female; 629 male). After controlling for confounding clinical covariates, we found that over 80% of genes were only associated with survival in either men or women, but not both. There is, therefore, strong clinical and molecular rationales to explore sex-associated differences in somatic mutation profiles.

There are thus wide-spread sex-differences in cancer. This proposal aims to understand at a broad level the underlying genomic differences between tumours in men and tumours in women.

Timelines & resources dedicated to project

One FTE analyst is in-place, and we are currently looking to recruit a graduate-student to also work on these analyses. This project is, of course, dependent on the presence of data for the non-hormonal tumour types (*i.e.* it would not use prostate, ovarian, cervix or breast tumours), but would leverage both exomes and whole-genomes.

Several aspects of this work will depend on the sub-groups. We would use variant-calls made by the core. We would mimic the pathway-analysis methods used by the core (and decided in the AWG papers, which we are participating in). Most importantly, we would want to focus on the core set of driver-mutations for some analyses, and would want to incorporate the mutation-signatures as used by the core groups.

The key intermediate milestone would be identification of sex-specific mutations and pathway-level alterations.

Research proposal

We will analyze all tumours that are not of the sex-organs (e.g. no prostate, breast, cervix, or ovarian tumours will be included) to study sex-specific differences. Analyses will be done first independently, by tumour-type, and then as a pool to identify tumour-type-independent drivers. Our analyses will fall into three specific aims:

Aim 1: sex-specificity of driver mutations
For each of these analyses we plan to employ the tools being used by the core mutation-calling group, but if those are not available (*e.g.* custom code) we will use a combination of MuTect for SNVs, GISTIC for CNAs and standard statistics for genomic-rearrangements. We may, however, need to develop multivariate versions of these statistical techniques to control each of our analyses for differences in clinical presentation (*e.g.* higher average stage of male tumours).
1A) We will look at sex-differences in the rate of mutations (particularly ncSNVs, but of all types) identified as pan-cancer common and identify differential rates, both in individual tumour types and between them
1B) We will analyze each tumour type separately and all tumour-types together to identify sex-specific drivers.

Aim 2: sex-specificity of mutational trends
Next, we will look at genome-wide trends and compare them using standard methods, using pathway methods devised by the AWG and standard statistical approaches.
2A) rates of genomic aberrations: number of CNAs, rearrangements, fusion proteins, SNVs
2B) pathway differences between male & female tumours
2C) differences in mutational signatures between male & female tumours

Aim 3: aberrations in sex-associated genes/pathways
Lastly we will select a set of pathways and genomic already associated with sex-differences.
3A) We will confirm and extend reported differences in mutation rates/types in the X and Y chromosomes, and particularly in the inactive X chromosomes
3B) we will investigate differences in estrogen and androgen signaling pathways, particularly the fraction of EREs and AREs with ncSNVs or other somatic disruptions and alterations in downstream target genes

Legacy plans

All analyzed data will be made available, of course. This work is anticipated to be done in the shared cloud-resource, so we expect to be producing a VM with a single perl-script that launches all analysis jobs and creates all visualizations. That should create a completely replicable environment. We do not use Rstudio or knitr, but do create .tex documents directly as the output of our analysis pipelines, so that should allow ready independent replication. We anticipate working with the DCC team to try to create a web-searchable interface do our results database so that users can easily query sex-specific and sex-independent results.

Paul C. Boutros

Ontario Institute for Cancer Research
Tel: (416) 673-8564
Paul.Boutros@oicr.on.ca

1. Education

University of Toronto (2004-2008)
PhD, Department of Medical Biophysics
Integrated Molecular Prediction of Patient Prognosis
Supervisors: Dr. Linda Z. Penn & Dr. Igor Jurisica
University of Waterloo (2000-2004)
B.Sc, Chemistry, Honours Co-operative Education

2. Employment (Current)

Assistant Professor (11/2012 to Present)
Department of Pharmacology and Toxicology, University of Toronto
Assistant Professor (11/2011 to Present)
Department of Medical Biophysics, University of Toronto
Principal Investigator (11/2010 to Present)
Biocomputing Platform, Ontario Institute of Cancer Research

3. ICGC Contributions

7. Bioinformatics Team Lead, Canadian Prostate Cancer Genome Network
8. Lead PI, The ICGC-TCGA DREAM Somatic Mutation Calling Challenge
9. Member ICGC Bioinformatics AWG
10. Member Disease Working Group, Esophageal TCGA
11. Member Disease Working Group, Head & Neck TCGA
12. Member Analysis Working Group, Prostate TCGA

4. Selected Submitted ICGC/TCGA Manuscripts

The Cancer Genome Atlas Consortium (2013) "Comprehensive genomic characterization of head and neck squamous cell carcinomas"

Boutros PC *et al.* (2013) "A Molecular Portrait of Gleason 7 Prostate Cancer"

Lalonde *et al.* (2013) "Heterogeneity Analyses Using Genomics and the Tumour Microenvironment Predict Prostate cancer Recurrence"

Govind *et al.* (2013) "ShatterProof: operational detection and quantification of chromothripsis"

Chong *et al.* (2013) "SeqControl: Optimizing DNA Sequencing Experiments By Modelling Library Complexity Using Quality Metrics"

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Determinants of Genome Stability and Instability

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Dr. Paul C. Boutros, Principal Investigator, Informatics & Biocomputing, Ontario Institute for Cancer Research; Canadian Prostate Cancer Genome Network (CPC-GENE: Canadian Prostate ICGC)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Dr. Kenneth Chu	

Background and preliminary data

Cancer is a disease of genomic alterations, and much of the PanCan studies will be directed at understanding specific alterations – where they are, how they influence tumour progression, and what is the interplay amongst aberrations of different types. However one of the most fundamental hallmarks of cancer is decreased genomic stability. The overall ability of a cell to repair mutations declines, and the total mutation load increases. We propose here to study this phenomenon in-depth.

Within our ICGC project, we started such studies in the context of a single tumour-type: prostate cancer. In work currently under review (Lalonde *et al*. Submitted), we have shown that genomic stability at the level of double-strand breaks (as estimated by copy-number aberrations count) is an independent prognostic variable. It predicts patient outcome with high fidelity (Hazard Ratio = ~4 in two separate cohorts). Further, it was predictive of metastases and was associated with morphological differences in tumour grade. We were able to identify specific genes whose mutation status was associated with increased genomic instability. Finally, we extended this work to breast cancer, using both the TCGA and META-BRIC cohorts. We found again an association between genomic stability and outcome, as well as genes predictive of it. Importantly a subset of these genes were common between breast and prostate cancers, suggesting that there are tumour-type specific and tumour-type independent determinants of genomic stability.

We propose to extend this work, identifying determinants of genomic stability between and across tumour-types. We are also ready to perform targeted validation (e.g. using cell-line models) depending on the findings made.

Timelines & resources dedicated to project

This project has 1 FTE of a research associate already in-place working on these analyses on TCGA data, along with contributions from the PhD student who performed the original studies and a biostatistician. All are fully-funded and in-place for this analysis.

We do not anticipate any early-stage milestones during this project, but it will require the finalized mutation-calls for each tumour-type along with the per-tumour pathway-activation values as required inputs.

Research proposal

We will look at genomic instability at three levels, and for each make the same set of associations. The levels at which we will assess genomic stability are:

- point-mutations
- putative double-stranded breaks (CNAs + Rearrangements; either the sum or separately)
- total mutational load (number of genes altered at any level; this may be generated by a core or would be created ourselves using techniques like those in TCGA papers and in our own ICGC project)

To ensure independent validation of the associations we see, we will generally use WGS data to "learn" associations and exome data to validate them. Rearrangements will be derived from TCGA low-pass WGS where possible. Additionally we will try to exploit non-ICGC datasets (e.g. META-BRIC) for validation.

The analyses we will do for each metric of genome-instability are:

- identify specific genes whose mutation is correlated with genomic stability
- identify specific pathways whose activation level is correlated with genomic stability

These two analyses will both be done using MANCOVA-type mixed-models, allowing adjustment for key covariates. In particular, we will include tumour-type as a covariate in each model to help ensure our analysis is not dominated by differences between tumour types. We will also run our modeling separately for each tumour-type to help identify genes that are associated with stability for only a subset of tumour types.

Experimental validation may be attempted on a limit subset of hits by over-expressing mutant forms and assaying changes in double-stranded breaks (via sequencing or cell-level imaging of gamma-H2AX positive foci) in cell-line models.

Legacy plans

All analyzed data will be made available, of course. This work is anticipated to be done in the shared cloud-resource, so we expect to be producing a VM with a single perl-script that launches all analysis jobs and creates all visualizations. That should create a completely replicable environment. We do not use Rstudio or knitr, but do create .tex documents directly as the output of our analysis pipelines, so that should allow ready independent replication.

Paul C. Boutros

Ontario Institute for Cancer Research
Tel: (416) 673-8564
Paul.Boutros@oicr.on.ca

---

1. Education

University of Toronto (2004-2008)
PhD, Department of Medical Biophysics
Integrated Molecular Prediction of Patient Prognosis
Supervisors: Dr. Linda Z. Penn & Dr. Igor Jurisica
University of Waterloo (2000-2004)
B.Sc, Chemistry, Honours Co-operative Education

2. Employment (Current)

Assistant Professor (11/2012 to Present)
Department of Pharmacology and Toxicology, University of Toronto
Assistant Professor (11/2011 to Present)
Department of Medical Biophysics, University of Toronto
Principal Investigator (11/2010 to Present)
Biocomputing Platform, Ontario Institute of Cancer Research

3. ICGC Contributions

13. Bioinformatics Team Lead, Canadian Prostate Cancer Genome Network
14. Lead PI, The ICGC-TCGA DREAM Somatic Mutation Calling Challenge
15. Member ICGC Bioinformatics AWG
16. Member Disease Working Group, Esophageal TCGA
17. Member Disease Working Group, Head & Neck TCGA
18. Member Analysis Working Group, Prostate TCGA

4. Selected Submitted ICGC/TCGA Manuscripts

The Cancer Genome Atlas Consortium (2013) "Comprehensive genomic characterization of head and neck squamous cell carcinomas"

Boutros PC *et al.* (2013) "A Molecular Portrait of Gleason 7 Prostate Cancer"

Lalonde *et al.* (2013) "Heterogeneity Analyses Using Genomics and the Tumour Microenvironment Predict Prostate cancer Recurrence"

Govind *et al.* (2013) "ShatterProof: operational detection and quantification of chromothripsis"

Chong *et al.* (2013) "SeqControl: Optimizing DNA Sequencing Experiments By Modelling Library Complexity Using Quality Metrics"

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 31st December, 2013

Title of abstract

The landscape of viral associations across human cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

Vincent Ferretti, Ontario Institute for Cancer Research (OICR), Canada

Name(s) & institute(s) of junior investigators	Name(s) & institute(s) of non-ICGC collaborators
Ivan Borozan, (OICR) **Stuart Watt**, (OICR)	**Philip Branton**, McGill University, Canada **Jacques Archambault**, Institut de Recherche Clinique de Montreal (ICRM), Canada

Background and preliminary data

Specific viruses have been proved to be etiologic agents of human cancer and cause 15% to 20% of all human tumors worldwide (zur Hausen. H. 2006). Moreover, epidemiological studies indicate that new oncogenic pathogens are yet to be discovered (Javier, RT, Butel, JS. Cancer Res 2008). In this context the ICGC pan-cancer dataset will enable us to address the following five questions:

Q1. Which viral species are present in the Pan-Can cancer tumor genomes, transcriptomes and their matched control samples?

Q2. What is the scale of viral integration and their location in the host genomes?

Q3. Which of the viral species found are specifically associated with individual cancer types and in which proportions?

Q4. To which extent viral integration has an impact on transcription and methylation?

Q5. To which extent can we identify entirely new viruses?

The work will be based on our CaPSID platform for identifying viral genomes in NGS data (Borozan et al. BMC Bioinformatics. 2012). CaPSID's results have been validated both in vitro and in silico, and show high precision and sensitivity even on substantially divergent viral sequences with up to 15% overall sequence mutation rate (Borozan et al. PLOS ONE. 2013). Data generated by the CaPSID pipeline is stored in its scalable MongoDB database that greatly facilitates retrieval, organization and analysis using pluggable external tools. The scale and nature of the Pan-Can datasets (matched tumor-control, different cancer types, in addition to different data types within the same cancer type) will enable CaPSID to establish powerful and previously unavailable results for Q1-Q3.

Our prior work applying CaPSID to more than 300 tumor genome and transcriptome samples from different cancer types has enabled us to establish and validate filtering criteria capable of effective discrimination between the viral signal present in the sample and the background noise (e.g. sequencing artifacts, genomes with no obvious relation to cancer phenotype, viral sequences present in the human reference, etc.) CaPSID's web interface also enables us to review the results with our panel of expert virologists. We also have a preliminary implementation of statistical analysis tools to identify those viruses that are significantly associated with a particular cancer type or group.

CaPSID also enables reads that do not map to known viral sequences in our database or the human reference genome to be analyzed using a de novo viral characterization pipeline developed in our laboratory. Additional approaches based on information theory (currently under development in our laboratory) will be also used to answer Q5. We note that Q4 and Q5 are more exploratory in nature compared to Q1-Q3.

Timelines & resources dedicated to project

Key milestones. (Subject to amendment based on project start)

M1. January 2013. Deploy cloud-enabled CaPSID pipeline and sharded database in PanCan data centre.

M2. May 2013. Complete analysis of selected cancer types. Mid-project meeting to review/update plan.

M3. November 2013. Primary analysis complete. Review meeting to define additional post-hoc analysis.

Analyses and data sets.

Initial work (to M2) will focus on cancer types with better-known viral risks, e.g., cervical, head-and-neck, liver. To M3, data sets will include all WGS tumour/normal pair data, RNAseq/cDNA microarray data, methylation data, and clinical data. The additional exome and cancer genome data will not be required.

Research proposal

There are three phases of work in this proposal.

Phase 1. Initially, the project will move CaPSID to the PanCan data centre, scaled to the data sizes and types required by the full data. This is a brief task, requiring only a small amount of data of each type, sufficient to ensure processes are valid. This concludes at milestone M1.

Phase 2. Next we propose to use complete data sets in four types of cancer, focusing on those with strong evidence of viral impact, e.g., cervical, head-and-neck, to about 10% of the full data set size (200 samples plus controls). During this stage, we will refine CaPSID's statistical and analytic systems to assess evidence of viral involvement (as specified in Q1, Q2 and Q5), in close collaboration with two teams of virologists headed by Prof. Philip Branton (McGill) and Dr. Jacques Archambault (IRCM). We will also chose one cancer type with strong evidence of viral integration in host genomes (conditional on the availability of RNA-Seq and methylation data) to address Q4. This enables us to refine the processes addressing Q1-Q5. This concludes with a mid-project review meeting at M2.

Phase 3. The final phase extends this to the full data set. Further development of analysis techniques might be required at this point to address Q3 and Q4. The computational analysis will be well-proven at this stage, and we will move on to address Q3 and Q4. To enable this, we will begin viral sequence alignments for the WGS data as it becomes available in the data centre. As reported in (Borozan et al. 2013) this is 6-10x faster than aligning against a human genome, therefore computationally reasonable in the PanCan data centre.

Phases 2 and 3 will follow a similar pattern. Using the CaPSID pipeline's digital subtraction techniques, four metrics and visualization tools will be used to find and characterize evidence of pathogens in the WGS and RNA-seq data. Samples will be statistically ranked to identify the likely viral genomes and their relevance to cancer, grouped by the NCBI taxonomic tree. These will be filtered to remove viral genomes that relate to sequencing artifacts or have no obvious relation to cancer phenotypes, based on a set of criteria including CaPSID's metrics, genome relative abundance, taxonomic information and genome annotations. Furthermore RNA-Seq data will be used to validate the WGS results. Post-hoc analysis will assess possible associations between viral evidence and clinical data.

These candidate viral sequences will then be used to guide the search for viral integration sites in tumor genomes using the pipeline described in VirusFinder (Wang et al. PLOS ONE. 2013). Viral integration sites will then be compared to human gene boundaries (annotated by the NCBI RefSeq database) to find genes that are directly disrupted or those that are potentially affected by viral integration (within 15kb of integration sites). Time permitting, we will also assess the positional bias of viral integration sites in respect to sequence features of the human genome such as repeat families, recombination hot spots and segmental duplications.

As with past work using CaPSID, reads that do not map to viral sequences or the human reference genome will be analyzed using our de novo viral characterization pipeline, which assembles unmapped reads into contigs, filters them for low sequence complexity and known sequence content, and then scans (using InterproScan, Quevillon et al. Nucleic Acids Res 2005) for the presence of known protein features. Finally the statistical analysis across different cancer types will be performed using data stored in the CaPSID database in combination with statistical tools and analysis techniques developed in R and Python.

The effect of viral integration on transcription (RNA-Seq data) will be determined by comparing the expression profiles, in the vicinity of viral integration sites, between tumors (or non tumors) with inserted viral genomes to those with unaltered genomes (tumor and non-tumor). Changes in DNA methylation in the vicinity of viral integration sites and methylated genes, reported by (Leonard et al. Carcinogenesis. 2012), will also be assessed between tumors (or non tumors) with inserted viral genomes to those with unaltered genomes (tumor and non-tumor).

Legacy plans

Additional public outputs.

- Open source CaPSID cloud-compatible pipeline platform, including deployment system for a sharded database capable of handling the dataset load (at M1).
- Refinements to CaPSID's web-based analysis environment will be incorporated through the project. Revisions will be delivered according to an agile roadmap determined by the team including the viral expert panel.
- Where appropriate, independent pipelines for viral and statistical analysis will be released independently of the core CaPSID software platform.

All software is published at Github (see: https://github.com/capsid), and future revisions will continue to be released there.

Ferretti, Vincent **Curriculum Vitae**

Current Employment
Associate Director, Bioinformatics Software Development & Senior Principal Investigator
Ontario Institute for Cancer Research (OICR), vincent.ferretti@oicr.on.ca

Education

Mathematical Research Center, University of Montreal, Canada	Postdoctoral	1995-1997	Computational biology
University of Montreal, Canada	Ph. D.	1989- 1994	Mathematics
University of Montreal, Canada	M. Sc.	1986-1989	Mathematics
University of Montreal, Canada	B. Sc	1983-1986	Mathematical Physics

Employment

Senior Principal Investigator and Senior Scientist	OICR	2013-Present
Associate Director, Bioinformatics Software Development	OICR	2011-Present
Principal Investigator and Senior Scientist	OICR	2008-2013
Adjunct Professor, School of Computer Science	McGill University	2005-2008
Chief Bioinformatician	Genome Quebec	2002-2008
Director of Bioinformatics	PhageTech Inc.	1998-2002
Director of Bioinformatics	Algene Inc.	1997-1998

Leadership

Leader of the ICGC DCC software infrastructure development and the Data Portal	2011-Present
Member of the ICGC Scientific Steering Committee	2011-Present
Member of the Governance Board of BioSHaRE	2011-Present
Co-Leader of the Maelstrom Research program (maelstrom-research.org)	2012-Present
Leader of the OBiBa project in collaboration with the Public Population Project in Genomics (P3G).	2007-Present

Honours

Honorary Research Fellow in the Genomic Epidemiology Research Group within the School of Translational Medicine, University of Manchester.	2007-2010

Current Grants

Title	Funding Source	Support Period
ICGC DCC	OICR	04-2013 to 05-2017
OICR Principal Investigator Awards	OICR	05-2013 to 05-2018
BBMRI-Large Prospective Cohorts	European Union FP7	09-2012 to 09-2016
Plateforme d'Harmonisation Québec-Europe	Quebec Ministry of Innovation and Economic Development	01-2012 to 01-2014
Biobank Standardisation and Harmonisation for Research Excellence in the European Union: BioSHaRE-EU	European Union FP7	01-2010 to 01-2015

Recent Publications (last author)

- Borozan I, et al. *Evaluation of Alignment Algorithms for Discovery and Identification of Pathogens Using RNA-Seq*. **PLoS ONE**. Oct 30 2013.

- Doiron D, et al. *Data harmonization and federated analysis of population-based studies: the BioSHaRE projec*t. **Emerging Themes in Epidemiology**. 2013, 10:12.

- Watt S, et al. *Clinical genomics information management software linking cancer genome sequence and clinical decisions.* **Genomics**. Sept 2013

- Borozan I, at al. CaPSID: *A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes*. **BMC Bioinformatics**. Aug 2012

Borozan, Ivan **Curriculum Vitae**

PRIMARY AFFILIATION AND CURRENT EMPLOYMENT
Scientific Associate, Informatics and Biocomputing, Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, Ontario, Canada, M5G0A3
Email: stuart.watt@oicr.on.ca

Education

Ph. D. 1997 Open University, Milton Keynes, UK, Cognitive Science (Psychology)
B. Sc. 1985 University of York, UK, Honours Degree in Computer Science

Recent Positions and Employment

1995 – 2002 Lecturer, Knowledge Media Institute and Psychology, The Open University, Milton Keynes, UK.
2002 – 2007 Reader and Director of Research, School of Computing Science and Digital Media, The Robert Gordon University, Aberdeen, UK.
2007 – 2011 Senior Product Developer, Information Balance, Toronto, Canada.
2011 – Present Scientific Associate, Informatics and Biocomputing, Ontario Institute for Cancer Research, Toronto, Canada.

Selected Honors and Awards

Summer 1997 British Telecom Short Research Fellowship, Martlesham, UK.
2004 – 2006 Open Mentor: text classification for tutor training ($80k), JISC, UK.
2006 – 2007 Open Comment: classification for formative feedback in the arts ($70k), JISC, UK.

Selected Publications

1. Clark, M., Ruthven, I., Holt, P. O'B., Song, D., & Watt, S. (2014, in press) You have e-mail, what happens next? Tracking the eyes for genre. Information Processing & Management, 50(1):175–198, doi:10.1016/j.ipm.2013.08.005.
2. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. PLoS ONE. Oct 2013; 8(10): e76935. doi:10.1371/journal.pone.0076935
3. Watt, S., Jiao, W., Brown, A. M., Petrocelli, T., Tran, B., Zhang, T., McPherson, J. D., Kamel-Reid, S., Bedard, P. L., Onetto, N., Hudson, T. J., Dancey, J., Siu, L. L., Stein, L., Ferretti, V. (2013) Clinical genomics information management software linking cancer genome sequence and clinical decisions. Genomics, Sept 2013, 102(3), doi:10.1016/j.ygeno.2013.04.007.
4. Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, Stein LD, Ferretti V, Watt S, Jiao W, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE. International Journal of Cancer. April 2013, 132(7), doi :10.1002/ijc.27817.
5. Borozan I, Wilson S, Blanchette P, Laflamme P, Watt SN, Kryzanowski P, Sircoulomb F, Rottapel R, Branton PE, Ferretti V. CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinformatics. 2012 Aug 17;13:206. doi: 10.1186/1471-2105-13-206.
6. Watt, S. (2009). Text categorisation and genre in information retrieval. In A. Goker & J. Davies (Eds.) Information Retrieval, John Wiley & Sons.
7. Watt, S. (2009). Can People Think? Or Machines. Invited chapter : R. Epstein, G. Roberts & G. Beber, Parsing the Turing Test, Springer.
8. Chakraborti, S., Mukras, R., Lothian, R. M., Wiratunga, R., Watt, S., & Harper, D. J. (2007). Supervised latent semantic indexing using adaptive sprinkling. International Joint Conference on Artificial Intelligence, IJCAI'07, pp. 1582-1587.

Watt, Stuart **Curriculum Vitae**

PRIMARY AFFILIATION AND CURRENT EMPLOYMENT

Scientific Associate, Informatics and Biocomputing, Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, Ontario, Canada, M5G0A3
Email: stuart.watt@oicr.on.ca

Education

Ph. D. 1997 Open University, Milton Keynes, UK, Cognitive Science (Psychology)
B. Sc. 1985 University of York, UK, Honours Degree in Computer Science

Recent Positions and Employment

1995 – 2002 Lecturer, Knowledge Media Institute and Psychology, The Open University, Milton Keynes, UK.
2002 – 2007 Reader and Director of Research, School of Computing Science and Digital Media, The Robert Gordon University, Aberdeen, UK.
2007 – 2011 Senior Product Developer, Information Balance, Toronto, Canada.
2011 – Present Scientific Associate, Informatics and Biocomputing, Ontario Institute for Cancer Research, Toronto, Canada.

Selected Honors and Awards

Summer 1997 British Telecom Short Research Fellowship, Martlesham, UK.
2004 – 2006 Open Mentor: text classification for tutor training ($80k), JISC, UK.
2006 – 2007 Open Comment: classification for formative feedback in the arts ($70k), JISC, UK.

Selected Publications

9. Clark, M., Ruthven, I., Holt, P. O'B., Song, D., & Watt, S. (2014, in press) You have e-mail, what happens next? Tracking the eyes for genre. Information Processing & Management, 50(1):175–198, doi:10.1016/j.ipm.2013.08.005.
10. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. PLoS ONE. Oct 2013; 8(10): e76935. doi:10.1371/journal.pone.0076935
11. Watt, S., Jiao, W., Brown, A. M., Petrocelli, T., Tran, B., Zhang, T., McPherson, J. D., Kamel-Reid, S., Bedard, P. L., Onetto, N., Hudson, T. J., Dancey, J., Siu, L. L., Stein, L., Ferretti, V. (2013) Clinical genomics information management software linking cancer genome sequence and clinical decisions. Genomics, Sept 2013, 102(3), doi:10.1016/j.ygeno.2013.04.007.
12. Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, Stein LD, Ferretti V, Watt S, Jiao W, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE. International Journal of Cancer. April 2013, 132(7), doi :10.1002/ijc.27817.
13. Borozan I, Wilson S, Blanchette P, Laflamme P, Watt SN, Kryzanowski P, Sircoulomb F, Rottapel R, Branton PE, Ferretti V. CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinformatics. 2012 Aug 17;13:206. doi: 10.1186/1471-2105-13-206.
14. Watt, S. (2009). Text categorisation and genre in information retrieval. In A. Goker & J. Davies (Eds.) Information Retrieval, John Wiley & Sons.
15. Watt, S. (2009). Can People Think? Or Machines. Invited chapter : R. Epstein, G. Roberts & G. Beber, Parsing the Turing Test, Springer.
16. Chakraborti, S., Mukras, R., Lothian, R. M., Wiratunga, R., Watt, S., & Harper, D. J. (2007). Supervised latent semantic indexing using adaptive sprinkling. International Joint Conference on Artificial Intelligence, IJCAI'07, pp. 1582-1587.

Branton, Philip Edward **Curriculum Vitae**

McGill University, Department of Biochemistry 1-514-398-8350 philip.branton@mcgill.ca

UNIVERSITY EDUCATION

1966 BSc, UofT, Dept. of Microbiology. 1966-1968 MSc, UofT, Dept. of Microbiology. 1968-1972 PhD, UofT, Dept. of Medical Biophysics. 1972-1974 PDF M.I.T.

STAFF POSITIONS

1974-1975	Université de Sherbrooke, Assistant Professor, Dépt. de biologie cellulaire
1975-1990	McMaster University, Assistant/Associate/Full Professor, Pathology
1987-1990	Coordinator, NCIC Viral Oncology Group, McMaster University
1990-present	Professor, McGill University, Dept. of Biochemistry
1990-2000	Chair, McGill University, Dept. of Biochemistry
1994-present	Professor, Dept. of Oncology, McGill University
2000-2008	Scientific Director, Institute of Cancer Research (CIHR)
2001-present	Member, Goodman Cancer Research Centre, McGill University

AWARDS

1996-present	Gilman Cheney Professor of Biochemistry, McGill University
2002-present	Member, Academy of Science of the Royal Society of Canada
2005	R.M. Taylor Medal, Canadian Cancer Society and the NCIC
2011	CCRA Award for Exceptional Leadership in Cancer Research
2013	Queen Elizabeth II Diamond Jubilee Medal

CURRENT OPERATING GRANTS

2010-2015	CIHR ($221,455/yr) E4orf6/E1B55K E3 ubiquitin ligase in the adenovirus life cycle
2009-2014	CIHR ($145,894/yr) Mechanism of action of the adenovirus death protein E4orf4

PANELS, BOARDS AND COUNCILS (Recent)

2006-present	Chair, International Advisory Board, University Hospital Network, U of T
2008-present	Chair, Scientific Advisory Committee, Terry Fox Research Institute
2010-present	Member, Advisory Board, Canadian Tumour Repository Network
2010-2012	Member, Review Board, SIRIC, INCa, Paris, France

RECENT PUBLICATIONS

119 BRANTON, P.E. and R. C. Marcellus. 2011. Adenoviruses, In "Fundamentals of Molecular Virology" (2nd edition). Ed. N. Acheson. John Wiley and Sons, New York, NY. pp. 274-284.

124 BOROZAN, I., S. Wilson, P. Blanchette, P. Laflamme, S.N. Watt, P.M. Krzyzanowski, F. Sircoulombe, R. Rottapel, P.E. Branton and V. Ferretti. 2012. CaPSID: A bioinformatics platform for computational pathogen sequence identification in **human genomes and transcriptomes. BMC Bioinformatics 13, 206 (on-line 13, 206).**

125 BLANCHETTE, P., P. Wimmer, F. Dallaire, C.-Y. Cheng, P.E. Branton, P.E. 2013. Aggresome formation by the E1B55K product is not conserved among different adenovirus serotypes and is not required for ligase mediated degradation of substrates. J. Virol. 87, 4872-4881.

126 CHENG, C.Y., T. Gilson, P. Wimmer, G. Ketner, T. Dobner, P.E. Branton and P. Blanchette. 2013. The role of E1B55K in E4orf6/E1B55K E3 ligase complexes formed by different human adenovirus serotypes. J. Virol. 87, 6232-6245

127 CABON, L., N. Sriskandarajah, M. Mui, J. Teodoro, P. Blanchette and P.E. Branton. 2013. The adenovirus E4orf4 protein induces G1 arrest and death of tetraploid and diploid $p53^{-/-}$ human cancer cells. J. Virol. 87, 13168 (on-line)

128 MUI, M., A.M. M. Kucharski, M.J. Miron, F. Dallaire, W. Hur, A. Berghuis, P. Blanchette and P.E. Branton. 2013. Identification of the Adenovirus E4orf4 Protein Binding Site on the B55α and Cdc55 Regulatory Subunits of PP2A: Implications for PP2A Function, Tumor Cell Killing and Viral Replication. PLoS Pathogens 9(11): e1003742, doi:10.1371 (on-line)

Archambault, Jacques **Curriculum Vitae**

PRIMARY AFFILIATION AND CURRENT EMPLOYMENT

Director, Molecular Virology Laboratory and Full IRCM Research Professor
Institut de recherches cliniques de Montréal (IRCM), CANADA, Jacques.Archambault@ircm.qc.ca

EDUCATION

Université de Montréal, Montreal, Canada	B.Sc.	1981-1984	Biochemistry
University of Toronto, Toronto, Canada (Mentor: James D. Friesen)	Ph.D.	1984-1991	Molecular and Medical Genetics
University of Toronto, Toronto, Canada Banting and Best Dept. of Medical Research (Mentor: Jack F. Greenblatt)	Postdoctoral	1992-1996	RNA polymerase II transcription

POSITIONS AND EMPLOYMENT

1996-1999	Research Scientist (HPV antiviral research program), Biological Sciences, Boehringer Ingelheim Ltd., Canada
1999-2000	Senior Research Scientist, Coordinator - HPV antiviral research program, Biological Sciences, Boehringer Ingelheim Ltd., Canada
2000-2002	Group Leader, Coordinator - HPV antiviral research program, Biological Sciences, Boehringer Ingelheim Ltd., Canada
2002-2003	Group Leader, Coordinator - HPV and HIV antiviral research programs, Biological Sciences, Boehringer Ingelheim Ltd., Canada
2003	Group Leader, Coordinator - HIV research program, Biological Sciences, Boehringer Ingelheim Ltd., Canada
2003-2012	Director, Molecular Virology Unit, and Associate Research Professor, IRCM, Canada
2005-2013	Research Associate Professor, Dept. of Biochemistry, Université de Montréal, Canada
2005-present	Adjunct Professor, Dept. of Medicine (Division of Experimental Medicine), McGill University, Montréal, Canada
2005-present	Associate Researcher, Institut de recherche en immunologie et cancérologie (IRIC), Université de Montréal, Canada
2009-present	Adjunct Professor, Dept. of Microbiology and Immunology, McGill University, Montréal, Canada
2010-present	Scientific Advisor, IRCM Chemical Biology Resource Platform, Montréal, Canada
2012-present	Director, Molecular Virology Unit, and Full IRCM Research Professor, Montréal, Canada
2013-present	Full Research Professor, Biochemistry and Molecular Medicine Dept., Université de Montréal

HONORS

1984-1990	Studentship, Medical Research Council of Canada (MRC), Canada
1992-1995	Postdoctoral Fellowship, Medical Research Council of Canada (MRC), Canada
2005-2008	Senior Research Scholarship, Fonds de la Recherche en Santé du Québec (FRSQ)
2006-2016	IRCM Chair (Molecular Virology), Canada

SCIENTIFIC ADVISORY BOARD

2004-present	*Anaconda Pharma*, "Institut Pasteur", Paris. Anaconda Pharma is dedicated to the discovery and development of small-molecule HPV antiviral drugs. (www.anacondapharma.com)

ASSOCIATE EDITOR

2013-present	The Open Virology Journal (TOVJ)

INTERNATIONAL MEETINGS ORGANIZATION

2010	*Organizing committee,* 26th International Papillomavirus Conference and Clinical Workshop, Palais des Congrès, Montréal, Qc, Canada.
2012	*Organizing committee,* DNA Tumor Virus Meeting 2012, Montréal, Qc, Canada.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Classification and comparison of somatic genome rearrangements across 2,000 whole-genome cancer samples

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

John McPherson, OICR

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
	Ravi Pandya, Microsoft Research Ma'ayan Bresler, Microsoft Research / UC Berkeley

Background and preliminary data

Many cancers involve significant genomic rearrangements, varying from specific chromosome fusions such as BCR-ABL, to chains of correlated rearrangements (chromoplexy), to widespread genomic fragmentation (chromothripsis). Reconstructing the topology of these genomic rearrangements from short-read next-generation sequencing data remains a computational and algorithmic challenge. Current methods show little concordance between their results, and can take significant computational resources to analyze modern datasets.

The collaborators have developed algorithms for analyzing structural variations that offer useful improvements in speed and accuracy. Preliminary results indicate that they can analyze a whole-genome dataset in about half a day on a single server. The algorithms evaluate the relative likelihood of alternative rearrangements using a probabilistic model that integrates information such as sequence similarity, coverage depth, and mate pair distances to select the most probable rearrangement. These algorithms have shown good results (unpublished) on regions with significant genomic duplication, repetition, and similarity.

Timelines & resources dedicated to project

Q1 2014: Refining methods for somatic structural variation analysis; evaluation against DREAM somatic mutation challenge, TCGA Benchmarks, and OICR pancreatic/prostate cancer data

Q2 2014: Evaluation against a broader range of ICGC/TCGA whole-genome samples; improving speed, accuracy, and handling of sequencing errors across a range of datasets

Q3 2014: Analysis of complete 2,000-genome dataset; integrating and organizing results

Q4 2014: Statistical and machine learning analysis of results; classifying and comparing rearrangements within and across tumor types; investigating underlying biological mechanisms

Q1-Q2 2015: Final analysis and publication

Microsoft will provide computing and storage resources required for this analysis

Research proposal

The first phase of the project involves extending the algorithms to analyze somatic rather than germline structural variation, as well as handling issues specific to tumor genomes such as high coverage depth, tumor/normal comparison, sample purity, and clonal evolution. The algorithms will be evaluated against a small number of high-quality datasets, such as the DREAM somatic mutation challenge, the TCGA benchmark data, and specific high-quality datasets such as pancreatic and prostate cancer data from OICR.

The second phase involves scaling the algorithm to a range of data from different tumor types, of varying sample quality and process protocols. This will require developing methods for measuring and incorporating these variations, and metrics for evaluating the quality of the results in the absence of specific validation data. It will also require engineering to ensure that the algorithms perform efficiently and scale well across the range of data available in ICGC.

Finally, we will run the structural variation analysis across the entire 2,000 whole-genome dataset, and analyze the results using statistics, graph theory, and machine learning. This will involve a range of interesting questions, such as:

- What are the distinct patterns or sub-patterns of structural variation?
- How do these patterns relate to distinct tumor types?
- Are there common patterns of genomic alteration and tumor evolution that could have generated these patterns?
- Do these patterns correlate with other classifications, such as network analysis?
- Are these patterns indicative of clinical outcomes such as survival time and drug response?

Legacy plans

The methods used in this analysis will be documented and published.

The structural variation analysis software will be released as open source for use within the research community, for general-purpose analysis of germline and somatic structural variations in whole-genome data.

SOFTWARE ARCHITECT — **MICROSOFT, REDMOND**

eSCIENCE RESEARCH GROUP *Mar 2012-Present*

- Co-developer of SNAP fast, scalable short-read genome sequence aligner
- Developed algorithms for high-performance IO, structural variation, de novo assembly
- Researching network models for cancer systems biology

CLOUD COMPUTING RESEARCH GROUP *Oct 2009-Mar 2012*

- Architect & technical leader for the Orleans distributed actor framework for client+cloud computing
- Built distributed optimistic transactions, persistence, fault tolerance; deployed to millions of users per day
- Built applications including a distributed graph database, Bayesian gene regulatory network analysis

TECHNICAL STRATEGY INCUBATION *Jun 2006-Oct 2009*

- Founding member of clean-slate operating system project reporting directly into Bill Gates & Steve Ballmer
- Built team of 15 stellar engineers, implemented core messaging & application model subsystems

WINDOWS SECURITY *Mar 2003-Jun 2006*

- Architect for Rights Management Services for Office enterprise users, and Windows Anti-Piracy technology
- On small team of senior architects advising Bill Gates & executives on company-wide security strategy

CONSULTANT — **COVALENT INDUSTRIAL TECHNOLOGIES, BURLINGAME**

- Built extensible molecular modeling workbench for 2-D and 3-D polymer design *Dec 2000-Mar 2003*

CHIEF TECHNOLOGY OFFICER — **EVERYTHINGOFFICE, SAN FRANCISCO**

- Co-founded company & raised venture capital for web procurement system *Oct 1997-Nov 2000*
- Designed innovative technology combining Java app server, Oracle Financials, EDI, wireless PalmPilots

VP OF ENGINEERING & SERVER OPERATIONS — **JANGO, SEATTLE**

- Took shopping agent from AI research to award-winning product & successful acquisition *Nov 1996-Sep 1997*
- Built strong engineering team, raised venture capital, managed board relationships & industry partnerships

DIRECTOR OF ENGINEERING — **NETMANAGE, BELLEVUE**

- Shipped several releases of award-winning ECCO personal & group information manager *Jul 1993-Oct 1996*

CHIEF ARCHITECT — **XANADU OPERATING COMPANY, PALO ALTO**

- Co-architect of a pioneering hypertext system that inspired the World Wide Web *Aug 1989-Jun 1993*
- Innovations included fine-grained linking, history tracking, versioning, access control, link filtering, indexing

MANAGER OF SOFTWARE ENGINEERING — **HYPERCUBE, WATERLOO**

- Developed HyperChem, the first molecular modeling system for Windows *May 1988-Apr 1989*
- Designed UI for molecular editing with constraint-based model builder and high-performance 3D graphics
- Integrated clusters of Inmos Transputers or Intel HyperCubes for scalable high-performance computing

RESEARCH INTERN — **XEROX PALO ALTO RESEARCH CENTER, PALO ALTO**

- Worked with creators of Smalltalk, the seminal object-oriented language & graphical UI *May-Aug 1984&85*
- Developed an object server, and a logic/constraint-based graphical programming environment

B.SC. MATHEMATICS SPECIALIST — **TRINITY COLLEGE, UNIVERSITY OF TORONTO**

- Mathematical foundations: algebra, analysis, complexity, geometry, optimization, statistics *Apr 1989*
- First year at University of Waterloo (on full scholarship, top 20 in Canada on Descartes math competition)

Ma'ayan Bresler

University of California, Berkeley
EECS Department
627 Soda Hall, Berkeley, CA 94720-1776
(217) 721-4106
mbresler@eecs.berkeley.edu

Research Interests

Computational biology, Machine Learning / Data Mining, Bayesian models and statistics, Algorithms for analysis of next-generation sequencing data.

Education

University of California, Berkeley, California, USA
Ph.D. candidate, Electrical Engineering and Computer Sciences, August 2007–Present
Advisor: Yun S. Song

Princeton University, Princeton, New Jersey, USA
B.A., Physics, June 2006
Certificate in Program for Applied and Computational Mathematics
Senior thesis advisor: Mung Chiang

Graduate Courses

Statistical Learning Theory, Theoretical Statistics, Bayesian Modeling and Inference, Genome Project Lab, Randomness and Computation, Topology and Analysis, Probability Theory I and II, Information Theory and Coding, Linear System Theory, Random Processes and Systems.

Publications

Ameet Talwalkar, Jesse Liptrap, Julie Newcomb, Christopher Hartl, Jonathan Terhorst, Kristal Curtis, Ma'ayan Bresler, Yun S. Song, Michael I. Jordan, David Patterson: SMASH: A Benchmarking Toolkit for Variant Calling. arXiv:1310.8420. 2013.

Guy Bresler, Ma'ayan Bresler, David Tse: Optimal assembly for high throughput shotgun sequencing. BMC Bioinformatics. 2013.

Ma'ayan Bresler, Sara Sheehan, Andrew H. Chan, Yun S. Song: Telescoper: de novo assembly of highly repetitive regions. Bioinformatics. 2012.

Ma'ayan Bresler, Koby Sheffy, Giora Pillar, Meir Preiszler, Sarah Herscovici: Differentiating between light and deep sleep stages using an Ambulatory Device Based on Peripheral Arterial Tonometry. Physiol Meas. 2008.

Jiayue He, Martin Suchara, Ma'ayan Bresler, Jennifer Rexford, Mung Chiang: Rethinking internet traffic management: from multiple decompositions to a practical protocol. CoNEXT. 2007.

Jiayue He, Ma'ayan Bresler, Mung Chiang, Jennifer Rexford: Towards Robust Multi-Layer Traffic Engineering: Optimization of Congestion Control and Routing. IEEE Journal on Selected Areas in Communications. 2007.

Honors and Awards

National Science Foundation Graduate Research Fellowship 2009–2012
Kusaka Memorial Award in Physics 2006

Teaching Experience

Graduate Student Instructor. Berkeley EECS, EE 40: circuits. Fall 2007–Spring 08.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Somatic variant detection using assembly graphs

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Jared Simpson, Ontario Institute for Cancer Research

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Most mutation detection approaches rely on mapping reads to a reference genome and calling variants where consistent differences between the reads and reference genome are found. This reference-based approach is fast and effective for small differences, like simple point mutations, in regions of the genome that otherwise are very similar to the reference. However, as the complexity of the mutation increases, for instance when sequence is inserted or deleted, it becomes increasingly difficult to find the mutation. In addition, reads may not be aligned correctly to the reference genome where there are polymorphic indels, repeats or copy number differences between the sequenced individual and the reference genome. Reads placed incorrectly on the reference genome are the primary source of false positive variant calls.

To address this limitation we have developed variant calling methods based on direct comparison of read sets in an assembly graph. For somatic variant detection in cancer, we construct an assembly graph that consists of sequences present in the tumour and matched normal sample. Structures in the graph that are unique to the tumour indicate possible somatic variation. We have developed a set of algorithms to find and assemble these structures.

Our approach, implemented in the SGA assembler, is part of the core variant calling pipeline for the 1000 Genomes project. As a pilot for the pan-cancer project, we have called somatic mutations on 13 Pancreatic tumour/normal pairs. We found 3000-5000 SNVs and 250-2000 indels per genome. The average time required to go from raw sequence data to somatic variant calls was 250 CPU hours per genome. The average memory high-water mark was 105GB.

Timelines & resources dedicated to project

As each genome is processed independently, this pipeline can be run as data is uploaded to the pan-cancer cloud. I propose three milestones for this project:

- M1: Somatic calls for 100 tumour/normal pairs
- M2: Somatic calls for 500 tumour/normal pairs
- M3: Somatic calls for 2000 tumour/normal pairs

At each milestone the variant calls will be assessed and compared to somatic call sets from parallel projects. At this time algorithmic improvements or parameter adjustments may be made to improve call set quality or reduce compute time.

The resources dedicated to this project are the lead PI (Jared Simpson) at 50% time. A software engineer may be hired to assist with pipeline engineering.

Research proposal

The basic principle of the sga somatic variant caller is direct comparison of tumour reads against matched normal reads using a de Bruijn graph. A de Bruijn graph is defined over the set of *k*-mers (subsequences of length *k*) present in a collection of reads. To detect somatic mutations, we construct a de Bruijn graph from the union of the *k*-mers present in the tumour and the *k*-mers present in the matched normal. Each *k*-mer in the graph is marked to indicate whether it came from the tumour, the matched normal or both. The *k*-mers that belong only to the tumour reads contain candidate variants. Our algorithm searches for such *k*-mers then locally explores the de Bruijn graph to assemble the variant *k*-mers into haplotypes. These candidate haplotypes are aligned to the reference and variants are recorded. The evidence for each variant is evaluated in a Bayesian framework by realigning the tumour and normal reads to the candidate haplotypes. The final output is a VCF file that contains variants that are strongly supported by the tumour reads only.

Assembly graphs for human genomes can contain billions of vertices, so memory usage is a primary concern. We represent the sequence reads using an FM-index, which is a compressed data structure. The use of a compressed data structure limits memory usage while still allowing efficient string queries to be performed against the read collection. The FM-index for a 40X human genome data set requires approximately 32GB of memory.

The analysis pipeline begins by pre-processing the raw sequence reads to quality trim and filter reads using their quality scores. An FM-index is created for the tumour reads and an FM-index is created for the matched normal reads. These steps require around 50 CPU hours and 40GB of memory per 40X genome. The third step of the pipeline is to read the pair of index files, along with a reference genome, and generate somatic variant calls.

The quality of the generated call sets will be continually evaluated throughout the project but in particular at each milestone stated above. This will include both computational analysis (for example swapping the labels on the tumour and matched normal to estimate a false positive rate for a given sample) and experimental validation as part of the larger ICGC technical working group efforts and DREAM competition.

Legacy plans

Our software implementation is open-source and freely available on github. As part of the pan-cancer project we will develop a set of best practices for somatic variant detection and documentation for use by the wider cancer research community. The FM-index data structures that are constructed will be retained to facilitate post-production querying of the raw sequence data as need should arise.

Jared Simpson

E-mail: jared.simpson@oicr.on.ca Phone: (416) 471-1257

EDUCATION

University of Cambridge SEPTEMBER 2008 – SEPTEMBER 2012
Cambridge, UK

Wellcome Trust Sanger Institute
Doctor of Philosophy

University of British Columbia SEPTEMBER 1999 – MAY 2004
Vancouver, Canada

Bachelor of Science, Computer Science

PROFESSIONAL EXPERIENCE

Fellow MAY 2013 –PRESENT
Ontario Institute for Cancer Research
Toronto, Canada

Consultant JANUARY 2012 – JANUARY 2013
Moleculo, Inc
San Francisco, USA

Computational Biologist APRIL 2007 – SEPTEMBER 2008
Genome Sciences Centre
Vancouver, Canada

Software Engineer JUNE 2004 – JUNE 2006
Electronic Arts Canada
Burnaby, Canada

ACADEMIC SERVICE

Program Committee Member - RECOMB-Seq 2012, IEEE HiCOMB 2010
Reviewer – Nature Methods, Nucleic Acids Research, Bioinformatics, BMC Bioinformatics

PUBLICATIONS

List available online: http://scholar.google.com/citations?user=jcbxkKwAAAAJ&hl=en

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

A Classifier for Pan-Cancer Tumor Type

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Lincoln D. Stein, Ontario Institute for Cancer Research, member, ICGC bioinformatics working group

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Wei Jiao, Ontario Institute for Cancer Research	

Background and preliminary data

The recent application of high throughput genomic sequencing techniques to the study of cancer has effected a paradigm shift in recent years from one in which each cancer type is thought of as a distinct biological entity to a recognition of substantial overlaps among the tumor types at the molecular level. For example, basal-like breast cancers are more similar at the molecular level to high-grade serous ovarian tumours than they are to other types of breast cancer [Cancer Genome Atlas Network *Nature* **490:**61 2012], while bladder cancers, head and neck cancers, and lung tumours co-cluster with many other different tumour types, indicating extensive heterogeneity in those diseases [Kandoth C *et al. Nature* **502:**333 2013].

We anticipate that at least one Pan-Cancer analysis group will propose to perform unsupervised clustering of tumours based on patterns of somatic variation in order to identify groups of related subtypes. This project will take the opposite approach. Using supervised learning approaches, we will create a classifier that uses mutation data to predict which cancer type a tumour is derived from. Reframing the problem in this way has certain advantages. First, it addresses the feasibility of creating a diagnostic tool capable of identifying "tumours of unknown origin," the clinical conundrum that occurs when a patient presents with a distant metastasis but no obvious primary source. Second, by examining the rules learned by the classifier over the course of training, we can determine which mutations carry the strongest signal of type-specificity and which the least. This will provide a window into the pathways that most strongly distinguish one tumour type from another.

It may be that mutations of individual genes and regulatory regions will be insufficient to create a robust classifier due to low overall recurrence rates. For this reason, we will investigate whether the classifier performs better when presented with features corresponding to predicted changes in higher-level biological pathways. This method, which has shown promising results in reliably distinguishing a variety of tumours, developmental stages and disease processes [Altschuler GM *et al Genome Med* **5:**68 2013], will build upon the work described in the linked abstract *Metabolic and Regulatory Network Rewiring in WGS Pan-Cancer Data Sets* (PIs Gerstein and Stein).

The Stein lab has extensive experience in cancer genome analysis [Biankin A *et al. Nature* **491:**399-405, Peltekova VD *et al. Int J Cancer* Oct 23 20113, Sawey ET *et al Cancer Cell* **8:**347 2011], network analysis of normal and abnormal genomes [Altschuler *op cit*], [Wu G *et al Genome Biol* **11:**R53 2010], and machine learning in cancer data sets [Wu G and Stein L *Genome Biol.* **13:**R112 2012],. Very recently, we have applied machine learning techniques (logistic regression, random forests) to the ICGC somatic mutation-calling benchmarking exercise, and show that a combination of the outputs of three of the SNV callers can reduce the misclassification rate of somatic mutations by half.

Timelines & resources dedicated to project

We will dedicate one research associate (Wei Jiao) to the project for a period of six to nine months. The computational load needed for this project is relatively small, allowing us to use the ICGC cloud compute resources without impacting other research projects; if necessary we can use the 8,000-core OICR compute cluster for the needed computations.

Research proposal

We will develop and evaluate tumour type classifiers built with a variety of supervised machine learning algorithms. We will start by assembling training and testing sets from the 2000 ICGC and TCGA tumour pairs that have whole genome sequencing available, using ground truth based on the histological classification of the tumours. Only tumour types that have at least 50 exemplars will be used. As the project progresses, other PanCancer research groups that are performing unsupervised clustering of tumour types may encounter outliers and unexpected clusterings that differ from the histological classification; when this occurs, we will adjust the training sets to either exclude or relabel these tumours.

The greatest challenge to the project will be selecting and testing the features that provide the greatest power to distinguish one tumour type from another. Broadly speaking, we will investigate the following types of feature:

- *Overall mutation rates.* We will use the overall rate of mutations of various types, including single nucleotide substitutions, copy number changes and structural variations, as the first set of features for training.
- *Nucleotide substitution frequencies*. These will be taken from the frequency distributions derived by other working groups (e.g. via non-negative matrix factorization), and reflect the nature of the processes causing DNA damage in the tumour as well as the state of DNA repair pathways in the tumour cells.
- *The presence of high impact mutations in each putative cancer driver gene*. The list of driver genes will be derived from those produced by the current Pan-Cancer project as well as the earlier exome-based TCGA project.
- *The presence of mutations in conserved non-coding regions.* These will be taken from the results of the groups looking at the distribution of functional non-coding mutations, such as the project described in the Gerstein and Stein abstract.
- *The presence of characteristics structural rearrangements.* These features will be derived from recurrent structural rearrangements, such as those affecting particular chromosome arms.
- *Affected pathways.* The list of pathways altered in each tumour will be derived from the network modeling project described in Gerstein and Stein.

We will initially train and test classifiers based on random forests, naïve bayes, and logistic regression; we have chosen to start with these three systems because they are simple to implement and are relatively robust. Their disadvantage is that they require binning of continuous value features such as mutation rate. If initial results are unsatisfactory, we will evaluate machine learning systems that accept continuous values such as SVMs. We will use a ten-fold cross-validation strategy to evaluate the performance of each classifier, and keep those that provide the best performance (as measured by area under the ROC curve, AUC), and stability across multiple training runs.

Provided that we obtain a satisfactory tumour type classifier, we will examine the weights of each of the features that contribute to the model in order to identify those with the greatest influence. We will then relate these features to the biological properties of the tumours.

Legacy plans

The classifier(s) and all new software related to them will be released under an Open Source license that allows for unrestricted use and redistribution.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Lincoln Stein	Director, Bio-computing Platform and Senior Principal Investigator, Ontario Institute for Cancer Research
eRA COMMONS USER NAME LINCOLNSTEIN	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Johns Hopkins University	B.A.	1978-1982	Biology
Harvard Medical School, Boston, MA	M.D.	1982-1989	Medicine
Harvard University	Ph.D.	1982-1989	Cell Biology
Brigham & Women's Hospital	Residency	1989-1992	Anatomic Pathology

A. Positions

1995-1997	Instructor, Harvard Medical School, Dept. of Pathology, Brigham and Women's Hospital
1992-1997	Director, Informatics Core, MIT Genome Center, Whitehead Institute of Biomedical Research, M.I.T., Cambridge, MA
1997	Director Information Systems, CuraGen Corporation, New Haven, CT
1998-2004	Associate Professor, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2004-present	Professor, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2007-present	Director, Informatics and Biological Computing Platform & Senior Principal Investigator, Ontario Institute for Cancer Research, ON
2009-present	Professor, Department of Molecular Genetics, University of Toronto, Ontario

B. 10 Selected Peer-Reviewed Publications (From 141 peer-reviewed publications total)

1. Trinh QM, Jen FY, Zhou Z, Chu KM, Perry, MD, Kephart E, Contrino S, Ruzanov P, and **Stein LD**, Cloud-Based Uniform ChIP-Seq Processing Tools for modENCODE and ENCODE. BMC Genomics, 1. 2013 14:494.
2. Watt S, Jiao W, Brown AM, Petrocelli T, Tran B, Zhang T, McPherson JD, Kamel-Reid S, Bedard PL, Onetto N, Hudson TJ, Dancey J, Siu LL, **Stein L**, Ferretti V. Clinical genomics information management software linking cancer genome sequence and clinical decisions. Genomics. 2013 Apr 17.
3. Wang L, **Stein LD**. Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process. BMC Evol Biol. 2013 Feb 28;13:57. doi: 10.1186/1471-2148-13-57. PubMed PMID: 23448166.
4. Wu G, **Stein L**. A network module-based method for identifying cancer prognostic signatures. Genome Biol. 2012 Dec 10;13(12):R112.
5. Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, **Stein LD**, Ferretti V, Watt S, Jiao W, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE. Feasibility of real time next generation sequencing of cancer genes linked to drug response: Results from a clinical trial. *Int J Cancer*. 2012 Sep 5
6. Haw R, Hermjakob H, D'Eustachio P, **Stein L.** Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics*. 2011 Sep; 11(18):3598-613.
7. Feng X, Grossman R, **Stein L**. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*. 2011 May 9;12:139.
8. Wu G, Feng X, **Stein L.** A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*. 2010;11(5):R53.
9. Gerstein MB, *et al*, **Stein L**, Lieb JD, Waterston RH. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*. 2010 Dec 24;330(6012):1775-87.
10. modENCODE Consortium, *et al*, **Stein LD**, White KP, Kellis M. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*. 2010 Dec 24;330(6012):1787-97.

WEI JIAO

200 Beverley St. • Toronto, ON • M5T 1Z1
Tel: (647) 922-8216 • Email: wei.jiao@mail.utoronto.ca

Profile

I have a thorough training in both engineering and life sciences. During the last three years, I further enhanced my skills in computational biology. With an open minded and easygoing attitude, I believe my research experience would greatly help me to fulfill my duty as a research associate.

Education background

September 2010 – November 2013. M. Sc. University of Toronto
September 2008 - 2009 M. Res. University of Glasgow
September 2002 - July 2006 B. Eng. BioEngineering, Xi'an Jiaotong University

Knowledge Structures

Molecular genetics: Proficient
Machine learning and statistical modeling: Proficient
Programming and Database: Proficient

Research Experience

April 2012 – November. 2013 Thesis project
Morris Lab, Stein Lab, University of Toronto
In collaboration with another Student and Post Doc in the lab, we are working together to model the clonal evolution of cancer by an advanced Bayesian statistical method. Detailed work were summarized in [1].

January 2011 – March 2012 Thesis project
Morris Lab, Stein Lab, University of Toronto
Attempted to build up a number of statistical models to training on RNAi screen data set, and tried to make predictions about essential genes in unscreened cancer cell lines.

September 2010 – December 2010 Rotation projects
Rommens Lab, Huhges Lab, Stein Lab, University of Toronto
I was exposed to a couple of research projects that include database implementation, and high throughput biological data analysis. My part of the work was included in the following publications [2-5].

June 2009 – August 2010 Research projects
Morris Lab, Lipshitz Lab, Boon Lab, University of Toronto
I was involved in a number of research projects in the area of computational biology across multiple model organisms. The topics includes gene interaction network, post-transcriptional gene regulation. Part of the work was published in [5].

Technical Skills

Programming Language: Java, C++, SQL, and XHTML, PERL
Statistic Tools : R, Matlab
Operating Systems: UNIX, Windows
Microsoft Office suites

Publications

[1] **Jiao, W**, Vembu, S., Deshwar, A., Stein, L. and Morris, Q. (2013). Modeling the clonal evolution of cancer from next generation sequncing data. Under review.

[2] Watt, S., **Jiao, W.**, Brown, A., Petrocelli, T. Tran, B. Zhang, T., McPherson, J., Kamel-Reid, S., Bedard, P., Onetto, N., Hudson, T. Dancey, J., Siu, L., Stein, L., Ferretti. V. (2013). Clinical genomics information management software linking cancer genome sequence and clinical decisions. **Genomics**

[3] Watt, S., **Jiao, W.**, Brown, A., Petrocelli, T. Tran, B. Zhang, T., Dancey, J., Siu, L., Stein, L., Ferretti. V. (2012). Designing a web application for personalized medicine trials. **Journal of Clinical Oncology.**

[4] Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, Stein LD, Ferretti V, Watt S, **Jiao W**, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE. (2012). Feasibility of real time next generation sequencing of cancer genes linked to drug response: Results from a clinical trial. **International Journal of Cancer.**

[5] Magtanong, L., Ho, C. H., Barker, S. L., **Jiao, W**., Baryshnikova, A., Bahr, S., et al. (2011). Dosage suppression genetic interaction networks enhance functional wiring diagrams of the cell. **Nature biotechnology**,

E-learning and Certificates

Introduction to Machine Learning – Coursera Certificate
Probabilistic Graphical Model – Coursera Certificate

Interests

Sports take most of my spare time. I enjoy in participating and organizing sports activities, from which I have developed strong communication skills.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Inferring Subclonal Evolution from Primary Tumours

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Lincoln D. Stein, Ontario Institute for Cancer Research, member, ICGC bioinformatics working group

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Shankar Vembu, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto Amit G. Deshwar, Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto	Quaid Morris, Donnelly Centre for Cellular and Biomolecular Research and Departments of Molecular Genetics/Computer Science/Electrical and Computer Engineering, University of Toronto

Background and preliminary data

High-throughput sequencing allows the detection and quantification of frequencies of somatic single nucleotide variants (SNV) in heterogeneous tumor cell populations. In some cases, the evolution history and population frequency of the subclonal lineages of tumor cells present in the sample can be reconstructed simply from the SNV frequency measurements and we have recently developed a new statistical model to reconstruct these subclonal evolutionary structures from these frequencies. Our model, *PhyloSub* (in second review after minor revision), uses a Bayesian nonparametric prior over trees that groups SNVs into major subclonal lineages. PhyloSub automatically estimates the number of lineages and their ancestry. We sample from the joint posterior distribution over trees to identify evolutionary histories and cell population frequencies that have the highest probability of generating the observed SNV frequency data. When multiple phylogenies are consistent with a given set of SNV frequencies, PhyloSub explicitly represents the uncertainty in the exact phylogeny.

The output of the PhyloSub algorithm is a set of possible phylogenic reconstructions that are consistent with the measured SNV frequencies. Each reconstruction consists of i) a discrete clustering of SNVs into those associated with each subclonal lineage; ii) a tree describing the phylogenic relationships of the subclonal lineages; iii) cellular frequencies for each lineage. Each reconstruction is also associated with a posterior probability which is a measure of how consistent it is with the sequencing data. This set is generated through a Markov Chain Monte Carlo procedure from which we derive samples from the posterior distribution over reconstructions based on our generative model of the SNV frequency data. In some cases a set of SNV frequencies are only consistent with a handful of highly similar phylogenies; allowing a high resolution in phylogeny reconstruction. However, even where there is substantial posterior uncertainty in the correct reconstruction, useful information can still be extracted from this set, such as the posterior probability that i) that the phylogeny branches, i.e., that there are multiple independent, expanding subclonal populations; or ii) that an SNV appears in a lineage that is ancestral to that of another SNVs; or iii) a specific set of SNVs is present within at least one subclonal lineage (or whether they occur on different branches of the phylogeny).

Experiments on a simulated dataset and two real datasets comprising tumor samples from acute myeloid leukemia and chronic lymphocytic leukemia patients have demonstrated the efficacy of PhyloSub. PhyloSub can successfully infer not only simple tree structures like chains, but also tree structures with branching from single and multiple tumor samples. PhyloSub-inferred phylogenies are consistent with ground truth, where available, for these datasets. In one case, we were able to correctly infer three separate subclonal lineages from SNV frequencies from a single tumor sample.

Timelines & resources dedicated to project

The project is dependent on accurate simple somatic mutations (SSMs) and copy number calls, so our analysis will start after alignment and primary variant calling is complete. However, we will begin assessing our algorithm on simulated 50x coverage WGS data immediately while, in parallel, making modifications to allow it use copy number calls as additional sources of information on subclonal evolutionary structure. We will devote a total of 1 FTE (full time equivalent) postdoctoral fellow to and one graduate student to this project for a period of six months.

Research proposal

Extending the PhyloSub model to take SSM frequencies instead of SNV frequencies as input is trivial. The main computational challenge this project faces is that the average genome coverage in the PanCancer study is 50X and is much shallower than the coverage in the data sets we have previously worked on (>1000X). This will greatly decrease the accuracy of measured SSM frequencies; which ultimately limits the resolution of phylogenic reconstruction. Counterbalancing this, there will be many more SSM events in whole genomes than in the exome and targeted-sequencing data sets that PhyloSub was initially tested against. We conjecture that the increased numbers of SSMs associated with each lineage will compensate for the lower accuracy in their estimated frequencies and will recover the lost resolution. We also conjecture that we can further recover resolution using subclonal lineage information available in copy number variations (Oesper et al, Genome Biology 2013; PMID: 23895164). Incorporating larger numbers of SSMs will require no modifications to PhyloSub; incorporating data on copy number changes will require a minor modification of the algorithm which we expect will take less than three months of effort.

We will test these conjectures in synthetic data sets where the ground truth is known. The synthetic data sets will be designed to mimic the sequencing characteristics of primary cancers in the ICGC/TCGA PanCancer set. The modified algorithm will then be tested against a single-cell pancreatic cancer sequencing set in development at OICR (Lincoln Stein, personal communication) where there is matched flow-sorted bulk tumor sequence available as well **that provides ground truth for the multiple variant subclonal genotypes and population frequencies.**

Regardless of the amount of resolution that we can recover by extending PhyloSub, the algorithm can still provide useful information about the evolutionary history of the tumor, as indicated in the background section. Furthermore, an advantage of using a Bayesian framework and sampling from the posterior over phylogenies rather than simply taking the most probable reconstruction is that the decreased accuracy in SSM frequencies does not increase the chance of a highly incorrect phylogenic reconstruction, rather it simply increases the uncertainty in the reconstruction and makes it more likely that subclonal lineages with low frequency will be merged with their parental or daughter lineages.

We will run the modified PhyloSub software against the entire ICGC/TCGA PanCancer data set, to provide a comprehensive view into the subclonal heterogeneity and evolutionary patterns of multiple tumor types. By mapping putative driver mutations (derived by other working groups) onto the derived subclonal evolutionary trees, we should be able to distinguish, in many cases, early events from later ones, and infer which mutations resulted in a selective advantage for the subclone in which it arose. Also, we can determine whether some tumor types have tendencies for deeper phylogenies, more branching, or more mutations per lineage.

The software, subclonal frequencies, and evolutionary trees will be deposited back into the PanCancer cloud for use by other working groups.

Legacy plans

All software and algorithms developed from this project will be open source using a BSD license selected based on discussions with the working group. The subclonal phylogenies derived by this project will be contributed to the PanCancer data set available to the public.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Lincoln Stein	Director, Bio-computing Platform and Senior Principal Investigator, Ontario Institute for Cancer Research
eRA COMMONS USER NAME LINCOLNSTEIN	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Johns Hopkins University	B.A.	1978-1982	Biology
Harvard Medical School, Boston, MA	M.D.	1982-1989	Medicine
Harvard University	Ph.D.	1982-1989	Cell Biology
Brigham & Women's Hospital	Residency	1989-1992	Anatomic Pathology

C. **Positions**

1995-1997	Instructor, Harvard Medical School, Dept. of Pathology, Brigham and Women's Hospital
1992-1997	Director, Informatics Core, MIT Genome Center, Whitehead Institute of Biomedical Research, M.I.T., Cambridge, MA
1997	Director Information Systems, CuraGen Corporation, New Haven, CT
1998-2004	Associate Professor, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2004-present	Professor, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2007-present	Director, Informatics and Biological Computing Platform & Senior Principal Investigator, Ontario Institute for Cancer Research, ON
2009-present	Professor, Department of Molecular Genetics, University of Toronto, Ontario

D. **10 Selected Peer-Reviewed Publications** (From 141 peer-reviewed publications total)

11. Trinh QM, Jen FY, Zhou Z, Chu KM, Perry, MD, Kephart E, Contrino S, Ruzanov P, and **Stein LD**, Cloud-Based Uniform ChIP-Seq Processing Tools for modENCODE and ENCODE. BMC Genomics, 1. 2013 14:494.
12. Watt S, Jiao W, Brown AM, Petrocelli T, Tran B, Zhang T, McPherson JD, Kamel-Reid S, Bedard PL, Onetto N, Hudson TJ, Dancey J, Siu LL, **Stein L**, Ferretti V. Clinical genomics information management software linking cancer genome sequence and clinical decisions. Genomics. 2013 Apr 17.
13. Wang L, **Stein LD**. Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process. BMC Evol Biol. 2013 Feb 28;13:57. doi: 10.1186/1471-2148-13-57. PubMed PMID: 23448166.
14. Wu G, **Stein L**. A network module-based method for identifying cancer prognostic signatures. Genome Biol. 2012 Dec 10;13(12):R112.
15. Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, **Stein LD**, Ferretti V, Watt S, Jiao W, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE. Feasibility of real time next generation sequencing of cancer genes linked to drug response: Results from a clinical trial. *Int J Cancer*. 2012 Sep 5
16. Haw R, Hermjakob H, D'Eustachio P, **Stein L.** Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics*. 2011 Sep; 11(18):3598-613.
17. Feng X, Grossman R, **Stein L**. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*. 2011 May 9;12:139.
18. Wu G, Feng X, **Stein L.** A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*. 2010;11(5):R53.
19. Gerstein MB, *et al*, **Stein L**, Lieb JD, Waterston RH. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*. 2010 Dec 24;330(6012):1775-87.
20. modENCODE Consortium, *et al*, **Stein LD**, White KP, Kellis M. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*. 2010 Dec 24;330(6012):1787-97.

Shankar Vembu
Post Doctoral Fellow in the Donnelly Centre
University of Toronto, Toronto, Ontario
Email: shankar.vembu@utoronto.ca

RESEARCH SUMMARY

- Hands-on experience (7+ years) in the design, analysis and applications of machine learning algorithms.
- Current research focus is at the intersection of computational biology and machine learning, specifically in phylogenetic inference, motif finding and gene function prediction.
- Published papers at highly selective machine learning conferences[1] and journals.

ACADEMIC BACKGROUND

Jan 2012 – Present Post Doctoral Researcher, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada
Oct 2009 – Dec 2011 Post Doctoral Researcher, BioCircuits Institute, University of California San Diego and Department of Computer Science, University of Illinois at Urbana-Champaign, USA
Oct 2006 – May 2010 Ph.D. in Computer Science, University of Bonn, Germany
Oct 2002 – Mar 2005 M.Sc. in Information and Communication Systems, Hamburg University of Technology, Germany

WORK EXPERIENCE

Apr 2005 – Sep 2006 Research Scientist, Knowledge Management Lab, German Research Center for Artificial Intelligence, Germany
Feb 2001 – Aug 2002 Software Engineer, Premier Technology Group Pvt. Ltd., India (affiliated with E-Z Data Inc., Pasadena, CA, USA)

SELECTED ACCOMPLISHMENTS AND HONORS

- One of the top nine machine learning papers, European Conference on Machine Learning, 2012 (2.9% acceptance rate)
- Natural Sciences and Engineering Research Council of Canada pre-approved Industrial R&D Fellow, 2012
- One of the top seven machine learning papers, European Conference on Machine Learning, 2009 (3.3% acceptance rate)

SELECTED, RELEVANT PUBLICATIONS (total of ten peer-reviewed papers since 2009)

1. **Shankar Vembu**, Quaid Morris. An Efficient Algorithm to Integrate Network and Attribute Data for Gene Function Prediction. In *PSB 2014*
2. Wei Jiao*, **Shankar Vembu***, Amit Deshwar, Lincoln Stein, Quaid Morris. Inferring Clonal Evolution of Tumors from Single Nucleotide Somatic Mutations. In *second review following minor revision*. http://arxiv.org/abs/1210.3384
3. Abhishek Kumar*, **Shankar Vembu***, Aditya Krishna Menon, Charles Elkan. Beam Search Algorithms for Multilabel Learning. *Machine Learning Journal*, 2(1):65–89, 2013. (**among top nine machine learning papers of the ECML-PKDD 2012 conference**; 13/443=2.9% acceptance rate)
4. Thomas Gärtner*, **Shankar Vembu***. On Structured Output Training: Hard Cases and an Efficient Alternative. *Machine Learning Journal* 76(2):227–242, 2009. (**among top seven machine learning papers of the ECML-PKDD 2009 conference**; 14/422=3.3% acceptance rate)

---

[1] In the field of machine learning (and computer science in general), top-tier conferences are highly selective and peer-reviewed full-length papers in their proceedings are more highly regarded than journal papers.

Amit Deshwar

Ph.D. Student
Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto, Toronto, Ontario, Canada
amit.deshwar@utoronto.ca

Education

2013-Current	Ph.D. Electrical and Computer Engineering – University of Toronto
2010-2013	M.A.Sc. Electrical and Computer Engineering – University of Toronto Thesis: Tumor Gene Expression Purification Using Infinite Mixture Topic Models
2004-2010	B.Sc. Software Engineering / B.A. Psychology – University of Calgary

Selected Awards

Vanier Canada Graduate Scholarship (2013-2016)
Ontario Graduate Scholarship (2011)
NSERC Julie Payette Scholarship (2010-2011)
APEGGA Past Presidents Award (2010)
NSERC Undergraduate Student Research Award (2006)

Selected Work Experience

June 2010 – Current	Chief Developer, Counterpartmatch.com
June 2009 – August 2009	Software Engineering Intern, Google Inc.
May 2007 – April 2008	Platforms Engineering Intern, Google Inc.

Publications

W. Jiao, S. Vembu, **A.G. Deshwar**, L. Stein, Q. Morris, Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics (in second review).
A.G. Deshwar, G. Quon, Q. Morris, ISOpure: computational tumor gene expression purification. Nucleic Acids Research (submitted).
A.G. Deshwar, Q. Morris. PLIDA: Cross-platform normalization using perturbed topic models. Bioinformatics 2013.
G. Quon, S. Haider, **A.G. Deshwar**, A. Cui, P.C. Boutros, Q. Morris, Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Medicine 2013.
A.M. Mezlini, B. Wang, **A.G. Deshwar**, Q. Morris, A. Goldenberg, Identifying Cancer Specific Functionally Relevant miRNAs from Gene Expression and miRNA-to-Gene Networks Using Regularized Regression. PLOS One 2013.

Quaid Morris,
***Associate Professor* in the Donnelly Centre and Banting and Best Department of Medical Research (BBDMR).**
Cross-appointments in Departments of Molecular Genetics, Computer Science, and Electrical and Computer Engineering (ECE).
Director of the Collaborative Program in Genome Biology and Bioinformatics
University of Toronto (U of T), Toronto, Ontario

ACADEMIC AND TRAINING BACKGROUND

2010-Present	*Associate Professor*, Faculties of Medicine, Science, and Engineering. U of T
2005-2010	*Assistant Professor*, Faculties of Medicine and Science, U of T
2002-2005	*Post Doctoral Fellow* in Ontario Cancer Institute, Princess Margaret Hospital and in Computational Biology, BBDMR and Dept of ECE, University of Toronto
1996-2002	*PhD* in Brain and Cognitive Sciences (specialization Computational Neuroscience), Massachusetts Institute of Technology (MIT), Cambridge, USA
1999-2001	Graduate training and research in machine learning, Gatsby Unit in University College London, England
1991-1996	*Honours B.Sc.:* Department of Computer Science, University of Toronto

RELEVANT ACCOMPLISHMENTS, HONOURS, AND CONTRIBUTIONS

Other accomplishments include: three keynotes; >50 invited talks or seminars since 2008; extensive reviewing for major biology journals (Nature, Nature Biotech/Methods/Genetics, MSB, Genome Biology and Genome Research); computational journals (PLOS CompBio, Bioinformatics); conferences (NIPS, ISMB, RECOMB); and organization of seven sessions on machine learning in biology for NIPS and PSB.

2010-2014	*Early Researcher Award*, Ontario Ministry of Research and Innovation project: *Computer-assisted cancer diagnosis and treatment planning.*
2013-present	*Associate editor*, PLOS Computational Biology
2008-present	*Grant panel member:* CIHR Genomics panel
Feb 2012	*Grant panel member*: Canadian Breast Cancer Foundation, Panel A
March 2011	*Keynote speaker*, RECOMB satellite on Computational Cancer Biology “
2010-2013	*Associate editor*, BMC Bioinformatics

SELECTED, PEER-REVIEWED PUBLICATIONS (from >50 since 2008)

* indicates co-first author; † indicates co-corresponding author; underline indicates QM’s trainee or staff.

1. Jiao W*, Vembu S*, Deshwar A, Stein L, **Morris Q**. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *(in second review after minor revisions, full text of revised manuscript is available here:* http://arxiv.org/abs/1210.3384*)*
2. Ray D*, Kazan H*, Cook KB*, Weirauch MT*, Najafabadi HS*, Li X, (+ 27 more authors), **Morris QD**†, Hughes TR†. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013 Jul 11
3. Zuberi K*, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, **Morris Q**. GeneMANIA prediction server 2013 update. Nucleic Acids Res. 2013 Jul
4. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, **Morris Q**. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. 2013 Mar 28.
5. Qiao W*, Quon G*, Csaszar E, Yu M, **Morris Q**†, Zandstra PW†. PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput Biol. 2012 Dec
6. G Quon, **Q Morris**, (2009) ISOLATE: A computational strategy for identifying the primary origin of cancers using high throughput sequencing. Bioinformatics Epub 2009 June 19
7. IW Taylor, R Linding, D Warde-Farley, Y Liu, C Pesquita, D Faria, S Bull, T Pawson, **Q Morris**, JL Wrana. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009 Feb 27

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5 pm your local time). Explanatory notes follow the form.

Title of abstract

Metabolic and Regulatory Network Rewiring in WGS Pan-Cancer Data Sets

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Lincoln Stein, Ontario Institute for Cancer Research, Member of ICGC Bioinformatics Analysis Working Group.
Mark Gerstein, Yale University, Member of TCGA Prostate Analysis Working Group.

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Guanming Wu, Oregon Health Sciences University
Lucas Lochovsky, Yale University

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Kevin White, University of Chicago
Mark Rubin, Cornell U.

Background and preliminary data

Cancer exerts its effects via perturbations of several key biological pathways that regulate cell division, differentiation, metabolism and interaction with the microenvironment. This project brings together two groups with complementary expertise in gene network analysis--one at the post-translational protein-protein interaction level, and the other at the transcriptional regulatory network--to perform a comprehensive analysis of the network perturbations that occur in cancer.

The Stein group works at the post-translational level, having built a curated pathway knowledgebase called Reactome, which is now one of the most comprehensive open source pathway databases containing over 7000 human genes and 1500 pathways. This group has built on top of this knowledgebase a functional interaction (FI) network by integrating human curated pathways from Reactome with several other pathway databases and protein pair-wise relationships using a machine learning technique (Wu 2010). The current version of the FI network contains around 11,000 human gene products (54% of SwissProt) in 274,000 functional interactions. Based on the FI network, the Stein group has developed a suite of software algorithms and tools to perform network based data analysis and identify network alterations that have clinical prognostic significance (Wu 2012). The Stein group has also recently adapted the factor-graph approach for predicting the effect of multiple gene alterations on pathway activities (Vaske 2012) to run in a high-throughput, high performance environment.

The Gerstein group has developed extensive tools and algorithms for constructing and testing transcriptional regulatory networks (Gerstein 2012, Cheng 2011a,b) and building protein interaction networks (Kim 2006, Jansen 2003). The group has performed numerous studies to assess the impact of genetic variation on network architecture (Khurana *Science* 2013,, Khurana *PLoS Comput Biol* 2013, Xia 2009, Setlur 2007, Kim 2007). Much of this work has been carried out in the framework of the ENCODE Consortia. The Gerstein group has developed multiple tools for assessing network topology, for example finding points of centrality such as hubs and bottlenecks (Shou 2011, Yip 2006, Yu 2007, Bhardwaj 2010). PCAP will provide a window into somatic variation occurring in both coding and transcriptional regulatory regions, allowing us to perform a comprehensive assessment of the network effects at both the transcriptional and post-transcriptional levels.

References

Bhardwaj, N. *et al.* Science Signaling 3, ra79 (2010).	Kim, P. M. *et al.* Science 314, 1938–1941 (2006).
Cheng, C. Bioinformatics, 27: 3221-7 (2011).	Setlur *et al.* Cancer Res 67: 10296-303 (2007).
Cheng, C. PLoS Comput Biol, 7: e1002190 (2011).	Shou, C. *et al.* PLoS Comput Biol 7, e1001050 (2011).
Gerstein, M. B. *et al.* Nature 489, 91–100 (2012).	Vaske *et al*. Bioinformatics. 26(12):i237-45 (2012)
Jansen, R. Science 302, 449–453 (2003).	Wu G *et al.* Genome Biol 11:R53 (2010)
Khurana, E. *et al.* Science 342, 1235587–1235587 (2013.	Wu G *et al.* Genome Biol 13(12):R112 (2012)
Khurana, E., *et al.* PLoS Comput Biol 9, e1002886 (2013).	Xia, Y. PLoS Comput Biol 5: e1000413 (2009).
Kim, P. M. *et al.* Mol Syst Biol 4, (2008).	Yip, K. Y. *et al.* Bioinformatics 22, 2968–2970 (2006).
Kim, P. M. *et al.* PNAS 104, 20274–20279 (2007).	Yu, H. *et al.* PLoS Comput Biol 3, e59 (2007).

Timelines & resources dedicated to project

December–March 2014: Integration of transcriptional regulatory network with post-transcriptional interaction network.

January–March 2014: Await identification of selected driver mutations affecting cis-regulatory sites (selected regulatory driver mutations, SRDMs) and coding sequences (selected coding mutations, SCMs) from the efforts of other pan-cancer projects.

April-May 2014: Separate exploration of transcriptional and post-transcriptional alterations in network topology to develop broad hypotheses.
June-August 2014: Integration of transcriptional and post-transcriptional alterations via clustering and Reactome pathway overlay analysis.
September-November 2014: Integration of transcriptional and post-transcriptional alterations via factor-graph analysis.
We will devote a total of 1.5 full time postdoctoral fellows and 0.5 other research staff to this project.

Research proposal

1. Combine the transcriptional and post-transcriptional networks. We will combine the transcription factor network derived from the ENCODE project with the Reactome FI network to produce a representation of both transcriptional and post-transcriptional gene interactions.
2. Identify of driver coding and non-coding mutations. We will develop a list of driver mutations that affect transcriptional cis-regulatory sites as well as genic coding regions. For the former, we will use the *funseq* module developed by the Gerstein group (this work will be carried out in the context of another abstract the Gerstein group is submitting). For the latter, we will rely on lists of coding driver mutations developed by other groups participating in this project. Both simple somatic mutations and larger structural mutations (CNVs, rearrangements) will be used for this list.
3. Assess transcriptional and post-transcriptional network effects independently. Mutations that affect cis-regulatory sites and transcription factors will be projected onto the transcriptional regulatory network in order to identify network edges that are removed or strengthened due to mutations. We will also look for cases in which a mutation creates a new transcription-factor binding site, creating new edges in the regulatory network. Likewise, we will project mutations that are predicted to knock out or activate the coding regions of post-transcriptionally active proteins onto the functional interaction network. From the perturbations we observe, we will assess broad patterns of mutational effects in terms of various graph theoretic measures. In particular we will look at how hubs in the networks change, how the hierarchical organization of the factors changes, and how bottlenecks are subtly tweaked. We will also look at how the wiring of key oncogenes and tumour suppressors are changed.
4. Assess transcriptional and post-transcriptional network effects jointly. We will use driver mutations in cis-regulatory sites and coding regions to extract affected subnetworks from the integrated network created in (1), using one or two steps of network propagation based on random walks (Hofree et al, Mat Methods, 10:1108-15). We will then perform spectral partitioning on these subnetworks (Newman, Proc Natl Acad Sci USA, 2006, 103:8577-82) to identify groups of highly-interacting genes that are recurrently altered. These clusters will be annotated by overrepresentation among Reactome pathways. The resulting maps are expected to provide insights into recurrent patterns of pathway alteration, including mutual inclusion and exclusion, that are associated with the disease, as well as to distinguish pathways that are generally altered in all tumour types from those that are tumour type specific.
5. Integrate the effects of transcriptional and post-transcriptional mutations. Using the factor-graph approach (Vaske *et al.* Bioinformatics. 26(12):i237-45) we will perform an integration of each donor's set of mutations across the combined network. This will reduce each set of mutations to a small set of inferred pathway activity alterations. From this information we will cluster donors within and between tumour types in order to identify tumour subtypes and cross-type relationships. We will then search for correlations between integrated pathway effects and donor clinical characteristics.

Legacy plans

All of the software and non-identifying data sets that we will develop for the project will be made available on an open source and open access basis. We will provide detailed documentation to allow researchers to reproduce and build upon our results.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Lincoln Stein	Director, Bio-computing Platform and Senior Principal Investigator, Ontario Institute for Cancer Research
eRA COMMONS USER NAME LINCOLNSTEIN	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Johns Hopkins University	B.A.	1978-1982	Biology
Harvard Medical School, Boston, MA	M.D.	1982-1989	Medicine
Harvard University	Ph.D.	1982-1989	Cell Biology
Brigham & Women's Hospital	Residency	1989-1992	Anatomic Pathology

E. **Positions**

1995-1997	Instructor, Harvard Medical School, Dept. of Pathology, Brigham and Women's Hospital
1992-1997	Director, Informatics Core, MIT Genome Center, Whitehead Institute of Biomedical Research, M.I.T., Cambridge, MA
1997	Director Information Systems, CuraGen Corporation, New Haven, CT
1998-2004	Associate Professor, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2004-present	Professor, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2007-present	Director, Informatics and Biological Computing Platform & Senior Principal Investigator, Ontario Institute for Cancer Research, ON
2009-present	Professor, Department of Molecular Genetics, University of Toronto, Ontario

F. **10 Selected Peer-Reviewed Publications** (From 141 peer-reviewed publications total)

21. Trinh QM, Jen FY, Zhou Z, Chu KM, Perry, MD, Kephart E, Contrino S, Ruzanov P, and **Stein LD**, Cloud-Based Uniform ChIP-Seq Processing Tools for modENCODE and ENCODE. BMC Genomics, 1. 2013 14:494.
22. Watt S, Jiao W, Brown AM, Petrocelli T, Tran B, Zhang T, McPherson JD, Kamel-Reid S, Bedard PL, Onetto N, Hudson TJ, Dancey J, Siu LL, **Stein L**, Ferretti V. Clinical genomics information management software linking cancer genome sequence and clinical decisions. Genomics. 2013 Apr 17.
23. Wang L, **Stein LD**. Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process. BMC Evol Biol. 2013 Feb 28;13:57. doi: 10.1186/1471-2148-13-57. PubMed PMID: 23448166.
24. Wu G, **Stein L**. A network module-based method for identifying cancer prognostic signatures. Genome Biol. 2012 Dec 10;13(12):R112.
25. Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, **Stein LD**, Ferretti V, Watt S, Jiao W, Ng K, Ghai S, Shaw P, Petrocelli T, Hudson TJ, Neel BG, Onetto N, Siu LL, McPherson JD, Kamel-Reid S, Dancey JE. Feasibility of real time next generation sequencing of cancer genes linked to drug response: Results from a clinical trial. *Int J Cancer*. 2012 Sep 5
26. Haw R, Hermjakob H, D'Eustachio P, **Stein L.** Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics*. 2011 Sep; 11(18):3598-613.
27. Feng X, Grossman R, **Stein L**. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*. 2011 May 9;12:139.
28. Wu G, Feng X, **Stein L.** A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*. 2010;11(5):R53.
29. Gerstein MB, *et al*, **Stein L**, Lieb JD, Waterston RH. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*. 2010 Dec 24;330(6012):1775-87.
30. modENCODE Consortium, *et al*, **Stein LD**, White KP, Kellis M. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*. 2010 Dec 24;330(6012):1787-97.

Mark Gerstein

Education
Harvard College, AB Physics '89
Cambridge University, PhD Chemistry '93
Stanford University, postdoc '93-'96, Bioinformatics (advisor M Levitt)

Positions
2006- **AL Williams Prof. Biomedical Informatics, Yale**
2002- co-director Yale Computational Biology and Bioinformatics Program
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

Honors
'89-'93 Herchel-Smith Scholarship for PhD at Cambridge
'93-'96 Damon Runyon-Walter Winchell post-doctoral Fellowship
'09 AAAS Fellow

Consortia
Analysis co-chair: NHGRI modENCODE Project AWG ('07-), Brainspan Project ('09-), 1000 Genomes Functional Interpretation Group ('12-), ENCODE & Cancer Group ('13-) exRNA consortium ('13-)

Publications (senior author on all papers listed below, which are selected from a total of >460; H-index=116)

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, M Gerstein (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342:1235587

E Khurana, Y Fu, J Chen, M Gerstein (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comp Bio* 9:e1002886.

M Gerstein, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." *Nature* 489:91

A Abyzov, J Mariani... (16 authors)... M Gerstein, FM Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells." *Nature* 492:438

B Pei, C Sisu... (10 authors)... J Harrow, M Gerstein (2012). "The GENCODE pseudogene resource." *Genome Biol* 13:R51.

C Cheng, R Alexander... (16 authors)... M Gerstein (2012). " Understanding transcriptional regulation by integrative analysis of transcription factor binding data." *Genome Res* 22:1658.

DG MacArthur, S Balasubramanian... (50 authors)... M Gerstein, C Tyler-Smith (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." *Science* 335:823.

A Abyzov, AE Urban, M Snyder, M Gerstein (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21:974

A Sboner, L Habegger... (9 authors)... MA Rubin, M Gerstein (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." *Genome Biol* 11:R104.

HY Lam, XJ Mu, AM Stütz, A Tanzer, PD Cayting, M Snyder, PM Kim, JO Korbel, M Gerstein (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." *Nat Biotech* 28:47.

RP Alexander, G Fang, J Rozowsky, M Snyder, M Gerstein (2010). "Annotating non-coding regions of the genome." *Nat Rev Genet* 11:559.

KK Yan, G Fang, N Bhardwaj, RP Alexander, M Gerstein (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107:9186.

N Bhardwaj, KK Yan, M Gerstein (2010). "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels." *PNAS* 107:6841

M Gerstein, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330:1775.

Guanming Wu

Education

Ph.D.	Biochemistry and Molecular Biology, Peking University, 1995
M.S.	Biochemistry and Molecular Biology, Shanghai Institute of Biochemistry, Chinese Academy of Sciences, 1992
B.S.	Biochemistry, Nanking University, 1989

Positions and Employment

2013-present	Adjunct Assistant Professor, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Oregon, OR
2009 to present	Consultant to Ontario Institute for Cancer Research, Toronto, ON, Canada
2007 to Present	Senior Bioinformatics Developer, Oregon Health Science University, Portland, OR
2005 to 2009	Scientific Informatics Analyst II, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2003 to 2005	Scientific Programmer, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
2003 to 2005	Software Engineer, Physiome Sciences, Princeton, NJ
2000 to 2001	Computational Postdoctoral Fellow, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
1995 to 2000	Staff Fellow, Boston Biomedical Research Institute, Boston, MA

Editorial Boards

- Dataset Papers in Biology, 2012-present
- Ad hoc reviewer for PLoS One, Bioinformatics, BMC Bioinformatics

Other Activity

- Mentor, Google Summer of Code (GSoC), 2010, 2011, 2013

Selected Publications Related to this Project

1. Croft D*, Mundo AF*, Haw R*, Milacic M*, Weiser J*, **Wu G*** et al. (2014) The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. (Database Issue) (accepted). (* Joint first authors)
2. **Wu G**, Stein L. (2012) A network module-based method for identifying cancer prognostic signatures. *Genome Biol* **13**:R112. Sawey ET, Chanrion M, Cai C, **Wu G** et al. (2011) Identification of a therapeutic strategy targeting amplified FGF19 in liver cancer by oncogenomic screening. *Cancer Cell* **19**(3): 347-58.
3. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, **Wu G** et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*. **42**(7): 579-89,
4. **Wu G**, Feng X, Stein L. (2010) A Human Functional Protein Interaction Network and Its Application to Cancer Data Analysis. *Genome Biol* **11**: R53.
5. Grey F, Tiribassi R, Meyers H, **Wu** G et al. (2010) A viral microRNA down-regulates multiple cell cycle genes through mRNA 5'UTRs. *PLoS Pathog* **6**: e1000967.

LUCAS LOCHOVSKY

Address: 266 Whitney Ave, Bass 437, New Haven CT, 06520 / **Phone:** 203-432-5405
E-mail: lucas.lochovsky@yale.edu

EDUCATION

2008-present **Ph.D. candidate**, Computational Biology & Bioinformatics (CBB) track, Biological and Biomedical Sciences program, Yale University
Thesis supervisor: Prof. Mark Gerstein
Thesis topic: Understanding Noncoding and Systems-level Recurrent Variation in Cancer

2010 **M.Sc.** enroute to the Ph.D., Computational Biology & Bioinformatics (CBB) track, Biological and Biomedical Sciences program, Yale University

2006-2008 **M.Sc.**, Department of Computer Science. University of Toronto
Thesis supervisor: Prof. Thodoros Topaloglou & Prof. John Mylopoulos
Thesis topic: An Entity Resolution Framework For Deduplicating Proteins

2002-2006 **Honours Bachelor of Science**, Trinity College, University of Toronto
Graduated with ***High Distinction*** (cGPA of 3.5 or higher)
Programs of Study: Double Major in Computer Science and Human Biology

PUBLICATIONS

Oct 2013 *Khurana E et al.* **"Integrative annotation of variants from 1,092 humans: application to cancer genomics."** *Science* **342,** 1235587–1235587 (2013).

Sep 2012 *ENCODE Project Consortium* **"An integrated encyclopedia of DNA elements in the human genome."** *Nature* **489,** 57–74 (2012).

Apr 2011 *ENCODE Project Consortium* **"A user's guide to the encyclopedia of DNA elements (ENCODE)"** *PLOS Biology* (2011) at <http://scholarworks.boisestate.edu/bio_facpubs/84/>

Apr 2011 *Jee J, Rozowsky J, Yip KY,* ***Lochovsky L****, Bjornson R, Zhong G, Zhang Z, Fu Y, Wang J, Weng Z, Gerstein M* **"ACT: Aggregation and Correlation Toolbox for Analyses of Genome Tracks"** *Bioinformatics* (2011).

Jan 2011 *Xiong X, Song H, On T,* ***Lochovsky L****, Provart N, Parkinson J* **"Phylopro: A web based tool for the generation and visualization of phylogenetic profiles across Eukarya."** *Bioinformatics* 27, 877–878 (2011).

Dec 2010 *Gerstein, M. B. et al.* **"Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project."** *Science* **330,** 1775–1787 (2010).

June 2008 ***Lochovsky L***, *Topaloglou T* "**An Entity Resolution Framework for Deduplicating Proteins"** Proceedings of "Data Integration in Life Sciences (DILS), 2008"; Evry, France; pp. 92-107, June 25-27, 2008.

CONFERENCE PRESENTATIONS

June 2008 "**An Entity Resolution Framework for Deduplicating Proteins"**
Presented at "Data Integration in Life Sciences (DILS), 2008"; Evry, France; June 25-27, 2008.

RESEARCH EXPERIENCE

Noncoding disruptions of genetic disease, systems-level disruptions of genetic disease, software engineering of whole genome mutation analysis tools, database approaches to whole genome mutation analysis, cancer genomics, biological interaction networks, Peakseq signal aggregation and correlation for Encyclopedia of DNA Elements (ENCODE) project, saturation of genomic feature annotations across cell lines for ENCODE project, whole genome assembly algorithm design, biological pathway applications of graph matching algorithms, protein record deduplication, comparative genomics, phylogenetic analysis, protein-protein interaction software design, protein-protein interaction confidence measures.

Kevin P. White

Education

Yale University, New Haven, B.S./M.S., Biology 1993
Stanford University, Stanford, CA, Ph.D., Developmental Biology 1998
Stanford Genome Technology Ctr, Palto Alto, CA, Postdoc, Biochemistry & Genomics, 1998-2000

Professional Positions

2006-present	Director, Joint Institute for Genomics & Systems Biology, The University of Chicago and Argo National Laboratory
2006-present	James and Karen Frank Family Professor, Human Genetics and Ecology & Evolution, The Universit Chicago
2004-2006	Associate Prof. of Ecology & Evolutionary Biology (joint appointment), Yale University
2004-2006	Associate Professor of Genetics, Yale University School of Medicine
2001-2004	Assistant Professor of Genetics, Yale University School of Medicine

Publications Selected from 97 peer-reviewed publications

1. Michelle N. Arbeitman, Eileen E. M. Furlong, Farhad Imam, Eric Johnson, Brian H. Null, Bruce S. Baker, Mark A. Krasnow, Matthew P. Scott, Ronald W. Davis and Kevin P. White. Gene Expression During the Life Cycle of Drosophila melanogaster. **Science,** 297: 2270-2275, **2002.**
2. Giot L, Bader JS, Brouwer C, Chaudhuri, et al. A genome-scale protein interaction map of Drosophila melanogaster. **Science**, 302: 1727-36, **2003.**
3. Viktor Stolc*, Zareen Gauhar*, Christopher Mason*, Gabor Halasz, Marinus F. van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Barbano, Harmen J. Bussemaker, and Kevin P White. A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster. **Science**, 306:655-60, **2004.**
4. Scott Rifkin, David Houle, Junhyong Kim and Kevin P. White. A mutation accumulation assay reveals extensive capacity for rapid gene expression evolution. **Nature,** 438:220-3, **2005.**
5. Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed and Kevin P. White. "Expression profiling in primates reveals a rapid evolution of human transcription factors." **Nature,** 440:242-5, **2006**.
6. Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Nègre N, Ludwig M, Stricker T, Al-Ahmadie HA, Tretiakova M, Camp RL, Perera-Alberto M, Rimm DL, Xu T, Rzhetsky A, White KP. Analysis of Drosophila Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer. **Science,** 323:1218-22, **2009**.
7. Hua SJ, Kittler R, and White KP. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. **Cell.** 137:1259-71, **2009**.
8. modENCODE Consortium, et, al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. **Science**. 330:1787-97. **2010**
9. Nègre N*, Brown CD*, Ma L*, Bristow CA*, Miller S*, Kheradpour P, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony R, Chen J, Hwang L, Wagner U, Auburn R, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Bild NA, Grundstad AJ, Hanley D, Mannervik M, Venken K, Bellen H, White R, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. A cis-regulatory map for the Drosophila genome. **Nature**. 471:527-31. **2011**.
10. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. **Nature.**489:57-74. 2012.:
11. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumors, **Nature.** 490:61-70. 2012
12. Xiaochun Ni, Yong E. Zhang, Nicolas Negre, Sidi Chen, Manyuan Long and Kevin P. White. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. **PLoS. Biology**. 10(11):e1001420. 2012.
13. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. **Blood**. 121: 975-83. 2013.
14. Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP. A comprehensive nuclear receptor network for breast cancer cells. **Cell Rep.** 3:538-51. 2013.
15. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. **Cell**. Sep 26;155:70-80. 2013.

Mark A. Rubin

Education
B.S. University of Wisconsin, Madison, WI, 1984
M.D. Mount Sinai School of Medicine, NY, 1988

Positions

2007-	Professor of Pathology and Laboratory Medicine, Vice Chair for Experimental Pathology, Weill Cornell Medical College
2009-	Homer T. Hirst Professor of Oncology in Pathology, Weill Cornell Medical College
2013-	Director, Institute for Precision Medicine, Weill Cornell Medical College and New York-Presbyterian Hospital
2006-2009	Associate Member, Broad Institute of Harvard and MIT
2006-2007	Staff Physician, Dana Farber Cancer Institute
2002-2007	Associate Professor of Pathology and Chief of Genitourinary Pathology, Brigham and Women's Hospital, Harvard Medical School

Selected Honors 2007 Team Science Award (Co-Leader with Arul Chinnaiyan), American Association for Cancer Research 2012 GU ASCO Keynote Lecture, "Insights from Genomic Approaches to Oncology Discovery" 2012 Huggins Award, Society of Urologic Oncology 2013 Damon Runyon Cancer Research Foundation Clinical Investigator Award (Mentor for H. Beltran) 2013 Prostate Cancer Foundation Mentor of Excellence Award

Committees Chair, EDRN Prostate Group, NCI (2010-); Chair, PCRP EAB, Department of Defense (2011-); Executive Director, New York Genome Center Executive Committee (2012-)

Publications (selected from over 285)

Dhanasekaran SM…**Rubin MA***, Chinnalyan AM*. Delineation of prognostic biomarkers in prostate cancer. Nature 2001;412:822-826. (1489 Citations) *Co-senior author

Rubin MA…Chinnaiyan AM. alpha-Methylacyl-CoA racemase as a tissue biomarker for prostate cancer. JAMA 2002;287:1662-1670. (567 Citations)

Varambally S…**Rubin MA**, Chinnaiyan AM. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature. 2002 Oct 10;419(6907):624-9. (1342 Citations)

Shah RB…**Rubin MA***, Pienta KJ. Androgen-independent prostate cancer is a heterogeneous group of diseases: lessons from a rapid autopsy program. Cancer Res. 2004 Dec 15;64(24):9209-16. (356 Citations) *Co-senior suthor

Tomlins SA…**Rubin MA**, Chinnaiyan AM. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644-8. (1748 Citations)

Perner S…**Rubin MA***. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res. 2006;66(17):8337-41. (342 Citations) *Senior and corresponding author

Demichelis F…**Rubin MA**. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a Watchful Waiting cohort. Oncogene 2007. (360 Citations) *Co-senior and corresponding author

Berger MF…**Rubin MA***, Garraway LA*. The genomic complexity of primary human prostate cancer. Nature. 2011 Feb 10;470(7333):214-20. (317 Citations) *Co-senior and corresponding author

Rickman DS…Rubin MA. Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci U S A. 2012 Jun 5;109(23):9083-8. (22 Citations)

Demichelis F…**Rubin MA**. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A. 2012 Apr 24;109(17):6686-91. (5 Citations)

Barbieri CE…**Rubin MA***, Garraway LA*. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012 May 20;44(6):685-9. (118 Citations) *Co-senior and corresponding author

Beltran H…**Rubin MA***. Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. Cancer Discov. 2011 Nov;1(6):487-95. (34 Citations) *Senior and corresponding author

Mosquera JM…**Rubin MA**. Concurrent AURKA and MYCN gene amplifications are harbingers of lethal treatment-related neuroendocrine prostate cancer. Neoplasia. 2013 Jan;15(1):1-10.

Baca SC…**Rubin MA***, Garraway LA*. Punctuated evolution of prostate cancer genomes. Cell. 2013 Apr 25;153(3):666 77. *Co-senior author

Khurana E…**Rubin MA**, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1,092 humans: application to cancer genomics. Science. 2013 Oct 4;342(6154):1235587.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Pan-Cancer analysis of digestive system and urinary system cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Zhibo Gao, BGI Shenzhen, member of the International Cancer Genome Consortium (ICGC), key member of the China Cancer Genome Consortium (CCGC);
Lin Li, BGI Shenzhen, key member of the China Cancer Genome Consortium (CCGC).

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

The human body is made up of several organ systems that work together as one unit, within which the digestive system and the urinary system are closely connected and cooperating in the nutrients generation and waste removal process. Both of these two systems are bound by epithelial tissues tracts And contact with inner-tract components directly, tumors occur in organs within these two systems account for the majority of human tumors and leading to a large number of deaths every years.

For years, many sporadic research showed that tumor within the same organ system usually share common mutations, for example, *TERT* promoter mutations are common in urinary system tumors and *P53* and *EGFR* mutation are implicated in pathophysiology of esophageal and stomach cancers. In the recent TCGA's Pan-Cancer studies, molecular analysis showed that cancers of different organs have many shared features, and these similarities across cancers can have important implications for treatment. All of these inspired us to raised the questions " Did cancers occur in the same organ system possess similar mutations patterns? Are there any organ-system-specific mutations and what are the common mutations shared by different organ systems?"

In our prior study, we had perform WGS, WES and transcriptome sequencing across 8 tumor types in the two staple organ system(digestive system and urinary system) and finished basic data analysis of all the tumor types. However due to the relatively small data size we haven't carry out integrating analysis in organ system level. By joining into the Pan-Cancer Analysis Working Group and accessing to the pan-cancer Data sets, we hope that we could increase the statistical power to detect functional genomic determinants across tumors of the whole organ systems, and to identify both organ system specific mutations and intrinsic molecular commonalities across different organ systems.

Timelines & resources dedicated to project

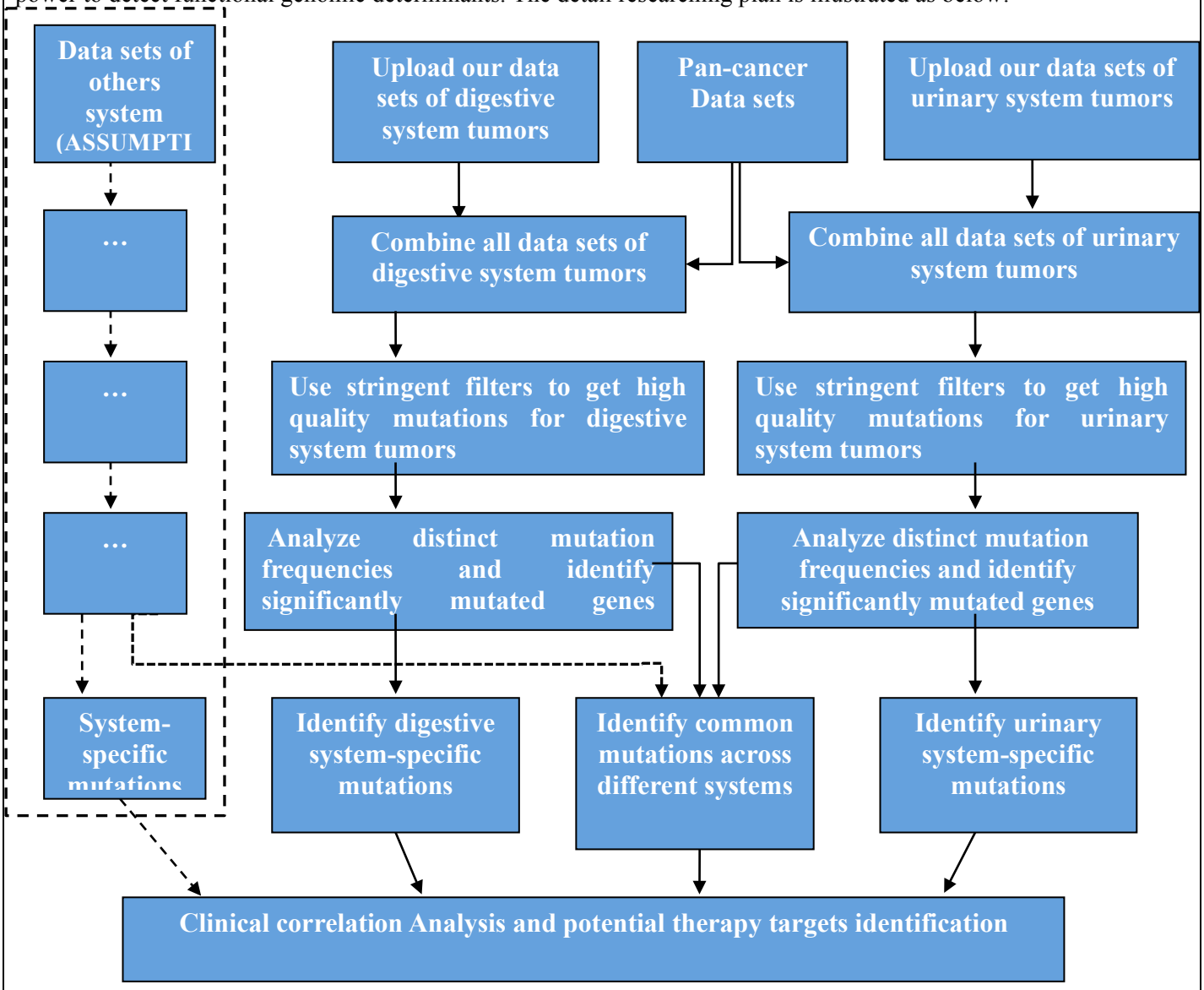
Timelines:

December, 2013-January,2014	Upload the data sets of digestive system and urinary system cancers.
January,2014- February,2014	Accessing to the pan-cancer Data sets, and combine our data sets with the pan-cancer Data sets.
February,2014-May,2014	Use stringent filters (Methods) to get high quality mutations for the 8 cancer types.
May,2014-July,2014	Analyze distinct mutation frequencies and identify significantly mutated genes.
July,2014 - November 2014	Analyze organ-system-specific mutations and common mutations shared by different systems.
November 2014- December 2014	Clinical correlation Analysis.
January 2015- February 2015	Manuscript preparation.
20th March 2015	Manuscript submission.

Data sets from the pan-cancer studies we would be particularly dependent on:
Data sets of digestive system and urinary system cancers from the pan-cancer studies.

Research proposal

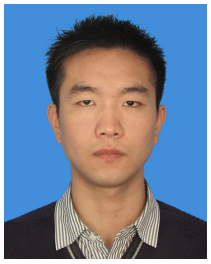
We plan to perform a systematic analysis of 1200 tumours across 8 tumor types in two staple organ system (digestive system and urinary system) of human body, however tumors within other organ systems also can be included in this study. In our prior study, we had performed WGS, WES and transcriptome sequencing of 600 tumours across 8 tumor types (esophagus cancer, gastric cancer, colorectal cancer, hepatocellular carcinoma, islet cell tumor, renal carcinoma, renal pelvis cancer and Bladder cancer) belong to digestive and urinary system respectively, and finished basic data analysis of all the tumor types. By joining into the Pan-Cancer Analysis Working Group and accessing to the pan-cancer Data sets, we hope that we could enlarge the sample set for each kind of tumor and increase the statistical power to detect functional genomic determinants. The detail researching plan is illustrated as below:



Legacy plans

We plan to develop algorithms to Standardization of mutation data across different tumor types for getting high quality mutations.
We would also commit ourselves to the development of algorithms to indentify organ-system-specific mutations and common mutations shared by different organ systems, and we hope that in the end of the research we could develop standard pipelines to facilitate the Pan-Cancer analysis in human body system.
And the algorithms or pipelines developed by us would be available to the research community.

Resume of Zhibo GAO □□□□□



Gender: Male
Address: BGI-Shenzhen, Yantian district
Cell: +86 13424201215
Shenzhen, Guangdong, China, 518083 Email: gaozhb@genomics.cn

Education

- *Aug. 2009 - Sep. 2013*

Degree: Doctor of Philosophy Medical Sciences Chinese University of Hong Kong (CUHK)

- *Sep. 2004 - Jul. 2008*

Degree: Bachelor of Engineering Bioinformatics Huazhong University of Science and Technology (HUST)

Work Experience

- Director of Medical Sciences Research, BGI Tech Solutions Co Ltd

Jul. 2013- Now

- Director of Cancer Research, BGI Tech Solutions Co Ltd

Feb. 2011- Jul. 2013

- Bioinformatics Analysis, BGI-Shenzhen

Jul. 2009- Feb. 2011

Biography

Dr. Gao received his doctor degree of bioinformatics from Chinese University of Hong Kong (CUHK). As a senior bioinformatician and director of medical science research division, he and his team are now working on collaborations all over the world, especially on cancer research, aiming to reveal the repertoire of oncogenic mutations, and enable the development of novel cancer therapies. The research field of him focuses on genomics, transcriptomics and epigenomics, particular integrated analysis. He is a member of the

International Cancer Genome Consortium (ICGC), and a key member of the China Cancer Genome Consortium (CCGC). Now, he mainly takes charge of the bioinformatics analysis of five ICGC cancer projects from China, including esophageal cancer, gastric cancer, liver cancer, colorectal cancer as well as nasopharyngeal cancer.

Main Publication

Yanan Cao,* **Zhibo Gao**,* Weiqing Wang,* Lin Li* … Jun Wang,‡ Guang Ning.‡ Whole Exome Sequencing of Insulinoma Reveals Recurrent T372R mutations in YY1. **Nat Commun.** 2013 Dec 10;4:2810. **Co-first author (2/20)**

Guangwu Guo,* Xiaojuan Sun,* Chao Chen,* Song Wu,* Peide Huang,* Zesong Li,* Michael Dean,* … **Zhibo Gao** … Yaoting Gui,‡ Jun Wang,‡ Zhiming Cai.‡ Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. **Nat Genet**. Published online 13 October 2013. **Contributing author (31/52)**

Noel FCC de Miranda,* Roujun Peng,* … **Zhibo Gao** … Qiang Pan-Hammarström.‡ DNA repair genes are selectively mutated in diffuse large B cell lymphomas. **J Exp Med.** 2013 Aug 26;210(9):1729-42. **Contributing author (8/20)**

Devendra Singh,* Joseph Minhow Chan,* Pietro Zoppoli,* Francesco Niola,* … **Gao Z** … Anna Lasorella,‡ Raul Rabadan,‡ Antonio Iavarone.‡ Transforming fusions of FGFR and TACC genes in human glioblastoma. **Science.** 2012 Sep 7;337(6099):1231-5. **Contributing author (12/24)**

Yongmei Son,* Lin Li,* Yunwei Ou,* **Zhibo Gao**,* Enmin Li,* Xiangchun Li* … Jun Wang,‡ Qimin Zhan.‡ Identifications of genomic alterations in esophageal squamous cell cancer. **Nature** (Manuscript under revision). **Co-first author (4/41)**

Chenguagn Li,* **Zhibo Gao**,* Yihua Sun,* Xiangchun Li,* Fei Li,* … Hongbin Ji,‡ Haiquan Chen,‡ Jun Wang,‡ Qingyi Wei.‡ Exome Sequencing Identifies Frequent Somatic Mutations in Cell-cell Adhesion Genes in Lung Squamous Cell Carcinoma. **Nat Genet** (Manuscript under review). **Co-first author (2/50)**

Resume of Lin Li

Gender: Male
Address: BGI-Shenzhen, Yantian district Cell: 15986799010
Shenzhen, Guangdong, China, 518083 Email: lilin@genomics.cn

Education

- *Sep. 2006 - Jul. 2010*

Degree: Bachelor of Engineering Computer Science and Technology Ocean University of China

Work Experience

- Bioinformatics Analysis, BGI Tech Solutions

Jul. 2013- Now

- Bioinformatics Analysis, BGI-Shenzhen

Aug. 2010- Jul. 2013

Biography

Mr. Li received his bachelor's degree of Computer Science and Technology from Ocean University of China. As a senior bioinformatician, he and his team are now working on cancer research, which focus on genomics, especially pipeline of cancer genome analysis. He is a member of the ICGC Data Coordination and Management Working Group, a key member of the China Cancer Genome Consortium (CCGC), and takes charge of ICGC-China data submission. Now, he is mainly responsible for the bioinformatics analysis of esophageal cancer project from China.

Main Publication

Yanan Cao*, Zhibo Gao*, **Lin Li*** … Jun Wang‡, Guang Ning‡. Whole Exome Sequencing of Insulinoma Reveals Recurrent T372R mutations in YY1. **Nat Commun.** 2013 Dec 10;4:2810. **Co-first author (3/20)**

Zhengyan Kan*, Hancheng Zheng*, Xiao Liu* …**Lin Li…**Yingrui Li‡, John M.Luk‡ and MaoMao‡. Whole genome sequencing identifies recurrent mutations in hepatocellular carcinoma. **Genome Research**. 2013 June 20;10.1101. **Contributing author(25/50)**

Guangwu Guo*, Xiaojuan Sun*, Chao Chen*, Song Wu*, Peide Huang*, Zesong Li*, Michael Dean* … **Lin Li** … Yaoting Gui‡, Jun Wang‡, Zhiming Cai‡. Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. **Nat Genet**. Published online 13 October 2013. **Contributing author (35/52)**

Yongmei Son*, **Lin Li***, Yunwei Ou*, Zhibo Gao***,** Enmin Li*, Xiangchun Li* … Jun Wang‡, Qimin Zhan‡. Identifications of genomic alterations in esophageal squamous cell cancer. **Nature** (Manuscript under revision). **Co-first author (2/41)**

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.caby ~~27th November~~31stDecember, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The landscape of RNA-editing in human cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

Xueda Hu, Cancer Institute & Hospital, Chinese Academy of Medical Science

Name(s) & institute(s) of junior investigators	Name(s)& institute(s) of non-ICGC collaborators
	Jialou Zhu, BGI

Background and preliminary data

RNA editing, which is defined as the nucleotide sequence change of RNA transcripts relative to that of the encoding DNA, could enhance the RNA diversity, and lead to changes in amino acid sequence and alternative splicing, thereby increasing the complexity of gene expression. RNA editing has been connected to cancer development and progression. Analysis of many editing sites in various cancer types is expected to provide new diagnostic and prognostic markers and might contribute to early detection of cancer and monitoring of therapy response.

With the merit of next generation sequencing, we could profile the RNA editing sites in an individual genome, and extend it to cancer research field. From 2009, many bioinformatic methods for detecting RNA editome have been developed, as well as one computational pipeline was exploited in our lab (Nat Biotechnol. 2012 30(3):253-60). However, most of previous global analysis was implemented on limited human peripheral blood or cell-line samples.

Here we would like to address the issue of RNA editome involved in human cancers. With the preliminary work, we have sequenced 2 hepatocellular carcinoma tissues with normal adjacent tissues and 2 cell lines, with whole genome and deep transcriptome. We identified 6810 RNA editing sites in all 6 samples and 2816 overediting sites in tumor tissues. We have re-identified AZIN1 which has been reported before (Nat Med. 2013 19(2):209-16), as well as find some novel editing events in liver cancers. We also discussed some feature of sequence context of editing sites and its relationship with abundance of RNA which carries editing variances. Logically, we consider if we could apply our method to more tumor types and more tumor samples. Take advantage of samples with matched exome/genome sequence data and RNA-seq data in ICGC dataset, we plan to make a comprehensive profile of RNA editing in a variety of human cancers.

Timelines & resources dedicated to project

2014.1~2014.3 Access the ICGC pan-cancer genome data, complete the RNA editome detection in each sample, and determine the cancer-related RNA editing profile in every tumor type, find the tumor-specific editing events.

2014.4~2104.6 Compare the RNA editings with other cancer-related variations (e.g. somatic mutations, copy number changes, RNA expression changes), analysis of the sequence and structural features of RNA editings, identified some RNA editing sites associate with pathological features and prognosis.

2014.6~2014.12 Choose some RNA editing sites and test if it affect tumorigenicity (e.g. tumor cell proliferation, progression and apoptosis).

2015.1~2015.3 Write the scientific article and finish the work.

Research proposal

In this study, we project to make a comprehensive profile of RNA editing in a variety of human cancers. The first assignment is global identification of RNA editing sites in the given dataset, record their editing degree, and get an annotation. Then, comparing editing degrees between tumor and normal tissue data, we will identify the overediting sites in tumors. Meanwhile, miRNAs also have possibility to carry editing events. Consequently, we will profile nucleotide sequence variants in miRNAs identified from the small RNA-Seq data. Basically, we will apply the RNA editing detection pipeline set previously in our lab (Nat Biotechnol. 2012 30(3):253-60). Considering it was argued about noncarnonical editing events (Nat Biotechnol. 2012 30(3):246-7 & Nat Biotechnol. 2013 31(1):19-20), we have updated our informatics algorithm with several adjustments and additional filters to facilitate the detection of bona fide editing from the RNA-Seq reads.

As we get the RNA editing and overediting loci, the profile will be characterized in each tumor type. We will calculate 12 types of differences between RNA and DNA sequence; depict the distributions of editing sites on a carnonical gene structure. We will analyze sequence features of predicted A-to-I editing sites and the flanking regions, including sequence preferences and conservation of flanking bases. We will compare altered nucleotides in Alu, repetitive non-Alu and nonrepetitive sequences. Furthermore, functional enrichment in transcripts with RNA editing and overediting will be performed both in biological pathway categories and gene ontology annotation categories.

Next, with the hypothesis that RNA editing could be a complementary factor for DNA sequence aberrance in tumorigenecity, we will do comparative analysis of RNA overediting in tumors with other cancer-related genomic variations, including somatic mutations, indels, and copy number alteration in tumor cells. We will also analyze the association of RNA overediting with RNA abundance perturbation or RNA structure changes specific in tumor tissues. Additionally, we will explore the correlation of gene expression with RNAs which occurred cancer-related editing events, try to discuss the molecular mechanism in their occurrence by their profile and preference. Furthermore, we will determine some putative RNA editing or overediting sites if they have prognostic value for a given cancer type, or associate with any pathological features.

As we know, the most frequent type of editing in humans is the conversion of A to I, which is catalysed by the double stranded RNA (dsRNA) specific ADAR family of proteins. We would like to infer attribution of ADARs expression in cancer-related RNA editing in human tumors, both in the scope of degree and spectrum. Moreover, we will detect relationship of tumor pathogenesis with altered gene specific editing patterns induced by the differentially expressed ADARs. Lastly, we will try to recognize a model for certain neighbouring nucleotides ADAR targets preferred in tumor or a specific tumor types.

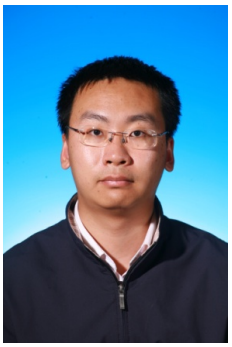
If any candidate RNA editing and overediting sites with specific clinical value or biological interest are selected, we will initiate some cell-line based experiment to discover whether the gain-of-function phenotypes are regulated in an editing–dependent manner. The edited type DNA sequence will be transduced into cells with appointed tumor types, and tumorigenecity phenotype will be assayed, including cell growth and clonality, cell invasion and migration, and apoptosis rate, etc.

Legacy plans

1. Describe the algorithm of RNA editing detection and submit the related analysis software, including steps of distinguish overediting loci in tumor tissues.

2. Submit all RNA editing sites and annotation information identified in this program, separating the individual and cancer-related overediting loci.

Xueda Hu *Curriculum Vitae*



Dr. Xueda Hu is an assistant professor from Cancer institute, Chinese Academy of Medical Sciences. My research major is genomic biology in a variety of cancers. When I was doing my PhD study and research in a former institute, my colleagues and I have done exome sequencing and analysis in bladder cancer and renal cell cancer. We detected frequent mutations in the ubiquitin-mediated proteolysis pathway (UMPP), and alterations in the UMPP were significantly associated with overexpression of HIF1a and HIF2a in the renal cell carcinomas. Additionally, we identified genetic aberrations of the chromatin remodeling genes (UTX, MLL-MLL3, CREBBP-EP300, NCOR1, ARID1A and CHD6) frequently in transitional cell carcinomas. I received my PhD degree from Chinese Academy of Science in 2011.

Employment & Education

9/2006 – 1/2012 PhD student, Beijing Institute of Genomics, Chinese Academy of Science, China

9/2006 – 4/2013 Staff Scientist, Beijing Genomics Institute (BGI), China

5/2013 – present Assistant Professor, Cancer Institute, Chinese Academy of Medical Sciences, China

Research Interests

Cancer Genomics

Selected Publications

Guo G*, Gui Y*, Gao S*, Tang A*, **Hu X***, et al., Frequent mutations of ubiquitin-mediated proteolysis pathway in clear cell renal cell carcinoma. **Nat Genet.** 2011 Dec 4;44(1):17-9.

Zhu J*, Jiang Z*, Gao F*, **Hu X***, et al., A systematic analysis on alterations of DNA methylation and expression of both mRNA and microRNA in bladder cancer. **PLoS One.** 2011;6(11):e28223.

Li X*, Chen J*, **Hu X***, Huang Y*, et al., Comparative mRNA and microRNA Expression Profiling of Three Genitourinary Cancers Reveals Common Hallmarks and Cancer-Specific Molecular Events. **PLoS One.** 2011;6(7):e22570.

Gui Y*, Guo G*, Huang Y*, **Hu X*,** Tang A*, et al., Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. **Nat Genet.** 2011 Aug 7;43(9):875-8.

Zhang G*, Guo G*, **Hu X***, Zhang Y*, et al., Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. **Genome Res.** 2010 May;20(5):646-54.

* Co-first author

Jialou Zhu *Curriculum Vitae*

Basic Information

I am a joint-trained PHD student under guidance of Prof. Huanming Yang and Prof. Xiuqing Zhang from BGI and Prof. Xiaodong Li from Wuhan University. I received my Bachelor of Science degree from Wuhan University in 2011.

Employment & Education

9/2011–Now PhD, Dept. of Cell Biology, Wuhan University, China
9/2011–Now Project Manager, Dept. of Research, BGI, China
9/2007-7/2011 B.S., College of Life Sciences , Wuhan University, China.

Research Interests

Cancer genomics, cancer epigenomics, etc.

Publications

Huang Y*, Gao S*.. **Zhu J**.. et al., Multilayered molecular profiling supported the monoclonal origin of metastatic renal cell carcinoma. **Int J Cancer**. 2013 Dec 6.

Wu S*, Lv Z*, **Zhu J***, et al., Somatic Mutation of the Androgen Receptor Gene Is Not Associated with Transitional Cell Carcinoma: A "Negative" Study by Whole-exome Sequencing Analysis. **Eur Urol**. 2013 Dec;64(6):1018-9.

Zhu J*, Jiang Z*, Gao F*, Hu X*, et al., A systematic analysis on DNA methylation and the expression of both mRNA and microRNA in bladder cancer. **PLoS One**. 2011;6(11):e28223.

Guo G*, Gui Y*.. **Zhu J**, et al.,. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. **Nat Genet**. 2011 Dec 4;44(1):17-9.

Major Honors & Awards

2011–2012 1 times National scholarships of Postgraduates.
7/2011 Outstanding BS. Graduate of Wuhan University
2007–2011 3 times National scholarships of Undergraduates.

Experiences

3/2009–3/2010 Vice-President of Student Union of College of Life Sciences in Wuhan University

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by-31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrated genome and transcriptome analysis to assess the impact of RNA editing on cancer progression

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Huanming Yang and Jun Wang, BGI (China)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s)& institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yong Hou and Xun Xu, BGI (China)	Qiang Pan-Hammarström, BGI and Karolinska Institutet (Sweden); Nina Papavasiliou, Rockefeller University (USA)

Background and preliminary data

Background

According to the current paradigm, cancer originates in dysfunction at the genomic level, and tumors are generally grouped by alterations observed at the DNA level, epigenetic or otherwise. We consider however that RNA editing, or the site-specific alteration in the sequence of RNA, may be a very important type of epigenetic alteration that is involved in driving cancer progression, and could serve as the basis for a new method to categorize tumors, thus better refining treatment options.

RNA editing is a molecular process that results in specific nucleotide alterations within mRNA, thus altering the informational content of the transcriptome. There are two main types of RNA editing. The first and most common one is the deamination of adenosine (A) to inosine (I - decoded as G), mediated by the Adenosine deaminases that act on RNA (ADARs). ADAR editing is widespread in human transcriptomes: it is mostly focused on 3'UTRs (often those that contain inverted Alu repeats), but has also been found in intron-exon junctions (where editing has been causal to alternative splicing). ADAR editing can also drive cancer progression through amino acid recoding[1]; recent binary comparisons between cancer genomes and transcriptomes have revealed additional examples of the involvement of ADAR editing in cancer progression (through the recoding of AZIN1[2]). The second type of RNA editing is the product of the deamination of cytosine (C) to uracil (U - decoded as T in cDNA) and is mediated by the APOBEC1 protein and its relatives. This type of editing is far less common, but it has also been found to play roles in cancer progression[1,3,4]. Here we propose that specific RNA alterations represent an untapped source of information that could potentially reveal associations that are more predictive of disease susceptibility than DNA changes and might also offer a direct molecular understanding of underlying mechanisms. BGI has developed a pipeline for detecting RNA editing sites and successfully applied to YH genome[5]. Together with our external collaborators, we plan to apply a modified version of pipeline to the pan-cancer dataset to explore the following questions:

1) How widespread is RNA editing in cancer? Toward this we will catalog both types of RNA editing events across different tumor types, as well as within a single type of tumor in the context of disease progression (healthy tissue, tumour, metastatic disease).
2) What is the contribution of RNA editing to altered gene expression in specific tumor types? Major tumor sequencing projects have been conducted to identify cancer related genes in which mutations occur more frequently than expected by random chance at the DNA level. Reports on RNA editing are rare[4] but intriguing. In pilot studies with healthy tissue we have demonstrated that RNA editing plays key roles in normal physiology by affecting the expression of related proteins and modulating the function of entire pathways[6] . We plan to explore the distribution of RNA editing sites in cancer related genes in great detail, to understand what aspects of gene expression might be impacted. For example, edits in coding regions might affect amino acid decoding or alter splicing; edits in non-coding regions might affect RNA stability (eg through miRNA target site editing), translational efficiency, localization etc.
3) How does RNA editing affect disease progression in specific cancers? We have found that RNA editing events are not randomly distributed in the transcriptome, but editing tends to target clusters of related transcripts whose altered expression could modulate entire pathways. We therefore plan to cluster edited

transcripts across different conditions, to derive RNA editing signatures that might be prognostic for disease susceptibility and progression.

Preliminary data

APOBEC-1 was originally identified as the enzyme that site-specifically edits the mRNA of the lipid metabolism protein apolipoprotein B (apoB) to convert the codon CAA for Gln2153 to the stop codon UAA in the small intestine[7]. Many other Apobec-1 editing sites have since been identified. A transcriptome-level assessment of C-to-U editing in healthy tissue (intestinal enterocytes) revealed that editing is primarily confined to 3' untranslated regions (UTRs[8]). In a follow-up study we have identified instances of RNA editing in additional types of healthy tissue (e.g. mouse macrophage populations) and have shown that there too (1) RNA editing targets almost exclusively 3'UTRs, that (2) C-to-U editing in 3'UTRs often results in a reduction in translation (in reporter assays) and (3) that editing often targets miRNA target sites within UTRs, thus altering miRNA:mRNA pairing.

Aside from instances of APOBEC-1 expression and editing in healthy tissue, there are reports in the literature of APOBEC-1 overexpression resulting in tumor formation (e.g. liver[9]). Conversely, ablation of APOBEC1 expression results in specific and drastic amelioration of cancer progression in mouse models of colon cancer[3] as well as in mouse models of testicular cancer[4]. These results clearly suggest a specific and driving effect of RNA editing on tumor progression.

In preliminary studies we have analyzed whole genome and transcriptome data from 25 lung cancer datasets with paired normal samples. We have found specific upregulation of APOBEC-1 in tumor (and metastatic) vs. healthy tissue, and are currently assessing these samples for APOBEC-1 specific RNA editing signatures. We have also analyzed whole genome and transcriptome data from 65 prostate cancer datasets with paired normal samples, and have already identified edit sites (to be further validated). These are the first attempts to analyze APOBEC-1 signatures and the contribution of RNA "mutation" in the context of human disease and we hope to continue our analysis toward the specific aims outlined above.

Once our analysis of the currently available datasets is completed and validated, RNA editing signatures can be assessed for their prognostic value by expanding the analysis to additional cohorts. Thus, the ICGC pan-cancer dataset would provide a unique resource for the study proposed and is expected to give new insights into the roles of RNA editing in cancer.

References

1. Mukhopadhyay, D. *et al.* C-->U editing of neurofibromatosis 1 mRNA occurs in tumors that express both the type II transcript and apobec-1, the catalytic subunit of the apolipoprotein B mRNA-editing enzyme. *Am J Hum Genet* 70, 38-50 (2002).
2. Chen, L. *et al.* Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat Med* 19, 209-16 (2013).
3. Blanc, V. *et al.* Deletion of the AU-rich RNA binding protein Apobec-1 reduces intestinal tumor burden in Apc(min) mice. *Cancer Res* 67, 8565-73 (2007).
4. Nelson, V.R., Heaney, J.D., Tesar, P.J., Davidson, N.O. & Nadeau, J.H. Transgenerational epigenetic effects of the Apobec1 cytidine deaminase deficiency on testicular germ cell tumor susceptibility and embryonic viability. *Proc Natl Acad Sci U S A* 109, E2766-73 (2012).
5. Peng, Z. *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* 30, 253-60 (2012).
6. Hamilton, C.E. *et al.* APOBEC-1 is a transcriptome-wide editor of 3'UTRs. *Submitted.* (2014).
7. Hamilton, C.E., Papavasiliou, F.N. & Rosenberg, B.R. Diverse functions for DNA and RNA editing in the immune system. *RNA Biol* 7, 220-8 (2010).
8. Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S. & Papavasiliou, F.N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat Struct Mol Biol* 18, 230-6 (2011).
9. Yamanaka, S. *et al.* Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proc Natl Acad Sci U S A* 92, 8483-7 (1995).

Timelines & resources dedicated to project

Timelines:

Abstract submission – 2013.12.30
Data and software uploads – 2014.01.01-2014.02.01
Core variant calling – 2014.02 - 2014.09
Scientific analyses and validation –2014.04 - 2014.12
Manuscript preparation – 2015.01 – 2015.02
Manuscript submission – 2015.03.20

Resources: RNA-sequencing data and matched WGS data for ~1000 cancer samples

Research proposal

1. Development of tumor RNA editing pipeline and validation of editing instances

The RNA-seq data will be aligned to the UCSC hg19 reference genome using TopHat and the aligned reads will then be processed using SAMTools to obtain a pileup of mismatches against the reference genome. Custom Python and BASH scripts will be deployed, filtering by editing rate, read depth, and read quality (we have already written and validated those). Structural variants will be excluded by eliminating entire neighborhoods of reads containing high rates of non-canonical editing (i.e. not A-to-I or C-to-U). Lists will also be further refined by filtering out C-to-T or A-to-G mismatches (SNVs) that are present in WGS from matched healthy tissue.

We will also use an additional bioinformatics approach to detect RNA editing (which we call the "vector approach", inspired by work from professor Fred Cross, a Rockefeller colleague who is a yeast geneticist). Specifically, we will use the pileup output from SAMtools to build vectors composed of the number of A's, T's, G's, C's, insertions, and deletions that are called at each genomic coordinate for the healthy and diseased matched transcriptome datasets. Thresholding by the vector magnitudes, coefficients of variation, and a minimum angle between the two vectors for a specific genomic coordinate where the transcriptome from diseased tissue shows a potential C-to-U transition and which is not present in matched healthy tissue, yields high confidence edits.

Putative edits that are called using both approaches will be of even higher confidence. Final putative lists of C-to-U edit sites will also be cross-referenced against published datasets, and annotated using Annovar (Wang et al., 2010). A random subset of the sites will be validated using Sanger sequencing to calculate a false positive rate. Gene ontology analysis will also be performed to determine if editing is enriched is specific functional pathways, with a special focus on editing that is found in established oncogenes or tumor suppressors, as well as in genes that are implicated in processes such as proliferation, migration, apoptosis, and differentiation.

Caveats: A major challenge in identifying bona fide RNA editing events is distinguishing noise due to sequencing error and alignment artifacts from actual edits (Bass et al., 2012). This can be rectified in part with longer or paired reads, as proposed here, and only using highly covered sites, especially for multiply edited loci. One advantage of studying APOBEC-1-mediated editing as well as (or instead of only) ADAR-mediated A-to-I editing is that highly processive C-to-U editing editing is much less prevalent, and so alignment and sequencing issues due to multiply edited reads are not as large a concern. Also, we plan on all possible precautions to minimize noise, such as only using uniquely mapping reads in alignments, setting strict quality thresholds, utilizing RNA-seq data from multiple biological replicates, etc.

2. Scientific analysis: RNA editing signatures and mechanism

The concept we will be testing in this proposal is that RNA editing is directly involved in driving tumour development and progression. This concept is supported by both gain- and loss- of function genetic experiments which have already been published in mouse, but where the key mechanistic link has not been made. The validation of our hypothesis will be the direct demonstration, using the comparative WGS/RNA-seq tools we have already generated, of the involvement of the RNA editors ADAR and APOBEC-1 in these processes. And further, of direct relevance to the medical community, our experiments will also delineate the oncogenic pathways altered by RNA editing in tissues for which there have been genetic implications of such an involvement. As an additional benchmark, our bioinformatic analyses of matched genome:transcriptome data, will establish relevance to human tumours.

We envision several steps to this analysis (delineated in specific aims, above): validation (using sanger sequencing from the tumor samples), clustering into loci of editing such as (a) CDSs - with the possibility of recoding potential like NF1 or AZIN1, which can be validated *in vitro*; (b) intron:exon junctions, with the possibility of alternative splicing, which can be validated both using RNAseq datasets and in using wet-lab approaches; and (c) UTRs, with the possibility of alterations in regulation (which can be tested using reporter assays). We have direct experience with all of these wet-lab techniques and do not anticipate roadblocks there.

Overall however, an important caveat to keep in mind is the possibility that a subset of tumor cells (or indeed, the subset within tumors of infiltrating immune cells) contribute the observed editing signature. The source of editing within the tumor does not invalidate our hypothesis. For example, editing could affect the small numbers of tumor cells that are stem-like, and could contribute to more aggressive growth (as would be expected from the mouse experiments). Alternatively, editing within infiltrating immune cells (distinct

from the normal cellular context) could contribute the editing signature: for example tumor-associated macrophages (TAMs) in the lung (which are known to promote oncogenesis) may contribute distinct editing signatures that alveolar macrophages which normally reside in lung tissue, do not (a possibility which we could test by assessing other types of TAM-containing solid tumours for equivalent signatures).

Should our hypothesis be correct we envision future involvement with clinicians to ascertain prospectively the outcomes of cases where tumor samples show evidence of editing vs. those that do not, thus potentially providing a new diagnostic tool for classifying tumors. We will therefore strive to disseminate this knowledge widely, through high profile publications as well as through seminars and collaborations. We will also aid future collaborators in the exploitation of this knowledge for the benefit of patients (e.g. the development of diagnostic tests to assess levels of editing that might be predictive of disease progression).

3. Expansion of our efforts to cataloguing RNA editing sites calling in pan-cancer dataset, if warranted by data thus far. (See discussion of TAMs above).
4. Upload genome data, pipeline and other software packages.
5. Writing manuscripts
6. Summit

Legacy plans

Tumor RNA editing detection software package.
Databases of tumor RNA editing sites for GBrowser.

CURRICULUM VITAE

Huanming YANG, Ph.D. BGI, China

ADDRESS

BGI-Headquarter, 11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China
Tel: (86) 755-2527-3620, Fax: (86) 755-2527-3620, E-mail: yanghm@genomics.org.cn

RESEARCH INTEREST & FIELD

Genomics/Genetics

EDUCATION

1988	Ph.D.	ıte of Medical Genetics, University of Copenhagen, Denmark
1982	M.Sc.	·tment of Biology, Medical School of Southeast University, Nanjing, China
1978		ge of Life Sciences, Zhejiang University, Hangzhou, China

PROFESSIONAL EXPERIENCES

-	ssor & President	:hina
-2007	ssor & Director	ıg Genomics Institute, Chinese Academy of Sciences
-2003		ın Genome Center, Institute of Genetics, Chinese Academy of Sciences

AWARDS AND HONORS

gn Academician	an National Academy of Sciences（Leopoldina）
gn Academician	n National Science Academy
N	hird World Academy of Sciences (TWAS)
emician	se Academy of Sciences (CAS), China
gn Fellow	pean Molecular Biology Organization (EMBO)

Publications

Dr. Huanming Yang has made a significant contribution to the international Human Genome Project (HGP), the HapMap Project, and the 1000 Genomes Project, as well as to the first Asian genomes, human pan-genome, ancient genomes, gut metagenomes, cancer genome, exomes and mythylome. He has also contributed to sequencing and analysizing genomes of rice, potato, maize, pigeonpea, soybean, cucumber, cabbage, tomato, foxtail millet and chicken, silkworm, panda, ants, naked mole rat, cynomolgus and Chinese rhesus macaque, yak, CHO and iPS cell lines, single cell sequencing in cancers, SARS virus, lethal *E.coli,* and many other organisms, as published in many peer reviewed scientific journals including *Cell, Nature* and *Science*.

CURRICULUM VITAE

Jun Wang
Birth date: June 4, 1976
Citizenship: P.R.China
Office phone: 0086-755-25273789
E-mail: wangj@genomics.org.cn
Fax: 0086-755-25273796

Title
Professor of Genomics and Bioinformatics

Education/training
1997.9 – 2002.7 College of Life Sciences, Peking University, Beijing, China
Ph.D. in life science
1992.9 – 1997.7 College of Life Sciences, Peking University, Beijing, China
B.Sc. in life science

Positions
2013- Visiting Professor, University of Hong Kong, Hong Kong, China.
2013- External Advisory Board Member, Institute for Genomic Biology, University of Illinois at Urbana-Champaign, USA
2013- Visiting Professor, King Abdulaziz University, Saudi Arabia.
2013- Visiting Professor, Macau University of Science and Technology, Macau, China
2013- Scientific Advisory Board Member, L'Oreal, France.
2013- Director and Professor, BGI, China

Honours and awards
"Major Award" (by Shenzhen Municipal Government) (2013)
"TanJiaZeng Life Science Innovation Award" (by Shanghai Bio-Medical Associations) (2013)
"Fortune's 40 under 40" (by Fortune Magazine) (2013)
"ZhouGuangZhao Award" (by ZhouGuangZhao Foundation) (2013)
"The Hottest Scientific Researchers of 2012" (by Thomson Reuters) (2012)
"Nature's 10; the year in Science" (by Nature) (2012)

Educational activities
Supervised 12 Ph.D., 10 Master Students. Presently, 35 graduate students are training in the laboratory.

Presentations
Numerous pre- and postgraduate educational activities at the University of Copenhagen, the University of Southern Denmark, the University of Aarhus, the Peking University, Chinese Academy of Sciences, the Zhejiang University, and the Zhongshan University. More than 100 invited lectures at international and national scientific meetings.

Publications
Presently 200+ peer-reviewed original scientific papers, of which several are published in high-impact journal including *Nature, Nature Genetics, Nature Biotechnology, Nature Review Genetics, Science, N Engl J Med*., *PloS Biology, Genome Research, Proceedings of National Academy of Science* and *Plant Cell.* Specifically, WJ published 100+ papers in *Cell, Nature* (including *Nature* series)*, N Engl J Med*., and *Science*, (25 as cover story); the first/co-first or corresponding/co-corresponding author for 68 of them (14 as cover story). Expert reviewer for 10+ international journals in genomics and bioinformatics, including *Nature, Nature Genetics,* and *Nature Biotechnology*.

YONG HOU

2nd F, Building No. 11 | Beishan Industrial Zone | Yantian District | Shenzhen 518083 | China
T: +86-755-22321495 | M: +86-13428739030 | F: +86-755-25037217 | E: houyong@genomics.cn

Work Experience

03/2010-09/2010 **Bioinformatician**, R&D, BGI, Shenzhen, China
09/2010-03/2012 **Project Manager,** Unit of Single Cell Manipulation & Omics, R&D, BGI, Shenzhen, China
03/2013- **Director**, Cancer Research, BGI-Research

Academic activity

10/2012 Single Cell Sequencing and Its Application on Tumor Study **30 min talk** on High Throughput Biology-Genomics &Epigenomics, Cold Spring Harbor Asia, Suzhou, China

05/2012 Single Cell Sequencing @BGI **30 min talk** on symposium of The Era of Next Generation Sequencing in Cancer, Imperial College London, UK

02/2012 Single Cell Sequencing and Its Application on Tumor Study **20 min talk** on the 2nd China-Japan Symposium on Cancer Research, Chiba-shi, Japan

Publications

- **Hou, Y**., Liu, Y., Chen, Z., Gu, N., and Wang, J. (2010). Manufacture of IRDye800CW-coupled Fe_3O_4 nanoparticles and their applications in cell labeling and in vivo imaging. **Journal of nanobiotechnology** *8*, 25.
- **Hou, Y**., Song, L., Zhu, P.,Zhang, B., Tao, Y., et al. (2012). Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. **Cell,** Volume 148, Issue 5, 873-885, 2 March 2012
- Xu, X., **Hou, Y**. *, Yin, X., Bao, L., Tang, A., et al. Single-cell Exome Sequencing to Nucleotide Level Reveals Novel Mutation Characteristics of Clear Cell Renal Cell Carcinoma. **Cell,** Volume 148, Issue 5, 886-895, 2 March 2012
- Li, Y; Xu, X; Song, L; **Hou, Y***; Li, F; Wu, K; Wu, H; Liang, J; Jian, M; Li, J; Zhang, X; Wang, J; Yang, H; Wang, J (2012): Single cell whole-exome sequences of bladder cancer from an individual. **GigaScience**, doi:10.1186/2047-217X-1-12
- Dong, Y, Xie,M, Jiang, Y, Xiao,N, Du,X, Zhang,W, …,**Hou,Y**,…Wang,W, et al.(2012) Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus), **Nature Biotechnology** (2012), Published online 23 December 2012, doi:10.1038/nbt.2478.

*Contribute the same to this work

CURRICULUM VITAE

Name: Xun Xu **Sex:** Male
Date of birth: Apr.3, 1984 **Unit:** Beijing Genomics Institute at Shenzhen
Title: Deputy Director of BGI-Research
Phone: 18688745853 **E-mail:** xuxun@genomics.org.cn
Address: Building NO.11, Beishan Industrial Zone, Yantian District, Shenzhen, China, 518083

Positions

2008–2010 Group leader of the field of plants, Bioinformatics center, Associate Operation Officer of the Science & Technology Department, Beijing Genomics Institute at Shenzhen (BGI-Shenzhen)

2011.1–2012.1 The head of the American Section, Beijing Genomics Institute at Shenzhen (BGI-Shenzhen)

2012.1-now Deputy Director of BGI Research

Education

2003.9--2007.7 Wuhan University, bachelor of biology

2007.9—2010.7 Kunming Institute of Zoology, Chinese Academy of Science, Master of Genetics.

Research & Publications

Xun Xu devoted himself to bioinformatics research, including genome assembly and annotation research for plant genomic, genetic polymorphism research based on the re-sequencing. At the same time, he engaged in explorations and studies on molecular breeding in plants, single cell operation and other experimental techniques, and built the experimental platform for micro-sequencing of single cell. 23 papers were published in high-impact journal including *nature* and *science* until now, he first or co-first author for 11 of them, and have undertaken scientific research projects 8 items, including National Science and Technology department “973” project, “863” project and the Ministry of Agriculture “948” project.

Name: Qiang Pan-Hammarström (Qiang Pan), born in 1970, Jiang Su, China; Swedish citizen

High education degree:
Area, year: Bachelor degree in Medicine, 1993
University: Sun Yat-Sen Medical University (China)

Doctoral degree (PhD):
Area, year: Clinical Immunology/Immunology, 1999-11-26. Supervisor, Prof. Erna Möller.
University: Karolinska Institutet (KI, Sweden)

Postdoctoral work:
2000: Postdoctoral fellow in the Div. of Gastroenterology, Harvard Medical School (USA).
2001-2003: Postdoctoral fellow in the Div. of Clinical Immunology, KI.

Previous positions other than as postdoctoral fellow:
1993-1994: Physician in Guangzhou Respiratory Disease Research Institute (China).
1994-1999: PhD student, Div. of Clinical Immunology, KI.
2004-2007: Research assistant, associate professor, Div. of Clinical Immunology, Dept. of Laboratory Medicine, KI.
2005-2009: Guest professor at the Dept. of Immunology, the School of Basic Medical Sciences, Peking University (China).
2008-2011: Senior researcher, associate professor, Div. of Clinical Immunology, Dept. of Laboratory Medicine, KI.

Current position:
2011-: Professor, group leader in the Div. of Clinical Immunology, Dept. of Laboratory Medicine, KI.
2012-: Visiting Professor, Sun Yat-Sen University Cancer Center (China).
2013-: Scientific advisor, Translational Cancer Research Center, Beijing Genome Institute (BGI, China).
2013-: Senior advisor for Junior Faculty at the Karolinska Institutet.
2013-: Visiting Professor, Rockefeller University (USA).

Supervision of doctoral candidates and postdoctoral researchers
-- Main supervisor for 3 and co-supervisor for 3 doctoral candidates who have completed their PhD thesis. Currently supervising 4 PhD candidates.
-- Main supervisor for 9 and co-supervisor for 3 postdoctoral fellows who have completed their studies. Currently supervising 4 postdoctoral fellows.

Selected distinctions and fellowships
2002: Jonas Söderqvists scholarship for basic research on Virology and Immunology.
2003: Research assistant position, 4 years, awarded by the Swedish Research Council.
2007: Young investigator prize, awarded by the Swedish Society of Medicine.
2009: ERC starting grant award, by European Research Council (ERC).
2010: Senior researcher position, 6 years, awarded by VR.
2011: Distinguished Alumni, Faculty of Medical Sciences, Sun Yat-Sen University.

Funding ID
Current research is supported by research grants awarded by Swedish Research Council, Swedish Cancer Society and ERC.

Scientific interest and productivity
Immunoglobulin gene diversifications, primary immunodeficiencies, DNA repair and recombination, cancer genetics. 84 papers published in these areas. The average impact of the papers published to date is 8.3. H index 27.

Curriculum Vitae

(Current December 2013)

F. Nina Papavasiliou, Ph. D.
Associate Professor
Head of Laboratory of Lymphocyte Biology
The Rockefeller University
1230 York Ave Box 39
New York, NY 10065

Phone: (212) 327-7857
Fax: (212) 327-7319
Email: papavasiliou@rockefeller.edu
http://mutation.rockefeller.edu

Date of Birth: February 6, 1971
Place of Birth: Thessaloniki, Greece
Visa Status: U.S. Permanent Resident

EDUCATION AND TRAINING

			Field of Study
Yale University, New Haven, CT	Ph.D.	1998-2001	Molecular Immunology

Postdoctoral Fellow of the Arthritis Foundation in the laboratory of Dr. David G. Schatz, Section of Immunobiology, Yale University Medical School.

			Field of Study
The Rockefeller University, NY, NY	Ph.D.	1992-1998	Molecular Immunology

Graduate student in the laboratory of Dr. Michel C. Nussenzweig (Ph.D. training). Thesis Title: "The role of B cell Receptor in Allelic Exclusion."

			Field of Study
Oberlin College, Oberlin, OH	B.A.	1988-1992	Biology

B.A. magna cum laude, (1992); Major in Biology and minor in German Literature.

HONORS AND AWARDS (2001-onward)

2011 NIH Director's Transformative Research Award
2009 Vilcek Prize for Creative Promise in Biomedical Research (finalist; Vilcek Foundation)
2007 Award for Outstanding Presentation from the Epigenetics Society (FEBS Workshop, Aussois, FR)
2005 Teaching Excellence Award, The Rockefeller University
2005 G. Jeanette Thorbecke Award, Society for Leukocyte Biology
2004 Sinsheimer Scholar Award
2003 W.M. Keck Foundation Distinguished Young Scholar in Medical Research
2003 Searle Scholars Award

GRANTS AND FELLOWSHIPS

ACTIVE

2013 STARR Foundation (PI:Papavasiliou) " *Role of RNA in Targeting AID to DNA in B Cell Immunity*"
2013 NIH/NIDA, R21 DA036365 (PI: Butelman/Stavropoulos/Papavasiliou) "*T.brucei:* next-generation platform for immunization against drugs of abuse".
2012 NIH/NIGMS, R01 GM084065 (PI: Alfonso/Papavasiliou) "tRNA editing by deamination: balancing affinity and specificity".
2011 NIH/NIAID, "transformative" R01 AI097127 (PI: Papavasiliou) "Building Novel Vaccines on a Borrowed Coat".
2010 NIH/NIAID, R01 AI085973 (PI: Papavasiliou) "Parameters that govern the initiation of VSG switching in T. brucei ".
2009 NIH/NCI, R01 CA098495 (PI: Papavasiliou) " The Regulation of Somatic Hypermutation"

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Methods for identification and analysis of non-coding driver elements

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Jakob Skou Pedersen[1,2] **& Asger Hobolth**[2]. 1) Department of Molecular Medicine, Faculty of Health, Aarhus University, Denmark. 2) Bioinformatics Research Centre, Faculty of Science and Technology, Aarhus University, Denmark. *NB*: After personal communication with Josh Stuart and Jan Korbel, we are submitting this proposal in the hope we can become part of the consortium despite not being at an ICGC or TCGA member institution.

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Malene Juul Rasmussen & Henrik Hornshøj. Department of Molecular Medicine, Aarhus University, Denmark.	**Torben Ørntoft.** Department of Molecular Medicine, Aarhus University, Denmark.

Background and preliminary data

In the last few years, point mutations (SNVs) and indels have been cataloged across thousands of cancer exomes, allowing systematic mutational screens of protein-coding (pc) driver genes. In general, the screens identify pc genes with surprisingly many mutations of expected functional impact compared to a null model. This approach has been extremely successful and rapidly extended the set of known driver genes as well as facilitated the molecular sub-classification of cancer types.

However, this approach has been blind to the role of regulatory elements outside pc genes, including most non-coding (nc) RNAs, in cancer development. With the advent of large sets of whole cancer genomes, it in principle becomes possible to extend the systematic mutational screens to also include nc elements. However, the existing methods for protein-coding genes assume well-annotated genetic elements of known extent and exploit that the expected functional impact of mutations can be evaluated based on the rules of splicing and translation. These assumptions do not hold in general for nc elements, which are functionally heterogeneous and often poorly annotated. There is therefore a need for new methodology tailored for nc elements, before general mutational screens can be conducted.

We propose to develop such methods for identifying nc elements with surprisingly (i.e. significantly) many mutations of functional impact and to apply these to screen the pan-cancer genomes for nc driver elements. We will both develop generic methods, which make few assumptions about the nature and extent of the functional elements, and more specific methods, tailored for individual classes of nc elements, such as functional RNA structures. In the lack of precise functional understanding of the elements, we will use evolutionary conservation, and comparative genomics more generally, to evaluate the functional impact of mutations. In the lack of precise annotations, we will identify mutational hotspots using statistical techniques capable of segmenting the genome (Hidden Markov Models (HMMs) and generalized random walks). For classes of well-annotated elements, we will apply significance evaluation methods resembling those used for pc genes, exploiting what is known about element function to evaluate mutational impact.

We have previously (1) analyzed the mutational landscape across whole cancer genomes (in revision at Cancer Cell); (2) developed a method for identification of mutational hotspots based on BLAST-like random walk theory, which was prototyped on exome SNV data, where it efficiently identified known oncogenes with clustered mutations, such as KDM6A; and (3) adapted MutSigCV (Lawrence et al., 2013) to identify regions with surprisingly many mutations in conserved elements, now being tested on WGS data (Alexandrov et al., 2013).

Both JSP and AH (PIs) have worked with statistical sequence analysis, evolutionary modeling, and comparative genomics for more than 12 years and cancer genomics for several years. JSP has developed one of the most cited methods for comparative RNA structure identification (EvoFold); contributed nc RNA analysis published in high impact journals (Nature, Cell, and elsewhere); and is currently involved in improving RNA structure maps genome-wide using experimental data with a recently developed method (ProbFold, in review).

Timelines & resources dedicated to project

JSP has a young group-leader grant from The Danish Research Council for identifying and characterizing nc RNAs and other nc elements in cancer by integration of multiple genomics resources. Becoming part of the WGS pan-cancer consortium would be a perfect fit and greatly facilitate the goals of the project. The project is dependent on access to WGS SNV and indel calls. The analysis would benefit from integration of expression and methylation data. The main milestones are: Jan 2014, first predictions on pilot data; August 2014, all new methods completed; Nov. 2014, predictions and interpretations completed; Feb. 2015, main contributions and companion papers completed.

The time dedication to the project is as follows JSP (50%), AH (33%), MJR (100%), HH (50%), nn post doc. with start spring 2014 (75%). Note that HH is a senior post doc. titled as ass. prof. and part of JSP's group. We have extensive local computational infrastructure, including a 2K-node cluster and a local public mirror of the UCSC GB database. If relevant, the department, represented by collaborator TØ, will contribute experimental and clinical validation of findings. The department houses an internationally unique biobank with thousands of tumor sample specimens from prostate, bladder, and colon cancer and employs more than 50 scientist and PhD students working with clinical and experimental aspects of cancer genomics and functional genomics.

Research proposal

We aim to improve the identification of nc driver elements in cancer, with a special focus on ncRNAs. Our proposal involves a tight connection between method development and application. Following the strategies outlined above, we will pursue the following two types of approaches:
Generic methods: These will make few assumptions on type of function and extent of the nc elements and work by screening through the genome, evaluating significance of SNV and indel observations across samples. Inspired by MutSigCV and other methods, we will model variation in mutational intensities at several levels: between samples, between context-dependent mutation types, and along the genome. A benefit of WGS analysis, compared to exome analysis, is that much more data is available to learn and model the mutational variation. We will use statistical methods similar to principle component analysis (PCA) to orthogonally decompose the dependencies on covariates along the genome. Similarly, we will explicitly model the mutational signatures present in individual samples. We will use evolutionary conservation as a proxy for functional importance and mutational impact. We will implement these strategies using both a random walk approach (prototype in place) and a HMM approach (under development).
Class-specific methods: These will take a more focused strategy, representing more specific hypotheses and likely have more power when applicable. Here we will (1) focus on classes of elements with known functions or (2) specifically evaluate sets of externally defined candidate elements. A central focus will be on ncRNAs and structural RNAs. 1) For individual ncRNAs, we will evaluate if structure-disrupting mutations are significantly enriched compared to our null models from above. Likewise, we will ask if mutations in miRNA binding sites or binding sites of RNA binding proteins are significantly enriched across individual ncRNAs (or UTRs), using known motif constraints and site-specific evolutionary substitution patterns to evaluate the functional impact of mutations. 2) We have adapted MutSigCV for nc elements (under evaluation), where conserved regions take the place of non-synonymous sites. This method will be applied to individual nc candidate elements. We will start by applying it to sets of lncRNAs that we shortlisted for their functional evidence using comparative and functional genomics (Nielsen et al, 2013). Similarly, we will specifically evaluate ncRNAs identified as differentially expressed in cancer (in-house or from the consortium), GWAS cancer susceptibility regions, etc.. Finally we will focus on the regulatory regions associated with known pc cancer driver genes. We will evaluate the robustness of our predictions using false discovery rate evaluation techniques and bootstrapping.

We expect the systematic application of the described methods on the extensive WGS pan-cancer data set will make significant contributions to the discovery of nc driver elements, which still remains sporadic. The identified elements will be characterized based on existing annotations, functional genomics data, and evolutionary analysis, with the option of both clinical and experimental local follow-up.

Legacy plans

We will make the developed methods, including source code, freely available to the academic community. We will make any genome-wide predictions available through our local UCSC mirror or the platform of choice by the consortium. Finally we will ensure reproducibility by releasing virtual machine images with input data, source code, binaries, scripts, outputs, and the necessary documentation to rerun and reproduce the published results.

Jakob Skou Pedersen

Associate Professor (Lektor)
Department of Molecular Medicine
Aarhus University, Denmark

PERSONAL INFORMATION

Born: 23 April 1975
Citizenship: Danish
Married, two children

CONTACT INFORMATION

Department of Molecular Medicine, Aarhus University Hospital, Skejby, Brendstrupgaardsvej 100, DK-8200 Aarhus N, Denmark
Phone: (+45) 8949 9412
E-mail: jakob.skou@ki.au.dk

RESEARCH INTERESTS

Non-coding RNA; Cancer Genomics; Gene regulation; Comparative Genomics; Statistical Modeling

RESEARCH GROUP

Current group: 2 senior researchers, 2 post docs (+ open position), 3 PhD students, 2 master students and 1 AC TAP.

EDUCATION

Aarhus University, Aarhus, Denmark
Ph.D., Biology (Bioinformatics), September 2004
Advisors: Jotun Hein and Freddy Bugge Christiansen
Dissertation: "Structured models of molecular evolution"

Aarhus University, Aarhus, Denmark
M.S., Biology (Bioinformatics), June 2001
Advisor: Jotun Hein
Dissertation: "Comparative Gene Finding"

PROFESSIONAL EXPERIENCE

Aarhus University, Aarhus, Denmark
Associate Professor (Lektor) at Department of Molecular Medicine **2010-present**

University of Copenhagen, Copenhagen, Denmark
Assistant Professor (Adjunkt) at The Bioinformatics Center **2007-2010**

University of California, Santa Cruz, CA, USA
Postdoctoral Fellow at Center for Biomolecular Science and Engineering **2004-2007**
Advisor: David Haussler

Aarhus University, Aarhus, Denmark
Teaching Assistant and occasionally *lecturer* **1999-2003**

SCIENTIFIC VISITS

University of Oxford, Oxford, United Kingdom
Visits Jotun Hein's group as part of Ph.D. studies, September 2002 - August 2003

TEACHING

Taught and co-taught courses and workshops on comparative genomics, phylogenetics, population genetics, Next Generation Sequencing, sequence analysis, statistical analysis, and non-coding RNA analysis.

PUBLICATIONS OVERVIEW

Publications in international scientific journals: 24
Publications in Nature: 6 (one shared first-author, three consortium papers)
Publications in Cell: 1 (second-author)

CITATIONS

First/last author citations: 646 (ISI)
Total citations: 6,000 (ISI)
h-index: 18 (ISI)

RECENT AWARDS AND RESEARCH SUPPORT

Project: "Regulatory networks of large intergenic non-coding RNAs in health and disease"
Source: DFF Sapere Aude group leader grant (#12-126439)
Amount: 7,034,354 DKK; Period: Jan. 2013 - Dec. 2016

Project: "Center for Computational and Applied Transcriptomics (COAT)"
Source: Danish strategic research council (Fellowship; #10-092320/DSF)
Amount (co-PI): 2,379,548 DKK; (Project total: 35,700,000 DKK) Period: Apr 2011 – Sep 2016

Project: "Novel medicines for treatment of cancer and neurological disorders"
Source: Advanced Technology Foundation (# 2008-2)
Amount (co-PI): 1,417,500 DKK (Project total: 43,367,666 DKK); Period: Sep. 2008 - Aug 2011

Curriculum Vitae for Asger Hobolth (November 2013)

Personal information

Name: Asger Hobolth. Born: September 16, 1972 in Århus, Denmark

Short biography

Asger Hobolth is Associate Professor in the Bioinformatics Research Center (BiRC) at Aarhus University. After finishing his PhD in statistical shape analysis in 2002, Asger Hobolth has worked in evolutionary genomics and bioinformatics. His main research area is statistical methods and probability models for analysing the evolution of DNA sequences. In particular he has formulated models and inference methods for taking context dependence and incomplete lineages sorting into account, and he is one of the pioneers in the emerging field of population genomics. Asger publishes as first author in both high-profile statistics and genomics journals, and collaborates with theoretical statisticians and computational biologists. Furthermore he has been a member of several international ape genome consortia, including the Orangutan, Gorilla, Bonobo and Great Apes projects (pulications in Nature in 2011, 2012, 2012 and 2013).

Employments

2010-present	Associate professor at BiRC, Aarhus University, DK
2008-2010	Assistant Professor at BiRC, Aarhus University, DK
2007-2008	Assistant Professor at North Carolina State University, USA
2005-2007	PostDoc at North Carolina State University, USA
2002-2005	PostDoc at BiRC, Aarhus University, DK

Education

1998-2002	PhD studies at Mathematics Institute, Aarhus University, DK
2000	3 months reseach stay at Brown University, USA
1999	6 months research stay at Leeds University, England

Other profesional activities

2008-present	Associate editor of Molecular Biology and Evolution

5 most important publications

Hobolth, A. and Jensen, J.L. (2011). Summary statistics for endpoint-conditioned continuous-time Markov chains. *Journal of Applied Probability*, **48**, 911-924.

Hobolth, A., Dutheil, J., Hawks, J., Schierup, M., and Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, **21**, 349-356.

Hobolth, A. (2008). A Markov Chain Monte Carlo Expectation Maximization algorithm for statistical analysis of DNA sequence evolution with neighbour-dependent substitution rates. *Journal of Computational and Graphical Statistics*, **17**, 138-164.

Hobolth, A., Christensen, O.F., Mailund, T. and Schierup, M.H. (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics*, **3**, 294-304.

Hobolth, A. and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in Genetics and Molecular Biology*, **4**, 18.

Publication record

Articles: 41; Monographs: 0; Book chapters: 3; Proceedings: 3 Other: 5.

MALENE JUUL RASMUSSEN

Malene Juul Rasmussen
Dept. of Molecular Medicine (MOMA)
Aarhus University Hospital
Brendstrupgaardsvej 100
8200 Aarhus N
Denmark

T + 45 784 55372
E malene.juul.rasmussen@ki.au.dk

Profile

Currently I work as a Ph.D. student in Bioinformatics at the Department of Molecular Medicine (MOMA) at Aarhus University Hospital in Denmark. I have a M.Sc. in statistics and a B.Sc. in mathematics from Aarhus University. Furthermore, I have three years of experience working as a statistical consultant in the Danish agricultural industry.

Professional experience

Ph.D. student, Aarhus University Hospital; Aarhus, Denmark **2013-present**

I am currently working on developing a statistical method for detecting mutational hotspots in genome-wide SNV data sets, including regulatory non-coding regions.

Innovation Consultant, AgroTech; Skejby, Denmark **2010-2013**

As a statistical consultant I primarily worked with statistical programming, data analysis, data presentation and design of experiments. Furthermore I had responsibility for writing funding applications, handling client contact and I also had the role as project manager on smaller projects.

Student teacher, Aarhus University; Aarhus, Denmark **2009-2010**

In this job I taught basic statistics to biology students at university.

Publications

The pattern of genomic instability during bladder cancer progression. Nordentoft, I., Lamy, P., Birkenkamp-Demtröder, K., Villesen P., Shumansky, K., Vang, S., Hornshøj, H., Rasmussen, M.J., Hedegaard, J., Thorsen, K., Høyer, S., Borre, M., Fristrup, N., Dyrskjøt, L., Shah, S., Pedersen, J.S., Ørntoft, T.F. (In revision at Cancer Cell).

Education

UC Berkeley Extension, California, USA: IDP in Project Management, 2013.
University of Auckland, NZ: Study abroad, Dept. of Statistics, 2012.
Aarhus University, Denmark: M.Sc. in statistics, 2010.
Aarhus University, Denmark: B.Sc. in mathematics, 2008.

Skills

Skills include but are not limited to **Statistics:** Monte Carlo based simulation methods, time series analysis, design of experiments, multivariate analysis, survival analysis. **Computer:** Structured statistical programming in R and SAS. **Business:** Project planning, project management, presentation skills. **Language:** Danish (native), English (fluent), German (basic), Spanish (basic).

Henrik Hornshøj

Nørregade 14, 8860 Ulstrup, Denmark
Phone: +45 51 92 15 12
E-mail: henrik.hornshoj@gmail.com
Website: www.hornshoj.net

Employment history

Associate Professor (2013 – current)
Department of Molecular Medicine (Aarhus University Hospital)
Post doc (2008 – 2013)
Department of Molecular Biology (Aarhus University)
Research Assistant (2002 – 2005)
Department of Genetics and Biotechnology (Aarhus University)
Bioinformaticist (2001)
Helixense, Singapore
Technical Coordinator (2000 – 2001)
Aventis CropScience, Copenhagen

Education

Ph.D Bioinformatics (2008)
Microarray platform development and analysis of tissue gene expression
M.Sc Molecular Biology (1999)
Receptor-like kinases in plant defense response

Research activities

Studies of mutations in cancer genomes by gene selection pressure, significant mutations in protein-coding genes and non-coding elements and identification of clusters and genomic mutation hotspots. System-wide integrative analysis of genetic variations affecting gene expression identified by next generation sequencing (NGS) and genome-wide detected sequence polymorphisms.

Areas of expertise

- Bioinformatics, Molecular Biology, High-throughput Data, Integrative Analysis, Method Development, Programming, Course Teaching, Ph.D supervision

Awards and funding

- DFF Sapere Aude Young Elite Scientist Award (2011), DFF Three-year Post Doc grant (2010), Aarhus University Research Foundation Ph.D Award (2009)

Publications

- 21 Peer Reviewed Research Publications

CV of Professor Torben F. Ørntoft
November 2013

Education/career

2008 -	Head of the Dept. Molecular Medicine, Aarhus University Hospital
2007 -	Centre Leader in Lundbeck foundation Res. Centre.
2004 -	Professor of Clinical Biochemistry at Aarhus University
2000 - 2008	Chairman, Dept. of Clinical Biochemistry, University of Aarhus, Skejby
1999 - 2004	Professor of Molecular Cancer Diagnostics/Clinical Biochemistry, Univ. of Aarhus
1996 - 2008	Chief Physician, Dept. of Clinical Biochemistry, Aarhus University Hospital, Skejby
1989-95	Education and training as Clinical Biochemist, Hospitals in Aarhus County
1984-89	Research Fellow, University of Aarhus and Danish Cancer Society
1981-84	Basic Clinical Education in Surgery, Internal Medicine, and Gynecology
1981	MD, University of Aarhus

Academic degrees

1990	Dr. med. Sci. thesis, University of Aarhus
1986	Gold Medal Award, University of Aarhus

Memberships, chairmanships, leadership

2011 -	European Academy of Cancer Sciences
2010 -	Member of the scientific advisory board, Novo Nordic Proteome Centre Copenhagen University
2009-	Chairman of the board at AROS
2007-2010	Member of the Scientific Board at WHO´s IARC in Lyon
2005-2008	Member of the board of East Jutland Innovation A/S, a public venture fund
2005-	Head of Nordic Centre of Excellence in Molecular Medicine (Danish Branch)
2002-2009	Head of the Danish Research School in Molecular Cancer Research
2001-	Member of expert committee on bladder cancer established by the NCI, USA
2001 -	Grant proposal consultant within the cancer area for the Finnish Academy and the Dutch Cancer Society
2000-2008	Director and co-founder of AROS Applied Biotechnology ApS

Editorial

2007-	Member of the editorial board Molecular Oncology
2004-	Associate Editor, Cancer Research
2001-	Member of the Editorial Board, Molecular and Cellular Proteomics
1994-	Member of the Editorial Board of Glycoconjugate Journal

Honours

2011	European Academy of Cancer Sciences
2010	Annual Research Award, Danish Society for Oncology
2005	General Consul Ernst Carlsens Fond honorary award for medical research
2002	Georg Gotfred Holmbach Torps honorary award for outstanding cancer researcher

Publications, invited lectures and other scientific activities

More than 300 scientific publications in international magazines, as Nature, Nature Genetics, New England J Medicine, Cancer Cell, Blood, Cancer research, JNCI, Journal of Clinical Investigation etc.

Invited lectures at international symposia, conferences and workshops.
Numerous professor-, associate Professor- and Ph.D. assessments.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Regulators of telomere length and composition, and of TERRA expression in 2000 cancer samples

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

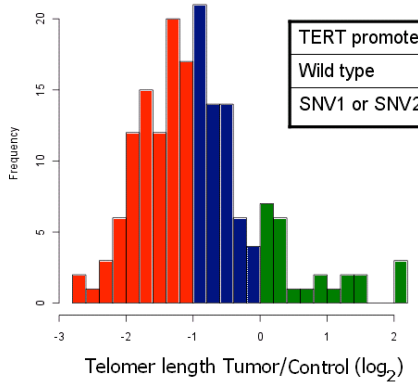
Benedikt Brors, Computational Oncology, Div. Theoretical Bioinformatics, DKFZ Heidelberg, Germany
(PI in: ICGC Prostate Cancer, ICGC Malignant Lymphoma, ICGC Pediatric Brain Tumors)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Lars Feuerbach, Computational Oncology, DKFZ; David T. W. Jones, Pediatric Neurooncology, DKFZ	Karsten Rippe, Genome Organization and Function, DKFZ Heidelberg

Background and preliminary data

Background A critical step in carcinogenesis is the development of mechanisms to escape the Hayflick-limit. This decouples the number of possible cell divisions from the limited length of telomeres by different mutational pathways. Some of these mechanisms are characteristic for particular cancer subtypes, e.g. the overrepresentation of recurrent SNVs in the hTERT promoter in medulloblastoma tumors of the SHH subtype [1-3]. Whole-genome sequencing (WGS) experiments produce data on the quantity, composition and location of telomere repeat (TTAGG) containing sequences. Similarly, RNA-sequencing data can be analyzed with respect to telomere-repeat containing RNA (TERRA, [4]). Estimations of telomere length derived by computational analysis of WGS data correlate with biological assays [5]. A number of covariates such as patient age convolve with tumor-specific alterations in determining telomere length [5]. Furthermore, mechanisms such as alternative telomere lengthening impact upon telomere composition [6]. For cancer types with inherently reduced tumor purity, such as prostate cancer, additional normalization is required. This is comparable to our previous work on tumor purity-adjusted mutation analysis [7].

Preliminary Data We implemented a computational pipeline for the identification of sequencing reads that contain motifs indicative for telomeres or alternative telomere lengthening in matched tumor/control pairs. These reads were stratified by their chromosomal location. Telomere length estimations were computed for 173 WGS medulloblastoma tumor/control pairs. A significant correlation between telomere length and the presence of somatic hotspot mutations in the TERT promoter were found.



TERT promoter	Strong attrition	Weak attrition	Lengthening
Wild type	84	51	21
SNV1 or SNV2	4	8	5

Significance:
Fisher Exact Test: 0.029
Chi-square : 0.043

Figure 1: Telomere statistics for 173 medulloblastoma samples and their association with known regulatory SNVs in the hTERT promoter (SNV1 or SNV2 [8]).

Timelines & resources dedicated to project

Timeline: Implementation of computational pipeline (done); Computation of telomere statistic for 2000 tumor/control WGS data (2 months); Identification of telomere repeat-containing RNA (TERRA) in ca. 1000 mRNA-seq datasets (1 month); Correlation of telomere length and composition to covariates and mutations (3 months).

Resources: 1 post-doc and 1 graduate student; 2000-core compute cluster (4x12core CPUs, 64-1024 GB RAM, 6 petabytes storage); wet-lab facilities and expertise for functional validation of findings

Research proposal

Measurement of telomere length and telomere composition from whole-genome sequencing data enables a targeted screening for mutations that contribute to cancer cells escaping the Hayflick limit.

The first aim of this project is to use a coherent dataset of 2000 tumor/control pairs to adjust an existing computational pipeline for determining telomere length from occurence of telemore repeat-containing sequences. We will implement adjustment for factors that influence telomere length measurements. Germline genome data will be used to identify the impact of covariates such as smoking habits and age. This step will result in unbiased (Figure 1) and normalized telomere statistics for each analyzed cancer type.

In a second step the computed telomere statistics and TERRA expression levels will be correlated to genes that are affected by somatic point mutations (coding or potentially regulatory), copy-number variations, structural variations and altered promoter methylation levels. To increase the power, the analysis will first be restricted to pathways that are implicated in telomere and alternative telomere lengthening, or that are frequently mutated in cancer. Lists that rank genes and recurrent aberrations by their putative impact on telomere length will be produced. In a third step significant associations will be stratified by tumor type and patient history to differentiate between subtype-specific and general events.

References

[1] Remke et al. TERT promoter mutations are highly recurrent in SSH subgroup medulloblastoma. Acta Neuropathol. (2013) 126: 917-929.
[2] Killela et al. TERT promoter mutations occure frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. USA (2013) 110:6021-6026.
[3] Tallet A et al. Overexpression and promoter mutation of the TERT gene in malignant pleural mesothelima. Oncogene (2013) [epub ahead of print].
[4] Luke, Lingner. TERRA: telomeric repeat-containing RNA. EMBO J. (2009) 28: 2503-2510.
[5] Parker et al. Assessing telomeric DNA content in pediatric cancer using whole genome sequencing. Genome Biol. (2012) 12:R113.
[6] Conomos et al. Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. J. Cell Biol. (2012) 199:893-906.
[7] Weischenfeldt*, Simon*, Feuerbach*, et al. Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer. Cancer Cell (2013) 23:159-170.
[8] Huang et al. Highly recurrent TERT promoter mutations in human melanoma. Science (2013) 339:957-959.

Legacy plans

We will make available the complete computational pipeline including the deconvolution model for further analysis. We will also provide data on telomere statistics and TERRA expression profiles within the entire pan-cancer data set for further exploration. For this, we will create a web-based database as well as downloadable data objects. Significant associations between telomere statistics, mutation profiles and tumor types/subtypes will be provided in an explorative, web-based tool that shows telomere-based groupings together with other sample-wise results from the ICGC PanCancer analysis.

Dr. Benedikt Brors

Group Leader Computational Oncology
Div. Theoretical Bioinformatics
German Cancer Research Center, Heidelberg

E-Mail: b.brors@dkfz.de

Degree

1989–1995	Diploma in chemistry, University of Düsseldorf, Germany
1999	Doctoral degree (Dr. rer. nat) in biochemistry, University of Düsseldorf

Scientific Career

1995–1999	Pre-doctoral research assistant, Inst. of Biochemistry, University of Düsseldorf
1999–2000	Postdoctoral Researcher, Div. Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
2000–2003	Postodoctoral Researcher, Intelligent Bioinformatics Systems, DKFZ
2003–present	Group Leader Computational Oncology, Div. Theoretical Bioinformatics, DKFZ
2003–present	Lecturer in bioinformatics, Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg
2007	Offer of a post as Full Reader in Medical Bioinformatics, Queen's University, Belfast, UK (not realized)
2008	Offer of a post as W2 professor of biostatistics (non-tenured), University Medical Center Göttingen, Germany (not realized)
2009–present	Tenured research position, DKFZ
2013	Offer of a position as full professor of applied bioinformatics, DKFZ and National Center for Tumor Diseases, Heidelberg (under negotiations)

Selected Publications

Jones, D.T.W.,* **Hutter, B.*, Jäger, N.***, (...), **Brors, B.**, (...), Eils, R., Lichter, P., Pfister, S.M., 2013. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet.** 45, 927-932

Weischenfeldt, J.*, Simon, R.*, **Feuerbach, L.***, (...), Sültmann, H.#, Sauter, G.#, Plass, C.#, **Brors, B.#**, Yaspo, M.-L.#, Korbel, J.O.#, Schlomm, T.#, 2013. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. **Cancer Cell** 23, 159–170.

Richter, J., **Schlesner, M.***, (...), Hummel, M#., Klapper, W.#, Rosenstiel, P.#, Rosenwald, A.#, **Brors, B.#**, Siebert, R.#, 2012. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. **Nat. Genet.** 44, 1316–1320.

Jones, D.T.W.*, **Jäger, N.***, (...), **Brors, B.**, (...), Eils, R., Pfister, S.M., Lichter, P., 2012. Dissecting the genomic complexity underlying medulloblastoma. **Nature** 488, 100–105.

Oberthuer, A., Hero, B., Berthold, F., **Juraeva, D.**, (...), **Brors, B.**, Fischer, M., 2010. Prognostic impact of gene expression-based classification for neuroblastoma. **J. Clin. Oncol.** 28, 3506–3515.

(* contributed equally; # contributed equally as senior authors)

Relevant Projects

- PI (bioinformatics analysis) in ICGC-PedBrain
- PI (bioinformatics analysis) in ICGC – Early Onset Prostate Carcinoma
- PI (bioinformatics analysis) in ICGC – Molecular Mechanisms in Malignant Lymphoma
- Member of ICGC Bioinformatics Work Group and Mutation Consequences and Pathways WG

Lars Feuerbach – Curriculum Vitae

Phone: +49 6221 42 3603 E-mail: l.feuerbach@dkfz.de

RESEARCH POSITIONS

10/2011-	Research Fellow – ICGC Early onset Prostate Cancer Computational Oncology, Theoretical Bioinformatics German Cancer Research Center, Heidelberg, Germany
10/2007-09/2011	Graduate Research Assistant Computational Biology and Applied Algorithms Department Max-Planck Institut für Informatik, Saarbrücken, Germany

EDUCATION

10/2007 –	PhD in Bioinformatics (Thesis submitted – 08/2013) Max-Planck Institut für Informatik, Saarbrücken, Germany
10/2005 – 09/2007	Master of Science (MSc) with Honor's Degree - Bioinformatics Center for Bioinformatics, University of Saarland, Germany
09/2002 – 07/2005	Bachelor of Science (BSc) - Bioinformatics Free University of Berlin, Germany

SELECTED PUBLICATIONS

Joachim Weischenfeldt*, Ronald Simon*, Lars Feuerbach*, Karin Schlangen*, et al.
Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer
Cancer Cell, 2013, 23(2):169-170

Lars Feuerbach, Konstantin Halachev, Yassen Assenov, Fabian Müller, Christoph Bock, Thomas Lengauer
Analyzing epigenome data in context of genome evolution and human diseases
Methods Mol. Biol. 2012,856:431-67

Malay Bhattacharyya, Lars Feuerbach, Tapas Bhadra, Thomas Lengauer, Sanghamitra Bandyopadhyay
MicroRNA Transcription Start Site Prediction with Multi-objective Feature Selection
Statistical Applications in Genetics and Molecular Biology, 2012, 11(1) 1–25

Lars Feuerbach, Rune B. Lyngsoe, Thomas Lengauer, Jotun Hein
Reconstructing the ancestral germline methylation state of young repeats
Molecular biology and evolution 2011;28(6):1777-84

Pavlo Lutsik, Lars Feuerbach, Julia Arand, Thomas Lengauer, Jörn Walter, et al.
BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing.
Nucleic Acids Research, May 11, 2011, 39(Web Server issue):W551-6

David T. W. Jones, PhD - Curriculum Vitae

Nationality: British **Phone:** +49 6221 424594 **E-mail:** davidjones@cantab.net

Career

09/10 – Post-Doctoral Scientist, Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
07/09 – 07/10 Post-Doctoral Research Fellow, Dept. of Pathology, University of Cambridge
09/05 – 06/09 Studying for the degree of PhD, University of Cambridge
09/04 – 06/05 Research Assistant, Dept. of Pathology, University of Cambridge

Education

2005-2009 Department of Pathology, University of Cambridge, UK – PhD in Molecular Genetics
2001-2004 Clare College, University of Cambridge, UK – B.A. (Hons) in Natural Sciences

Selected Publications

Jones DTW*, Hutter B*, Jäger N* *et al*. and Eils R, Lichter P & Pfister SM
Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma
Nat Genet 2013 Aug;45(8):927-32

Reuss DE*, Piro RM*, Jones DTW* et al. and Pfister SM & von Deimling A
Secretory meningiomas are defined by combined KLF4 K409Q and TRAF7 mutations
Acta Neuropathol 2013 Mar;125(3):351-8

Northcott PA, Jones DTW, Kool M, Robinson GW, Gilbertson RJ, Cho YJ, Pomeroy SL, Korshunov A, Lichter P, Taylor MD, Pfister SM
Medulloblastomics: the end of the beginning
Nat Rev Cancer 2012 Dec;12(12):818-34

Sturm D, Witt H, Hovestadt V, Khuong Quang D-A, Jones DTW, *et al* and Plass C, Jabado N & Pfister SM
Distinct hotspot mutations define epigenetic and biological subgroups of glioblastoma
Cancer Cell. 2012 Oct 16;22(4):425-37

Jones DTW*, Jäger N* *et al*. and Eils R, Pfister SM & Lichter P
Dissecting the Genomic Complexity Underlying Medulloblastoma
Nature 2012 Aug 2;488(7409):100-5

Schwartzentruber J, Korshunov A, Liu XY, Jones DTW, *et al* and Pfister SM, Jabado N
Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma
Nature. 2012 Jan 29;482(7384):226-31

Rausch T*, Jones DTW*, Zapatka M*, Stütz AM* *et al*. and Lichter P, Pfister SM, Korbel JO
Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations.
Cell 2012 Jan 20;148(1-2):59-71

Ichimura K, Pearson DM, Kocialkowski S, Bäcklund LM, Chan R, Jones DTW & Collins VP
IDH1 mutations are present in the majority of common adult gliomas but rare in primary glioblastomas.
Neuro Oncol 2009 Aug;11(4):341-347

Jones DTW, Kocialkowski S, Liu L, Pearson DM, Bäcklund LM, Ichimura K & Collins VP
Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas.
Cancer Res 2008 Nov;68(21):8673-8677

PD Dr. Karsten Rippe

(*1964)

E-Mail: Karsten.Rippe@dkfz.de
phone: 06221-54-51376
http://malone.bioquant.uni-heidelberg.de

German Cancer Research Center (DKFZ)
and BioQuant Institute
Research Group Genome Organization
& Function
Im Neuenheimer Feld 280
69120 Heidelberg

Biographical sketch

Head of Research Group Genome Organization & Function at the DKFZ	since 2007
Group leader, Kirchhoff-Institut für Physik, Germany	2001 – 2007
Habilitation, University of Heidelberg, Germany	2000
Scientist at the DKFZ, Germany	1994 – 2001
Postdoctoral fellow, University of Oregon, Eugene, USA	1992 – 1994
PhD, University of Göttingen, Germany	1989 – 1991

Research profile and selected publications

Karsten Rippe is conducting interdisciplinary research that combines molecular/cell biology and physics investigate the relation between chromatin states like heterochromatin and telomeric chromatin and cellular functions. In the area of telomere biology he focuses on the alternative lengthening of telomeres (ALT) mechanism active in some cancer types. The Rippe group is

- dissecting the ALT mechanism by applying a high-content imaging-based analysis in conjunction with deep sequencing methods (RNA-seq, ChIP-seq)
- conducting cell biology studies to elucidate the dynamics and structure of PML nuclear bodies and their complexes with telomeres, and
- elucidating the mobility of telomeres in an ALT-positive human cell line.

Chung, I., Osterwald, S., Deeg, K. & **Rippe, K.** (2012). PML body meets telomere: The beginning of an ALTernate ending. *Nucleus* 3, 263-275.

Osterwald, S., Wörz, S., Reymann, J., Sieckmann, F., Rohr, K., Erfle, H. & **Rippe, K.** (2012). A three-dimensional colocalization RNA interference screening platform to elucidate the alternative lengthening of telomeres pathway. *Biotechnol. J.* 7, 103-116.

Chung, I., Leonhardt, H. & **Rippe, K.** (2011). De novo assembly of a PML nuclear subcompartment occurs through multiple pathways and induces telomere elongation. *J. Cell Sci.* 124, 3603-3618.

Lang, M., Jegou, T., Chung, I., Richter, K., Udvarhelyi, A., Münch, S., Cremer, C., Hemmerich, P., Engelhardt, J., Hell, S. W. & **Rippe, K.** (2010). Three-dimensional organization of PML nuclear bodies. *J. Cell Sci.* 123, 392-400.

Jegou, T., Chung, I., Heuvelmann, G., Wachsmuth, M., Görisch, S. M., Greulich-Bode, K., Boukamp, P., Lichter, P. & **Rippe, K.** (2009). Dynamics of telomeres and promyelocytic leukemia nuclear bodies in a telomerase negative human cell line. *Mol Biol Cell* 20, 2070-2082.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Effect of non-coding somatic mutation in CpG Islands and regulatory elements on gene expression

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Benedikt Brors, DKFZ Heidelberg

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

- Carl Herrmann, DKFZ Heidelberg
- Lars Feuerbach, DKFZ Heidelberg

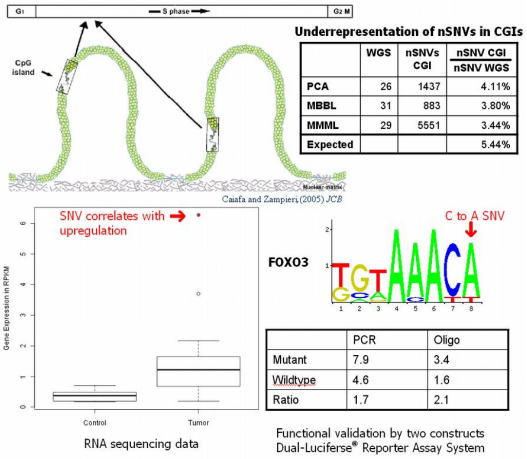
Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

- Ewan Birney (EBI-EMBL Hinxton)

Background and preliminary data

Biological functional CpG islands are genome regions that show low levels of DNA methylation, and are enriched in histone marks of open chromatin and frequently occupied polymerase II binding sites. They are co-localized with origins of replication and are passively protected from methylation induced cytosine-deamination, and thus, display a reduced background mutation rate. Due to the high accessibility for trans-factors, non-coding somatic point mutations (nSNVs) that create or disrupt transcription factor binding sites in these regions have an elevated probability to convey biological function.

Preliminary data: 86 nSNVs datasets produced by the DKFZ pipeline from three different cancer entities were overlapped with a high resolution CpG island annotation [1]. The analysis confirmed a 63-75% underrepresentation of nSNVs. In the ICGC Prostate cancer dataset 36 nSNVs fell into the strongest category of CpG islands (CGI cores) and were located in a gene promoter. One of these nSNVs correlated with an upregulation of the adjacent oncogene and generated a FOXO3 binding site. Functional validation by a luciferase assay confirmed that the nSNV induces a 2-fold upregulation in a cell line.
[1]http://cgihunter.bioinf.mpi-inf.mpg.de/



	WGS	nSNVs CGI	nSNV CGI / nSNV WGS
PCA	26	1437	4.11%
MBBL	31	883	3.80%
MMML	29	5551	3.44%
Expected			5.44%

	PCR	Oligo
Mutant	7.9	3.4
Wildtype	4.6	1.6
Ratio	1.7	2.1

Functional validation by two constructs Dual-Luciferse® Reporter Assay System

Timelines & resources dedicated to project

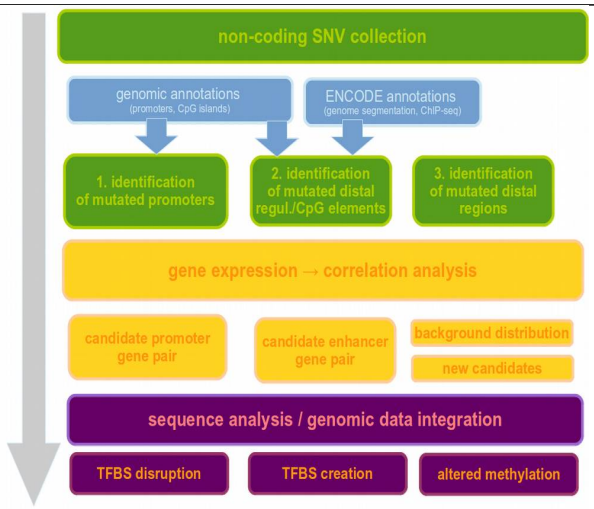
Human Resources:
- 2 Junior investigators (C. Herrmann & L. Feuerbach)
- 1 PhD Student (C. Chan, starting December 2013)
- collaboration with DKFZ group (C. Plass) for functional validation

Timeline
- nov. 2013 - Q1 2014: pilot study on focused in-house datasets (SNV & RNA-seq)
- Q1 2014 : identification of mutational hotspot for all entities (as they become available)
- Q2-Q4 2014: correlation analysis with expression data.

Research proposal

To maximize power, the impact of nSNVs on gene expression is analysed in a 3 step approach:

1. **focused approach**: Analysis of nSNVs in gene promoters. Mutations that correlate with gene expression changes of adjacent genes are analyzed for impact on TFBS.
2. **annotation driven approach**: Analysis is restricted to potential enhancer regions (cis-effect) and non-coding RNAs (trans-effect). High resolution annotations of **CpG islands**, which include cancer relevant CpG island shores (Irizarry et al., 2009), will be combined with **regions predicted as enhancers** by the ENCODE project (high density of TF binding, DHS, histone marks,...) to enrich for functional elements. Correlation with gene expression will be assessed for genes within distances typical for enhancer-promoter pairs (Jin et al., 2013).
3. **unbiased approach**: Fixed sized windows of 1 to 10 kb covering the non-coding genome are applied to identify nSNVs hotspots. Special attention will be given to the identification of non-functional artifacts. This approach will yield a background distribution of correlation values to judge significance of discoveries in step 1 and 2. Furthermore, identification of frequently mutated functional elements outside of predicted regulatory regions is enabled.



In all cases, correlation analysis will be performed using linear models, within a **bayesian framework** designed to distinguish possible covariate confounder effects from observed (cancer subtype, patient information if available) or hidden variables (Stegle, Parts, Durbin, & Winn, 2010). Correlation of hidden effects with known parameters will allow to gain better insight into these effects.

An interesting question is whether altered gene expressed is mainly caused by disruption or *de novo* creation of TFBS and which motifs are most frequently affected. As differential methylation patterns can be induced by altered protein binding-patterns, local DNA methylation levels will be included to judge nSNV function, when available. A special focus will be put on the relation between mutations, differential methylation and gene expression, similar to (Gutierrez-Arcelus et al., 2013) in relation with the particular cancer type.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., Dermitzakis, E. T. (2013). *eLife*, *2*, e00523. doi:10.7554/eLife.00523

Irizarry, R. a, Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Feinberg, A. P. (2009). *Nature genetics*, *41*(2), 178–86. doi:10.1038/ng.298

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Ren, B. (2013). *Nature*, 1–5. doi:10.1038/nature12644

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010). *PLoS computational biology*, *6*(5), e1000770. doi:10.1371/journal.pcbi.1000770

Legacy plans

- Identified mutational hotspots will be made available through a UCSC Genome Browser **Track Hub** as a public resource. A particular effort will be made towards **dynamic vizualization tool**, based e.g. on javascript graphic libraries.
- For all relevant mutations, a database integrating all available information (e.g. affected TFBS, methylation values, expression of nearest gene,...) will be set up, allowing to generate a **comprehensive report for each mutation**. This database could be queried in a gene-centric, region-centric or TF-centric manner.
- All scripts used for the analysis will be made available.

Dr. Benedikt Brors

Group Leader Computational Oncology
Div. Theoretical Bioinformatics
German Cancer Research Center, Heidelberg

E-Mail: b.brors@dkfz.de

Degree

1989–1995	Diploma in chemistry, University of Düsseldorf, Germany
1999	Doctoral degree (Dr. rer. nat) in biochemistry, University of Düsseldorf

Scientific Career

1995–1999	Pre-doctoral research assistant, Inst. of Biochemistry, University of Düsseldorf
1999–2000	Postdoctoral Researcher, Div. Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
2000–2003	Postodoctoral Researcher, Intelligent Bioinformatics Systems, DKFZ
2003–present	Group Leader Computational Oncology, Div. Theoretical Bioinformatics, DKFZ
2003–present	Lecturer in bioinformatics, Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg
2007	Offer of a post as Full Reader in Medical Bioinformatics, Queen's University, Belfast, UK (not realized)
2008	Offer of a post as W2 professor of biostatistics (non-tenured), University Medical Center Göttingen, Germany (not realized)
2009–present	Tenured research position, DKFZ
2013	Offer of a position as full professor of applied bioinformatics, DKFZ and National Center for Tumor Diseases, Heidelberg (under negotiations)

Selected Publications

Jones, D.T.W.,* **Hutter, B.*, Jäger, N.***, (...), **Brors, B.**, (...), Eils, R., Lichter, P., Pfister, S.M., 2013. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet.** 45, 927-932

Weischenfeldt, J.*, Simon, R.*, **Feuerbach, L.***, (...), Sültmann, H.#, Sauter, G.#, Plass, C.#, **Brors, B.#**, Yaspo, M.-L.#, Korbel, J.O.#, Schlomm, T.#, 2013. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. **Cancer Cell** 23, 159–170.

Richter, J., **Schlesner, M.***, (...), Hummel, M#., Klapper, W.#, Rosenstiel, P.#, Rosenwald, A.#, **Brors, B.#**, Siebert, R.#, 2012. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. **Nat. Genet.** 44, 1316–1320.

Jones, D.T.W.*, **Jäger, N.***, (...), **Brors, B.**, (...), Eils, R., Pfister, S.M., Lichter, P., 2012. Dissecting the genomic complexity underlying medulloblastoma. **Nature** 488, 100–105.

Oberthuer, A., Hero, B., Berthold, F., **Juraeva, D.**, (...), **Brors, B.**, Fischer, M., 2010. Prognostic impact of gene expression-based classification for neuroblastoma. **J. Clin. Oncol.** 28, 3506–3515.

(* contributed equally; # contributed equally as senior authors)

Relevant Projects

- PI (bioinformatics analysis) in ICGC-PedBrain
- PI (bioinformatics analysis) in ICGC – Early Onset Prostate Carcinoma
- PI (bioinformatics analysis) in ICGC – Molecular Mechanisms in Malignant Lymphoma
- Member of ICGC Bioinformatics Work Group and Mutation Consequences and Pathways WG

Dr. Carl Herrmann, PhD

1. General Information

Date/Place of Birth :	20th December, 1971, Nantes (France)
Gender :	Male
Citizenship :	French /German
Family position :	Maried, 4 daughters (10, 8, 5, 5 years)
Office addres:	IPMB/BioQuant – Universität Heidelberg
	Im Neuenheimer Feld 364 , D-69120 Heidelberg
Phone :	+49 6221 423612
Email :	c.herrmann@dkfz.de
Current position :	Associate Professor ("Akademischer Rat"), Universität Heidelberg

2. Scientific career

since 2013	:	Akademischer Rat, Universität Heidelberg & Department of Theoretical Bioinformatics, DKFZ Heidelberg (Germany)
2012 – 2013	:	Visiting scientist, EMBL Heidelberg (Germany)
2003 – 2013	:	Associate Professor ("maître de conférences"), Université de Marseille (France)
2001 – 2003	:	Postdoc at the Department of Theoretical Physics, University Turin (Italy)
1999 – 2001	:	Postdoc at the Department of Theoretical Physics, Universität Halle-Wittenberg (Germany)

3. Education

1996 – 1999	:	PhD thesis at the Centre de Physique Théorique, Marseille (France)
1994 – 1995	:	Postgraduate degreee in theoretical physics ("Diplome d'études approfondies") at the Ecole Normale Supérieure, Paris (France)
1991 – 1994	:	Study of engineering and applied mathematics at the Ecole Nationale des Ponts et Chaussées, Paris (France)

4. Scientific responsabilities

- scientific coordinator of the Center for Bioinformatics Teaching and Ressources (CRFB, Université Marseille (2007 – 2010)
- member of the scientific advisory committee of the Vital-IT biocomputing facility, Swiss Institute for Bioinformatics (2005 – 2009)

5. Publications

1. Darbo E, **Herrmann C**, Lecuit T, Thieffry D, van Helden J (2013) *Transcriptional and epigenetic signatures of zygotic genome activation during early Drosophila embryogenesis.* BMC Genomics 14: 226.

2. **Herrmann C**, Van de Sande B, Potier D, Aerts S (2012) *i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules*. Nucleic Acids Res 40: e114.

3. Potier D, Atak ZK, Sanchez MN, **Herrmann C**, Aerts S (2012) *Using cisTargetX to predict transcriptional targets and networks in Drosophila.* Methods Mol Biol 786: 291–314.

4. Thomas-Chollier M, **Herrmann C**, Defrance M, Sand O, Thieffry D, et al. (2012) *RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.* Nucleic Acids Res 40: e31.

5. Punnamoottil B, **Herrmann C**, Pascual-Anaya J, D'Aniello S, Garcia-Fernàndez J, et al. (2010) *Cis-regulatory characterization of sequence conservation surrounding the Hox4 genes.* Dev Biol 340: 269–282.

Lars Feuerbach – Curriculum Vitae

Phone: +49 6221 42 3603 E-mail: l.feuerbach@dkfz.de

RESEARCH POSITIONS

10/2011-	Research Fellow – ICGC Early onset Prostate Cancer Computational Oncology, Theoretical Bioinformatics German Cancer Research Center, Heidelberg, Germany
10/2007-09/2011	Graduate Research Assistant Computational Biology and Applied Algorithms Department Max-Planck Institut für Informatik, Saarbrücken, Germany

EDUCATION

10/2007 –	PhD in Bioinformatics (Thesis submitted – 08/2013) Max-Planck Institut für Informatik, Saarbrücken, Germany
10/2005 – 09/2007	Master of Science (MSc) with Honor's Degree - Bioinformatics Center for Bioinformatics, University of Saarland, Germany
09/2002 – 07/2005	Bachelor of Science (BSc) - Bioinformatics Free University of Berlin, Germany

SELECTED PUBLICATIONS

Joachim Weischenfeldt*, Ronald Simon*, Lars Feuerbach*, Karin Schlangen*, et al.
Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer
Cancer Cell, 2013, 23(2):169-170

Lars Feuerbach, Konstantin Halachev, Yassen Assenov, Fabian Müller, Christoph Bock, Thomas Lengauer
Analyzing epigenome data in context of genome evolution and human diseases
Methods Mol. Biol. 2012,856:431-67

Malay Bhattacharyya, Lars Feuerbach, Tapas Bhadra, Thomas Lengauer, Sanghamitra Bandyopadhyay
MicroRNA Transcription Start Site Prediction with Multi-objective Feature Selection
Statistical Applications in Genetics and Molecular Biology, 2012, 11(1) 1–25

Lars Feuerbach, Rune B. Lyngsoe, Thomas Lengauer, Jotun Hein
Reconstructing the ancestral germline methylation state of young repeats
Molecular biology and evolution 2011;28(6):1777-84

Pavlo Lutsik, Lars Feuerbach, Julia Arand, Thomas Lengauer, Jörn Walter, et al.
BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing.
Nucleic Acids Research, May 11, 2011, 39(Web Server issue):W551-6

Curriculum Vitae, Ewan Birney

Full Name: John Frederick William Birney
Date of Birth: 12 December 1972
Nationality: UK
Email: birney@ebi.ac.uk

77 Lancaster Road
London N4 4PL

Employment:

2012-Current : Associate Director, European Bioinformatics Institute
2000-2012: Head of Nucleotide data, European Bioinformatics Institute
Current supervisor for 4 PhD students

On a variety of SAB boards (includes Riken Institute, BCGSC, Leipzig MPI, Roslin Institute, IMP, TGAC)

1996-2000: PhD at the Sanger Centre (Supervisor, Richard Durbin)

Other positions held:

- A number of consultancy contracts, both strategic and technical in the biotech and pharmaceutical industry, including funding and finance orientated roles.
- Equity Research in SBC Warburg Pharmaceutical division (summer 1995).
- Freelance journalist (Economist) (1995).
- Research Assistant at Cold Spring Harbor Laboratory and EMBL Heidelberg.

Prizes and Awards

EMBO Member, Elected 2012
Winner of the Overton Award from the International Computational Biology Society, 2005
Winner of the Benjamin Franklin Award from Bioinformatics.org/BioIT in 2005
Winner of the Royal Society's Francis Crick Lecture in 2003

Patents:

US Provisional Patent Application 61/654295, *High-capacity storage of digital information in DNA*, filed 1 June 2012 (co-applicant with Nick Goldman)
Patent Cooperation Treaty Application PCT/EP2013/061300, *High-capacity storage of digital information in DNA*, filed 31 May 2013 (co-applicant with Nick Goldman)

Education:

1996-1999: PhD, St John's College Cambridge. Awarded a Scholarship
1992-1996: BA Biochemistry, Balliol College Oxford. 1st Class degree. Awarded a Scholarship

Publications

181 Peer reviewed publications, 23 in Nature (5 first/last author), 9 Science (1 last author). 1 Cell (joint last author). H-index: 83. Avg Citations/Paper 331. (Google Scholar)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Patterns of disregulation of splicing and alternative exon usage in cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Jan Korbel, EMBL Heidelberg; ICGC PedBrainTumor

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Simon Anders and Alejandro Reyes, EMBL Heidelberg (Genome Biology Unit)	Wolfgang Huber, EMBL Heidelberg (Genome Biology Unit)

Background and preliminary data

Cancer genome sequencing projects have discovered many recurrent mutations in genes involved in RNA processing, highlighting a previously underappreciated role of RNA-processing in tumourigenesis. In particular, recurrent mutations of splicing factors have been found, among them, recurrent mutations of the HEAT domain in the splicing factor SF3B1 in CLL [1, 2]. On the other hand, cancer transcriptome sequencing (RNA-Seq) studies have revealed a high prevalence of cancer-specific isoforms and, for some cases, specific transcript choices have been shown to promote the disease progression (reviewed in [3]). Currently ongoing follow-up work (among others, by our group) focuses on the impact of mutations is transcription-related genes like SF3B1 on the transcriptome and their downstream biological consequences.

Our group has many years of expertise on statistical methods for the analysis of RNA-Seq data (we developed, e.g., the widely used *DESeq* tool [4]), including testing for differential usage of exons between sample groups (the *DEXSeq* method [5]). Recently, we used our methodology to systematically assess the impact of the SF3B1 mutation on exon usage in CLL [6], and we have already established the core of the workflow necessary for the proposal.

[1] Quesada et al (2012): *Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia.* Nat Genet 44: 47-52.
[2] Wang et al. (2011): *SF3B1 and other novel cancer genes in chronic lymphocytic leukemia.* N Engl J Med 365:2497-2506.
[3] Zhang & Manley (2013): *Misregulation of pre-mRNA alternative splicing in cancer.* Cancer Discov 3: 1228.
[4] Anders & Huber (2010): *Differential expression analysis for sequence count data*. Genome Biol 11: R106.
[5] Anders, Reyes & Huber (2012): *Detecting differential usage of exons from RNA-seq data*. Genome Res 22: 2008-2017.
[6] Reyes, Blume, Pelechano, Jakob, Steinmetz, Zenz & Huber (2013): *Transcriptome analysis of chronic lymphoid leukemia reveals isoform regulation associated with mutations in SF3B1*. Submitted.

Timelines & resources dedicated to project

Our analysis will start immediately after gaining access to the preprocessed alignments. A graduate student (Alejandro Reyes) will devote 50% to the project, together with senior PhD-level researcher Simon Anders. They will be supported by frequent interactions with other group members with expertise in statistics, scientific computing, and cancer genomics. Recruitment of new group members who may join the project is planned during 2014.

While Jan Korbel is listed as PI, his role is scalable to his available time. The project is viable with scientific and managerial leadership by Wolfgang Huber (effort: 20%).

Research proposal

We aim to conduct a **systematic search for mutations in splicing factors or other genes related to isoform regulation that are statistically associated with differential exon usage** on the panel of combined genome and transcriptome sequence data available through PanCancer.

The statistical methodology and computational tools for this are in place, and here we propose to apply them to a large data set, and to conduct the integrative and follow-up analyses needed to substantiate the found candidate "hits" and patterns of hits.

Specifically, we will first screen the variant calls of all PanCancer samples with RNA-Seq data for mutations in genes that are annotated to have potential impact on isoform structure, including factors that can influence the choice between alternative splice sites, between alternative promotors, or polyadenylation sites. We will identify the cancer types for which mutations in these candidate genes are recurrent.

For these sample sets, we will compare the transcriptomes of the samples with and without the mutation using our *DEXSeq* method and so identify genes whose isoform regulation is affected by the mutation. Wherever such recurrently mutated candidate genes are found in several cancer types (e.g., the SF3B1 mutation has now also been found in other malignancies [7-10]), the usage of generalized linear models (GLMs) inherent to the DEXSeq methodology will allow for a combined analysis, which should improve inferential power and specificity. (We note that other authors unrelated to us, including Ref. [10], have employed *DEXSeq*, reflecting the its relevance for such tasks.)

The primary result of this analysis will be a list of associations between recurrently mutated factors and their affected downstream targets (specific exons coding for protein regions or UTRs), which we will compile across all analyzed data and present in suitable formats: (i) machine-readable for subsequent systematic analysis and (ii) in a human-friendly, visual, easily browsable, easy-to-drill-down form. We hope that this will also become a useful resource for other researchers with expertise in specific cancer types or mechanisms. In parallel, we aim to exploit found individual results or 'patterns' to contribute to the understanding of the role of transcript isoform de-regulation in cancer.

[7] Yoshida et al (2011): *Frequent pathway mutations of splicing machinery in myelodysplasia*. Nature, 478:64-69.
[8] Papaemmanuil et al (2011): *Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts*. N Engl J Med, 365:1384-1395.
[9] Harbour et al (2013): *Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma*. Nat Genet, 45:133-135.
[10] Furney et al (2013): *SF3B1 mutations are associated with alternative splicing in uveal melanoma*. Cancer Discov, advance online publication.

Legacy plans

The result of our analysis will take the shape of simple tables, which we will make publically available both as a browsable web site ("human-friendly"), and in text files or other suitable standardized data formats ("machine-readable").

It is standard practice in our research to make analysis workflows for our publications available as literate programming documents, i.e., code interspersed with detailed explanations of the computational steps. Such documents can be executed and will reproduce the complete analysis if provided with the original input data, which will therefore also be provided. We aim to use PanCancer cloud resources and/or services hosted by EMBL for the latter. Specifically, the literate programming approach should integrate well with the planned use of Sage Synapse to provide version linkage between data.

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci* USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

Simon Anders

Physicist (Dr. rer. nat.), working in bioinformatics/biostatistics

Contact

Genome Biology Unit, European Molecular Biology Laboratory, 69111 Heidelberg, Germany
Phone +49-6221-387-8632 (office), +49-176-96862925 (mobile), e-mail: sanders@fs.tum.de

Postdoc

2009—present: European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
2007—2009: European Bioinformatics Institute (EMBL-EBI), Hinxton/Cambridge, UK
Postdoctoral fellow; since 2012 Staff Scientist
Research groups of Dr. Wolfgang Huber and Prof. Lars Steinmetz

PhD Studies

2004—2007: Universität Innsbruck, Austria
2003: Ludwigs-Maximilans-Universität, Munich, Germany
Research group of Prof. Hans J Briegel
PhD in theoretical physics (Quantum information theory)

University Education

1997—2002: Technische Universität München, Munich, Germany
Study of Physics (Diplom degree, equivalent to Master of Science)
2002: thesis research work in experimental quantum optics (Group of Prof. Dirk Bouwmeester; University of Oxford, UK, and University of California, Santa Barbara, USA)

Selected recent publications

P Brennecke*, S Anders*, J K Kim*, A A Kołodziejczyk, X Zhang, V Proserpio, B Baying, V Benes, S A Teichmann, J C Marioni, M G Heisler: *Accounting for technical noise in single-cell RNA-seq experiments*." Nature Methods 10 (2013) 1093—1095.

A Reyes*, S Anders*, R J Weatheritt, T Gibson, L M Steinmetz, W Huber: *Drift and conservation of differential exon usage across tissues in primate species*. PNAS 110 (2012) 15377—15382.

S Anders*, A Reyes*, Wolfgang Huber: *Detecting differential usage of exons from RNA-seq data.* Genome Research 22 (2012) 2008—2017.

A Schlattl, S Anders, SM Waszak, W Huber, JO Korbel: *Relating CNVs to transcriptome data at fine resolution.* Genome Research 21 (2011) 2004-2013.

S Anders, W Huber: *Differential expression analysis for sequence count data*. Genome Biology 11 (2010) R106. (486 citations [PubMed].)

Expertise and skills

Current research interests: Statistical methods for the analysis of high-throughput sequencing data, especially for transcriptomics; alternative isoform regulation; somatic variant calling; cancer drug screens

Core skills: Computational sciences, statistics, numerics, scientific software development in various programming languages (C/C++, R, Python, etc.)

Languages: German (mother tongue), English (fluent), French (intermediate)

Alejandro Reyes

Curriculum Vitae (November 2013)

PERSONAL DETAILS

Birth	February 12, 1989
Address	Hausserstr 63 69126 Heidelberg
Phone	+49 6221 387-8169
Mail	alejandro.reyes@embl.de

EDUCATION

B.Sc. Genomics aug 2007 - june 2011
National Autonomous University of Mexico, Mexico
Graduated with Honours. 96/100

RESEARCH POSITIONS

Undergraduate research assistant nov 2009 - june 2010
National Autonomous University of Mexico, Mexico
Supervisors: Prof. Julio Collado-Vides and Prof. Enrique Morett

Kupcinet-Getz Science Summer School Student june 2010 - aug 2010
Weizmann Institute of Science, Israel
Supervisor: Prof. Doron Lancet

Traineeship aug 2010 - june 2011
European Molecular Biology Laboratory, Germany
Supervisor: Dr. Wolfgang Huber

Predoctoral Fellow sep 2011 - current
European Molecular Biology Laboratory, Germany
Supervisor: Dr. Wolfgang Huber

SKILLS

Software	R, PERL, C, PYTHON, MYSQL
Languages	Spanish (native), English (fluent), Italian (basic)

PUBLICATIONS AND SOFTWARE

[1] Reyes* A, Anders* S, and Huber W. Analyzing RNA-seq data for differential exon usage with the DEXSeq package. *Software: R/Bioconductor* 2012.

[2] Anders* S, Reyes* A, and Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Research* 2012.

[3] Olender T, Waszak SM, Viavant M, Khen M, Ben-Asher E, Reyes A, Nativ N, Wysocki CJ, Ge D, and Lancet D. Personal receptor repertoires: olfaction as a model. *BMC Genomics* 2012.

[4] Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, and Ule J. Direct competition between *hnRNP C* and *U2AF65* protects the transcriptome from the exonization of alu elements. *Cell* 2013.

[5] Reyes* A, Anders* S, Weatheritt RJ, Gibson TJ, Steinmetz LM, and Huber W. Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences* 2013.

[6] Reyes A, Blume C, Pelechano V, Jakob P, Steinmetz LM, Zenz T, and Huber W. Transcriptome analysis of chronic lymphoid leukemia reveals isoform regulation associated with mutations in *SF3B1*. *submitted*.

Wolfgang Huber

Contact
Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
Phone +49-6221-387-8823, e-mail: whuber@embl.de

Positions
2004—present: Research Group Leader at EMBL
2011—present: EMBL Senior Scientist
2009—present: Research Group Leader at Genome Biology Unit, EMBL Heidelberg
2004—2009: Research Group Leader at European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
2000—2004: Postdoc at German Cancer Research Centre (DKFZ), Heidelberg, Germany
1998—1999: Postdoc at IBM Research Almaden, San Jose, California
1994—1998: Teaching and Research Assistant, PhD Student at Faculty of Physics, University of Freiburg, Germany

Education
1998: Dr. rer. nat. (≙ PhD) in Theoretical Physics, University of Freiburg, Germany
1994: Diplom (≙ MSc) in Physics, University of Freiburg, Germany

Research overview and publications
http://www.embl.de/research/units/genome_biology/huber

Research highlights of the last five years
Transcriptomics and genetics: Developed computational analyses that led to discoveries in the fields of transcription (non-coding transcripts, bidirectional promoters) and meiotic recombination.

Genetic interactions and phenotyping: Established computational methods for reverse genetics by high-throughput RNAi and automated phenotyping by image analysis. I have applied these to the large-scale quantitative mapping of genetic interaction matrices by combinatorial RNAi.

Statistical methodology and computing: Developed new statistical methods and made contributions to Bioconductor, a large bioinformatics software project that has found widespread use in genome biology.

External academic services
Grant and project reviews: BBSRC, MRC, Wellcome Trust, Cancer Research UK, NSF, Ontario Research Fund, DFG, HFSP, Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), WWTF, Academy of Finland, others

Boards: Advisory Board, Bioconductor Project, NIH P41 (USA) • Executive Board, Systems Microscopy: EC FP7 Network of Excellence • Scientific Coordinator, Radiant: EC FP Collaborative Project • Scientific Advisory Board, Sophia Genetics S.A. (Switzerland) • Precision Oncology Programme Committee, DKFZ/NCT

Consulting: Genentech, Inc. (USA) • Evotec, Germany

Journal editor: Statistical Applications in Genetics and Molecular Biology (associate editor 2003–12) • Giga Science Journal (editorial board since 2011) • Bioinformatics (editorial board) • R-News (guest editor, special issue on bioinformatics, 2006)

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract	
Estimation of position-specific error profiles from aggregated genome sequencing cohorts	
Name, institute & ICGC/TCGA affiliations of principal investigator	
Jan Korbel (co-PI), Genome Biology Unit, EMBL Heidelberg, ICGC PedBrainTumor, ICGC Early-Onset Prostate Cancer, ICGC Malignant Lymphoma	
Name & institute of junior investigator	Name & institute of non-ICGC collaborator
Julian Gehring, Genome Biology Unit, EMBL Heidelberg	Wolfgang Huber (lead-PI), Genome Biology Unit, EMBL Heidelberg

Background and preliminary data

While the calling of somatic cancer variants is the workhorse of cancer genome sequencing, it remains challenging with regard to several aspects, including (i) due to subclones whose coverage is near the coverage-dependent detection limit, (ii) complex structural variants, (iii) and systematic biases of the technology. A diversity of approaches for controlling false positives exist. However, many of these approaches rely significantly on post-processing steps including manually tuned heuristics and manual inspection, in order to reduce the number of erroneous calls (*Alexandrov et al., 2013, Nature*). Over-optimization that reduces false positive rates can lead to suboptimal false negative rates. The need to find solutions to these problems is becoming more pressing as integrative studies of different cancer types across thousands of samples are pursued.

Errors are introduced at multiple levels, driven by processes dependent on sequence, library preparation, sequencing technology, and data analysis. A particular challenge is posed by the variety of sources of artifacts and so-called “batch effects” (*Leek et al., 2010, Nature Reviews Genetics*; *Taub et al., 2010, Genome Medicine*). The multitude of error sources prohibits mechanistic modeling, and indicates that statistical modeling based on empirical data is required to extend existing approaches (*Muralidharan et al., 2012, Annals of Applied Statistics*).

With these challenges in mind, we are currently analyzing the 1000 Genomes Project data set (*The 1000 Genomes Project Consortium, 2010, Nature*), and have developed a statistical and computational framework for applying this to large cohorts. The framework offers a HDF5-based infrastructure (*The HDF Group, 2000, http://www.hdfgroup.org/HDF5*) and is implemented in our 'h5vc' software package available from the Bioconductor repository (Pyl et al., in review). Our analyzes indicate that the simultaneous model fitting to data from thousands of samples is technically feasible. Regarding somatic variant calling, we have already observed in smaller cohorts that position-specific error rates aid in the assessment of called variants in multiple cancer types.

The WGS Pan-Cancer project provides an ideal setting for addressing the question of estimating biases from cohort studies, as it will give us access to a large number of samples, modern sequencing technology, and a standardized data processing pipeline. Furthermore, we believe that our statistical approaches will aid other projects within the Pan Cancer initiative in the process of identifying genomic alterations, and in overcoming the many technical challenges mentioned above.

Timelines & resources dedicated to project

Our analysis will start after gaining access to the whole genome alignments provided by the Pan-Cancer initiative. The graduate student Julian Gehring will devote 50% of his time to the project and will be supported by frequent interactions with other group members with expertise in statistics, scientific computing, and cancer genomics. While Dr. Jan Korbel is listed as PI, his role is scalable to his available time. The project is viable with scientific and managerial leadership by Dr. Wolfgang Huber.

Since this project has a strong methodological component, we are interested in exchanging our methods and results with other working groups at early stages.

Research proposal

We aim to develop and apply a stratified error model to the data that predicts the probability of common sequencing errors according to genomic position, base, and strand; further stratification criteria will be considered as the analyzes progress. We anticipate that error rates in large, aggregated datasets will vary with study- or batch-specific factors, and therefore intend to include explicit or implicit effects based on methodology that has already been successfully employed in analogous settings (*Leek and Storey, 2007, PLoS Genetics*; *Stegle et al., PLoS Computational Biology, 2010*). To our knowledge, this will constitute one of the first attempts at characterizing local and global biases on a genome-wide scale. We hope that the results will be instrumental as input for existing variant calling approaches (e.g. *Gerstung et al., 2011, Nature Communications*), and that they will provide a significant contribution to our ability to investigate rare somatic variants present at low frequencies in heterogenous tumor samples (*Yates and Campbell, 2012, Nature Reviews Genetics*).

After completion, this project will yield results at multiple levels:

Firstly, the estimated error profiles will be made available to other research groups within and outside of the Pan Cancer initiative. We hope that these will provide a valuable resource for complementing existing annotations and post-processing workflows. Additionally, an interactive web interface will be offered for flexible access to the results.

Secondly, we envisage that the application of the statistical methodology developed in this project will benefit other sequencing studies in the future.

Thirdly, a computational framework for efficient computation on, and storage of the error profiles, will be created based on the existing HDF5 infrastructure described before.

Legacy plans

A primary objective of the proposed project is to provide the framework of the statistical analysis and the data it generates to other scientific groups. The statistical framework will be distributed as a set of documented open-source software packages and maintained in close collaboration with the Bioconductor community (*Gentleman et al., 2004, Genome Biology*), which will allow researchers access to a scalable and user-oriented implementation. In addition, the estimated error profiles will be made publicly available as downloadable data tracks for integration with existing analysis pipelines, and will be additionally accessible through a web service for exploratory analysis of the data.

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci* USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

Curriculum Vitae
Julian Gehring

Personal Details

Birth	October 29, 1985, in Lahr, Germany
Address	Meyerhofstraße 1, 69117 Heidelberg
Phone	++49 6221 387-8224
E-Mail	julian.gehring@embl.de

Academic Degrees

Diploma in biology, awarded by the Alberts-Ludwigs University Freiburg, Germany (10/01/2011)
Grade: magna cum laude (1.1)

Research Positions and Education

10/2005 – 10/2010:	Studies of biology at the Alberts-Ludwigs University of Freiburg, Germany
11/2010-05/2011:	Research assistant at the Institute of Physics, University of Freiburg
07/2011-10/2011:	Visiting scientist at the group of Wolfgang Huber, EMBL Heidelberg, Germany
since 10/2011 :	Predoctoral fellow at the group of Wolfgang Huber, EMBL Heidelberg, Germany

Publications

Jenny Hansson, Mahmoud Reza Rafiee, Sonja Reiland, Jose M. Polo, Julian Gehring, Satoshi Okawa, Wolfgang Huber, Konrad Hochedlinger, and Jeroen Krijgsveld.
"Highly Coordinated Proteome Dynamics During Reprogramming of Somatic Cells to Pluripotency."
Cell Reports 2, no. 6 (December 27, 2012): 1579–1592.

Clemens Kreutz, Julian Gehring, Daniel Lang, Ralf Reski, Jens Timmer, and Stefan Rensing.
"TSSi - an R Package for Transcription Start Site Identification from 5' mRNA Tag Data."
Bioinformatics (Oxford, England) 28, no. 12 (June 15, 2012): 1641–1642.

Recent Bioconductor Software Packages

TSSi: Transcription Start Site Identification from Tag-Sequencing Data (2011)

proteinProfiles: Clustering of protein time course data (2011)

SomaticCancerAlterations: Collection of TCGA somatic mutation calls (2013)

COSMIC.67: Collection of COSMIC somatic mutational calls and commonly mutated genes (2013)

Wolfgang Huber

Contact
Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
Phone +49-6221-387-8823, e-mail: whuber@embl.de

Positions
2004—present: Research Group Leader at EMBL
2011—present: EMBL Senior Scientist
2009—present: Research Group Leader at Genome Biology Unit, EMBL Heidelberg
2004—2009: Research Group Leader at European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
2000—2004: Postdoc at German Cancer Research Centre (DKFZ), Heidelberg, Germany
1998—1999: Postdoc at IBM Research Almaden, San Jose, California
1994—1998: Teaching and Research Assistant, PhD Student at Faculty of Physics, University of Freiburg, Germany

Education
1998: Dr. rer. nat. (≙ PhD) in Theoretical Physics, University of Freiburg, Germany
1994: Diplom (≙ MSc) in Physics, University of Freiburg, Germany

Research overview and publications
http://www.embl.de/research/units/genome_biology/huber

Research highlights of the last five years
Transcriptomics and genetics: Developed computational analyses that led to discoveries in the fields of transcription (non-coding transcripts, bidirectional promoters) and meiotic recombination.

Genetic interactions and phenotyping: Established computational methods for reverse genetics by high-throughput RNAi and automated phenotyping by image analysis. I have applied these to the large-scale quantitative mapping of genetic interaction matrices by combinatorial RNAi.

Statistical methodology and computing: Developed new statistical methods and made contributions to Bioconductor, a large bioinformatics software project that has found widespread use in genome biology.

External academic services
Grant and project reviews: BBSRC, MRC, Wellcome Trust, Cancer Research UK, NSF, Ontario Research Fund, DFG, HFSP, Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), WWTF, Academy of Finland, others

Boards: Advisory Board, Bioconductor Project, NIH P41 (USA) • Executive Board, Systems Microscopy: EC FP7 Network of Excellence • Scientific Coordinator, Radiant: EC FP Collaborative Project • Scientific Advisory Board, Sophia Genetics S.A. (Switzerland) • Precision Oncology Programme Committee, DKFZ/NCT

Consulting: Genentech, Inc. (USA) • Evotec, Germany

Journal editor: Statistical Applications in Genetics and Molecular Biology (associate editor 2003–12) • Giga Science Journal (editorial board since 2011) • Bioinformatics (editorial board) • R-News (guest editor, special issue on bioinformatics, 2006)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract
TRIADE - Tumor-Related Infectious Agent Detection

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)
Peter Lichter, Molecular Genetics, German Cancer Research Center (dkfz), ICGC Roland Eils, Theoretical Bioinformatics, German Cancer Research Center (dkfz), ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Marc Zapatka, Molecular Genetics, dkfz, ICGC Matthias Schlesner, Theoretical Bioinformatics, dkfz	Adam Grundhoff, Heinrich-Pette-Institute for Experimental Virology

Background and preliminary data

It is widely accepted that the etiology of approximately 18% of human cancers is causally linked to infection with 7 viral and 1 bacterial species presently recognized as class 1 or 2a carcinogens. Based primarily on highly sensitive PCR detection methods, it has also been speculated that the fraction of human tumors associated with the above or other known pathogens may be substantially larger. Interestingly, however, a recent study of RNA-seq data from the TCGA panel found little evidence for the immediate involvement of known viruses in other cancers (Tang et al., 2013. Nat Commun. 4:2513). These findings suggest that, if additional infectious tumor agents exist, they may rather represent bacteria or, more likely, hitherto undiscovered viral species. While such agents may principally involve virus families not previously linked to tumorigenesis, we deem it more likely they represent novel human members of viral taxa already known to cause cancer. For example, the human gamma-herpesviruses EBV and KSHV have close relatives in the lymphocryptovirus and RV1 rhadinovirus lineages, respectively, of many primate species. In contrast, a human representative of the RV2 lineage of tumorigenic rhadinoviruses has thus far not been identified, even though evolutionary evidence strongly argues for its existence (Bruce et al., 2012. PLoS Pathog. e1002962). Likewise, the recent discovery of many novel human polyomaviruses, including the tumorigenic Merkel Cell Polyomavirus (MCPyV), suggests that additional cancer-associated polyomaviruses may exist in humans.

The WGS pan-cancer dataset offers a unique possibility to discover such agents. As a member of the German Centre for Infection Research (DZIF), Grundhoff et al. have recently developed a pipeline for the rapid and unbiased detection of known as well novel pathogens. In addition to the detection of primary nucleotide sequence homology, the pipeline was particularly designed to allow identification of novel viruses by pattern-based approaches. We already applied the pipeline samples from Sung et al. (Nat Gen. 44(7),765-9) detecting hepatitis B virus. We propose to use and further extend these methods to comprehensively interrogate the WGS pan-cancer dataset for the presence of infectious agents.

Timelines & resources dedicated to project

The proposed project depends on the availability of DNA-seq and RNA-seq datasets. If existent, availability of small-RNA-seq datasets would be an advantage. Clinical metadata, especially regarding patient's immune status, would be advantageous for follow up investigation.

The primary detection pipeline is integrated in a single software tool and employs a PostgreSQL database to store comprehensive results along with applied parameters and other metadata. The software is fully operational at the Heinrich-Pette-Institute, and we expect no difficulties in adapting it to cloud-based VM environment. We expect implementation of additional structural pattern-recognition algorithms (see proposal) to be complete by January 2014. However, since the pipeline is modular, primary analysis can commence as soon as DNA-/RNA-seq datasets become available.

Research proposal
We propose to perform a comprehensive classification of known or novel pathogens out of whole genome DNA- and RNA-seq data, using an already implemented detection pipeline termed DAMIAN (Detection and Analysis of Microbial Infectious Agents by NGS). The pipeline was specifically designed to analyze large numbers of complex datasets from clinical samples, with the goal of identifying signatures of pathogens that may represent etiologic disease agents according to the following steps that have been integrated in a single software tool: 1. Quality filtering & digital subtraction: The pipeline performs quality filtering and subtraction of reads originating from the host by Bowtie2 alignment to the human reference genome or transcriptome. Already hg19-aligned DNA-seq datasets can be directly fed into the pipeline. 2. *De novo* assembly and mapping. Non-host-aligned reads are assembled using SPAdes, and all reads are subsequently mapped back to the assembled contigs to optimize estimation of contig abundance. Since pathogens present at low frequency may only produce spurious reads, reads not assigned to contigs are additionally mapped to all RefSeq bacterial and viral genomes. 3. Taxonomic classification / identification of putative novel pathogens. De novo assembled contigs are complexity filtered and processed by successive classifier modules. First, contigs with extensive nucleotide homology are identified by alignment to the NCBI nr database using megablast, followed by blastn alignments for more distant relatives of known pathogens. Remaining contigs are then aligned to the nr database by blastx and additionally annotated for the presence of conserved aa motifs/domains by HMMER and Pfam database searches. Besides of primary sequence homology, the pipeline also annotates structural and architectural features characteristic for infectious agents, e.g. circularity or presence of overlapping open reading frames. Integration events can likewise be recognized by identification of reads that partially overlap with host sequences. 4. Cross-sample comparison. To identify etiologic agents present only or preferentially in disease samples, contigs from multiple sample sets are clustered by taxonomy annotation and relative abundance. Additionally, contigs can be subjected to cross-sample clustering according to sequence similarity (blastclust), allowing the identification of high priority candidates even among contigs that may have evaded any taxonomic classification. Significance sorted results are output in tabular form and can be visualized and further analyzed using an independently executed GUI-application. By implementing pattern-based classification methods and cross-sample clustering, our approach adds significantly to existing methods that employ digital subtraction, de novo assembly and blast homology searches (e.g., PathSeq, RINS, RITA). The existing pipeline can be readily extended to include additional classifiers, e.g. to identify non-coding signatures such as miRNAs that would be missed by conventional methods. We and others have shown that high-level expression of viral miRNAs is a frequent feature especially of transforming polyoma- and herpesviruses (reviewed in Grundhoff & Sullivan, 2011. Virology 411(2)). We have previously developed a computational method (VMir) for the *de novo* prediction of viral pre-miRNA structures from DNA and RNA sequences (Sullivan et al., 2005. Nature 435; Walz et al., 2010; J. Virol. 84(2)) and are currently implementing this and other tools (miRDeep2) to allow annotation of miRNA-encoding contigs. If necessary, other classifiers may be adjusted to favor discovery of potential cancer-causing agents, for example by lending more weight to the presence of motifs frequently present in transforming viruses (e.g, LxCxE or PDZ domain binding motifs). The DAMIAN pipeline has been recently applied to diverse and complex DNA-seq and RNA-seq sample sets (e.g., from gastroenteritis outbreaks) to verify its ability to identify etiologic agents and discover novel viruses (Grundhoff, Fischer, Alawi et al., in preparation). We thus consider it ideally suited to investigate the pan-cancer dataset.
Legacy plans
The developed software tools will be embedded in a virtual machine enabling their use in the context of the ICGC Pan-Cancer cloud computing infrastructure. We will also make the tools and comprehensive documentation publicly available upon publication of the results.

Peter Lichter

*(*1957)*

E-Mail: peter.lichter@dkfz-heidelberg.de

ResearcherID: I-3483-2013

Current Position

Head of Division Molecular Genetics (B060) since 1992
Full Professor at the Faculty of Medicine, University of Heidelberg since 2000

Research Topics

- Pathomechanisms of tumor development
- Tumor markers
- Molecular profiling of tumor cells
- Genome organization and gene function

Degree

Graduation (Biology) and PhD degree University of Heidelberg 1983-1986
Postdoc in the laboratory of Prof. D.C. Ward, Dept. of Genetics, School of Medicine, Yale University, New Haven, USA 1986-1990
Habilitation and venia legendi in Molecular Human Genetics, Faculty of Medicine, University of Heidelberg 1995

Previous appointments

Head of project group "Organization of complex genomes", DKFZ 1990-1992
Interim Director of the Management Board of the DKFZ 2003

Selected Publications

- Döhner H,..., **Lichter P** Genomic aberrations predict survival of patients with B-cell chronic lymphocytic leukemia. **New Engl. J. Med**. 343, 1910-1916 (2000)
- Rausch T, Jones DTW, Zapatka M, Stütz AM,..., **Lichter P***, Pfister SM*, Korbel JO* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. **Cell** 148, 59-71 (2012) * shared senior authorship
- Jones DTW, Jäger N,..., **Lichter P.** Dissecting the genomic complexity underlying medulloblastoma. **Nature** (2012) (DOI: 10.1038/nature11284)
- Jones DT, Hutter B, Jäger N,..., Eils R, **Lichter P**, Pfister SM; International Cancer Genome Consortium PedBrain Tumor Project. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet**.;45(8):927-32 (2013)
- Remke M, Hielscher T, ..., Taylor MD, **Lichter P**, Pfister SM. FSTL5 is a marker of poor prognosis in non-WNT/non-SHH medulloblastoma. **J Clin Oncol**.;29(29):3852-61 (2011)

Activities in the scientific community

- Member of Hinterzartener Kreis (DFG) 2004-2010
- Sci. Progr. Committee of the Eur. Soc. of Human Genetics 2004-2008
- Member of Wissenschaftsrat 2005-2011
- Member of Leopoldina since 2006
- Member of EMBO since 2008

Honors and awards

- "Karl-Freudenberg" Award of the Academy of Sciences, Heidelberg (1991)
- Award of the (German) Society for Human Genetics (1992)
- "Walther und Christine Richtzenhain" Award (1993)
- Deutscher Krebspreis (German Cancer Award, 2002)
- Award of Deutsche Krebshilfe (2003)
- Award of the European Society of Human Genetics (2012)

Curriculum vitae **Eils, Roland**

Date of birth: 26.05.1965
Place of birth: Krefeld, Germany
Work address: German Cancer Research Center – DKFZ
Division Theoretical Bioinformatics (B080)
Im Neuenheimer Feld 580
D-69120 Heidelberg, Germany
Tel.: +49 6221 42 3600, Fax: +49 6221 42 3620
e-mail: r.eils@dkfz.de

Academic education and qualifications:

1984-1990	Mathematics and Computer Science, RWTH Aachen (M.Sc. (Diplom), first class)
1988-1990	Southeast Asian Languages, Universities of Bonn and Cologne
1990-1992	Southeast Asian Languages, Universitas Padjadjaran, Bandung, Indonesia
1992-1995	Ph.D. study at Heidelberg University, Specialization: *Mathematics and Scientific Computing*
1995	PhD in Mathematics, first class

Professional appointments:

1996-1996	Guest researcher at the *Institut Albert Bonniot*, Université Grenoble, France (Host: Prof. M. Robert-Nicoud, Faculty of medicine).
1997 – 1999	Head of the Biocomputing group „Structure and function in cell biology“, IWR, University of Heidelberg, Germany
2000-2003	Head of the *Biofuture* Junior Group „Intelligent bioinformatics systems“ at the German Cancer Research Center (DKFZ)
Since 2003	Head of Division „Theoretical Bioinformatics“ at the German Cancer Research Center (DKFZ), Heidelberg
Since 2004	Ordinarius of Bioinformatics & Functional Genomics, Heidelberg University
Since 2006	Founding director of BIOQUANT- Centre for Quantitative Biology, University of Heidelberg (~250 scientists in 20 research groups).
2010-2011	Visiting Professor at Harvard Medical School, Harvard University

Awards and honours:

Appointment for Program Director and Full Professor for Quantitative Biology at Cold Spring Harbor Laboratory, New York (2011, declined)
Award for New Innovative Research by the Helmholtz Association “Systems Biology of Complex Diseases” (2005)
Microsoft Research Award “Computational Tools for Advancing Science” (2005)
BioFuture Prize from the German Ministry for Education and Research (1999)

Five most important publications since 2008 (# senior authorships)

Jäger, N, Schlesner, M, Jones, DTW, Raffel, S, Mallm, JP, Junge, KM, . . . Eils, R# Hypermutation of the inactive X chromosome is a frequent event in cancer. **Cell**, 155(3), 2013

Jones, DT, Hutter, B, Jäger, N, …., Eils, R#, Lichter, P#, Pfister, SM#: Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nature Genetics**, 45(8), 2013.

Jones, DT, Jäger, N,…, Eils, R#, Pfister, SM#, Lichter, P#: Dissecting the genomic complexity underlying medulloblastoma. **Nature**, 488(7409), 2012.

Neumann, L, Pforr, C, Beaudouin, J, Pappa, A, Fricker, N, Krammer, PH, Lavrik, IN, and Eils,R#: Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. **Molecular Systems Biology** 6:352, 2010.

Busch, H, Camacho-Trullio, D, Rogon, Z, Breuhahn, K, Angel, P, Eils, R#, and Szabowski, A# Gene network dynamics controlling keratinocyte migration. **Molecular Systems Biology** 4:199, 2008.

Curriculum vitae **Zapatka, Marc**

Date of birth: 12.01.1974
Place of birth: Essen, Germany
Work address: German Cancer Research Center – DKFZ
Division Molecular Genetics (B060)
Im Neuenheimer Feld 580
D-69120 Heidelberg, Germany
Tel.: +49 6221 42 4592, Fax: +49 6221 42 4639
e-mail: m.zapatka@dkfz.de
Researcher ID G-9896-2013

Academic education and qualifications:

1994 - 1999	Study of Biochemistry at the Ruhr-University, Bochum, Germany
1999	Diploma (Masters equiv.) in Biochemistry, Ruhr-University, Bochum, Germany
1999 - 2003	PhD study at the University Hospital, Bochum, Germany
2003	Dr. rer. nat. (PhD equiv.), Ruhr-University, Bochum, Germany

Professional appointments:

2003 – 2004	Postdoctoral fellow at University Hospital, Bochum, Germany
2005 – 2009	Postdoctoral fellow, Division Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
2010	Appointed as scientific head of research group Bioinformatics, Division Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany

Five most important publications *(* first authorship; § before joining DKFZ)*

Jones DT, Hutter B, Jäger N, ..., **Zapatka M**, …, Eils R, Lichter P, Pfister SM: Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet.**, 45(8):927-32, 2013.

Sturm D, Witt H, Hovestadt V, ..., **Zapatka M**, ...**,** Lichter P, Plass C, Jabado,N., Pfister, SM: Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. **Cancer Cell**, 22(4), 425–437, 2012

Rausch T.*, Jones DTW.*, **Zapatka M.*** …, Eils R, Lichter P, Pfister SM, Korbel JO: Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. **Cell** 148 (1-2), 59-71, 2012.

Findeisen P.*, **Zapatka M.***, et al.: Serum Amyloid A As a Prognostic Marker in Melanoma Identified by Proteomic Profiling. **Journal of Clinical Oncology** 27 (13), 2199-2208, 2009.

Zapatka M.*§, Zboralski D.*, et al.: Basement membrane component laminin-5 is a target of the tumor suppressor Smad4. **Oncogene** 26 (10), 1417-1427, 2007.

Curriculum vitae **Schlesner, Matthias**

Date of birth: 08.08.1978
Place of birth: Kiel, Germany
Work address: German Cancer Research Center – DKFZ
Division Theoretical Bioinformatics (B080)
Im Neuenheimer Feld 580
D-69120 Heidelberg, Germany
Tel.: +49 6221 42 3629, Fax: +49 6221 42 3626
e-mail: m.schlesner@dkfz.de

Academic education and qualifications:

1998 - 2003	Study of Human Biology at the University of Marburg, Marburg, Germany
2003	Diploma in Human Biology, University of Marburg
2004 - 2008	PhD study at the Max Planck Institute for Biochemistry, Department for Membrane Biochemistry (Prof. D. Oesterhelt), Martinsried, Germany
2008	Dr. rer. nat. (PhD), LMU Munich

Professional appointments:

2009 - 2011	Postdoctoral fellow and project group leader, Max Planck Institute for Biochemistry, Department for Membrane Biochemistry, Martinsried, Germany
2011 - 2013	Postdoctoral fellow, Division Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
2013	Appointed as head of Computational Oncology Group, Division Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany

Five most important publications since 2008

Jäger N, Schlesner M, Jones DTW, Raffel S, Mallm JP, Junge KM, . . . Eils, R[#] (2013) Hypermutation of the inactive X chromosome is a frequent event in cancer. **Cell** 155(3): 567-81

Alexandrov LB, …, Schlesner M, …, Stratton MR (2013) Signatures of mutational processes in human cancer. **Nature** 500(7463): 415-421

Jones DT, Hutter B, Jäger N, …., Schlesner M, …, Eils R[#], Lichter P[#], Pfister SM[#] (2013) Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet.**, 45(8):927-32

Richter J*, Schlesner M*, …, Siebert R (2012) Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. **Nat Genet.** 44(12):1316-20, *: shared first authors

Otto C, …, Schlesner M, …, Küppers R (2012) Genetic lesions of the TRAF3 and MAP3K14 genes in classical Hodgkin lymphoma. **Br J Haematol.** 157(6):702-8

Curriculum vitae

Name	Adam Grundhoff	Date of Birth	February 1, 1969

Affiliation			
Institution	Heinrich-Pette-Institute, Leibniz Institute for Experimental Virology (HPI)		
Dept./Institute	Research Group Virus Genomics		
Address	Martinistrasse 52, 20251 Hamburg, Germany		
Tel.	++49 (0)40 48051 275	Fax	++(0)40 48051 296
E-Mail	Adam.Grundhoff@hpi.uni-hamburg.de		

Academic Education	
1995	Diploma (Masters equiv.) in Biology, University of the Saarland, Germany
1999	Dr. rer. nat. (Ph.D. equiv.) in Biology, University of the Saarland, Germany

Professional Career	
1996-1999	Graduate studies, Inst. of Virology (Prof. Grässer), Univ. des Saarlandes, Germany
1999-2004	Postdoc, Howard Hughes Medical Institute (Prof. Ganem), UCSF, San Francisco, USA
2004-2005	Assistant Specialist, G.W. Hooper Foundation, UCSF, San Francisco, USA
2005-2011	Independent junior group leader, HPI, Hamburg
since 2011	Head of Research Group Virus Genomics, HPI, Hamburg
since 2013	Professor (tenured) of Virus Genomics, HPI, Hamburg

Major Research Topics
- Development of bioinformatic and experimental methods for NGS-based detection and analysis of infectious agents
- Mechanisms of persistence and transformation by human DNA tumor viruses
- Epigenetic control of viral latency
- Identification and characterization of viral miRNAs

Major Awards or Functions	
2000-2004	Howard Hughes Medical Institute Postdoctoral Fellowship
2012	Loeffler-Frosch Award of the Society of Virology (GfV)

Five Most Relevant Publications Pertaining to Proposal	
1	Sullivan CS, **Grundhoff A**, Pipas JM and Ganem D (2005). SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. Nature 435(7042):682-6.
2	Fischer N, Brandner J, Fuchs F, Moll I and **Grundhoff A** (2010). Detection of the Merkel cell polyomavirus (MCPyV) in Merkel cancer cell lines: cell line morphology does not reflect the presence of the virus. Int J Cancer 126(9):2133-42.
3	Gunther T, and **Grundhoff A** (2010). The epigenetic landscape of latent Kaposi sarcoma-associated herpesvirus genomes. PLoS Pathog 6, e1000935.
4	Walz N, Christalla T, Tessmer U, and **Grundhoff A** (2010). A global analysis of evolutionary conservation among known and predicted gammaherpesvirus microRNAs. J. Virol. 2010 Jan;84(2):716-28.
5	Bruce AG, Ryan JT, Thomas MJ, Peng X, **Grundhoff A**, Tsai CC, Rose TM (2013).Next-Generation Sequence Analysis of the Genome of RFHVMn, the Macaque Homolog of Kaposi's Sarcoma (KS)-Associated Herpesvirus, from a KS-Like Tumor of a Pig-Tailed Macaque. J Virol. 87(24):13676-13693.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

A population based reconstruction cancer genome evolution

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Martin Peifer, Department of Translational Genomics, University of Cologne, Germany

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yupeng Cun, José J Fernández-Melgarejo, Department of Translational Genomics	**N/A**

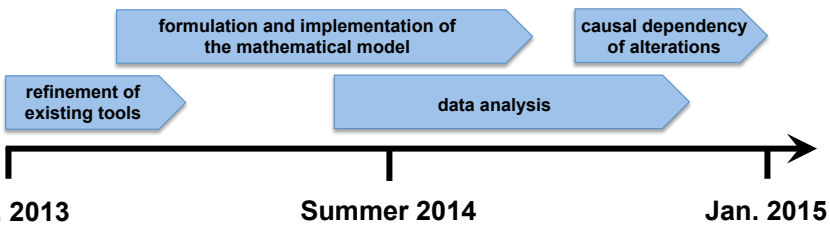
Background and preliminary data

Genomes of cancer cells undergo a succession of complex changes, such as point mutations, insertions, deletions, copy number alterations, and rearrangements. During tumorgenesis, these changes progress in time and are formed by evolutionary processes such as adaption and selection. Data derived form genome sequencing not only provides a comprehensive picture of the aforementioned alterations but also yields substantial insight into evolutionary forces reshaping cancer genomes. Computational methods designed to reconstruct tumor evolution from genome sequencing data have been proposed lately (e.g., Greenman CD, et al *Genome Research* 22, 2012). These methods are, however, mainly restricted to single samples. Instead, we propose a population-based method to reconstruct cancer genome evolution. To this end, we will formulate a mathematical model that is intended to capture evolutionary mechanisms underlying point mutations as well as structural changes. Estimates about tumor heterogeneity will also be incorporated into this model. Using this approach we aim to derive a *molecular time* for each analyzed tumor. As minimal requirement we impose that the real temporal order of serial biopsies is preserved. We envision that a well-defined *molecular time* yields information about the causal interplay between somatic mutations and may even contribute to the identification of new driver genes. The ICGC pan-cancer initiative is ideally designed for this project, since the large number of tumors across different lineages provides an optimal basis to formulate and calibrate the underlying mathematical model. Furthermore, consistency of the model can be evaluated by the ~200 additional cancer genomes related to different locations and time points.

Among our current cancer genome analysis framework (Peifer M, et al. *Nature Genetics* 44, 2012) we developed a method to derive a genome-wide picture of allelic states and tumor heterogeneity. In particular this module provides the foundation of our population-based evolutionary model. Furthermore, our initial analyses of serial biopsies suggest the feasibility to define a consistent estimate of the *molecular time* from genome sequencing data.

Timelines & resources dedicated to project

Timeline and key milestones:



Resources: The proposed analysis depends on mutation as well as rearrangement calls. To apply our methods to estimate tumor purity/ploidy and local patterns of heterogeneity, we need to process bam-files within the cloud-computing framework. Derived data will be further processed locally.

Research proposal

The basic concept behind our model is that cancer cells accumulate more and more genomic alterations during their lifespan. The overall mutational burden therefore should increase over time. However, just counting events is too simplistic and would result in an unavoidable bias of the *molecular time*. This is mainly due to the complex nature of the observed genomic alterations: sub-clonal events evolving differently from clonal, there is pronounced difference of mutation frequency between coding and non-coding regions, structural alterations are showing a rich variety of diverse patterns. It is therefore curial for the proposed project to derive a "microscopic" model that captures all known mechanisms describing the origins of the observed molecular events.
As data source of this model, we will use mutation and rearrangement calls provided by the ICGC pan-cancer project. To capture tumor heterogeneity as well as purity/ploidy corrected copy numbers, we are planning to refine and apply our own tools. In order to obtain these structural measures, we recently formulated a model that combines genotype with copy number (inferred by reads depth analysis) information. To improve the identification of copy number change points we are currently in the process to incorporate genomic rearrangements into the model as well. In addition, we plan to deploy a genotype-based segmentation approach to improve the resolution of allelic state boundaries, especially in heterogeneous parts of the tumor. Moreover, by adding somatic mutations to the model we aim to increase robustness of the purity and ploidy estimates (especially in copy number quiet tumors). As side effect, a better distinction between clonal- and sub-clonal mutations might also be achieved. To share results and to improve our methodology we are planning to cooperate tightly with the structural variations working group.
Next, we are aiming to derive a mathematical model that is capable to delineate the temporal succession of genomic events from the complete catalogue of somatic mutations. In particular, the model has to differentiate between clonal and sub-clonal events as well as between point mutations in coding and non-coding areas of the genome. Structural alterations are more difficult to incorporate into our model. To this end, we will translate known mechanisms of copy number changes (e.g., Hastings PJ, et al. *Nature Reviews Genetics* 10, 2009) into a mathematical description. Importantly, new modes of tumor evolution such as chromothipsis (Stevens PJ, et al. *Cell* 144, 2011) and chromoplexy (Baca, et al. *Cell* 153, 2013) require a separated mathematical description since their cause is unlikely to be linked to a continuous model of tumor evolution. After having formulated and implemented the model, parameters will be calibrated across the entire population of samples. This results in an estimate of the *molecular time* for each tumor included in the ICGC pan-cancer analysis and serves a basis to examine causal dependencies between observed genomic alterations. Please note that the proposed *molecular time* does not necessarily coincide with the "real" time span from tumor initiation to resection. It is only required that the temporal order of serial, untreated tumors is preserved. This consistency check will be performed by analyzing ICGC samples where different time points are available.

Possible pitfalls: In case we will not meet the planned time frame to define a consistent measure of *molecular time*, the ICGC pan-cancer analysis may benefit from our analysis by obtaining estimates about absolute copy numbers, allelic states, purity, ploidy, and heterogeneity. In addition, insights gained from the mathematical modeling might be useful for other working groups as well.

Legacy plans

Together with existing developments, we will provide the source code of our computational methods. All tools will be written under C++ and will run under different computer systems (e.g., Linux, Mac OS X). To ensure integrity of the source code, all developments will be made under a strict version control hosted by our institution. We will further provide a comprehensive documentation of the developed code including all mathematical details. In order to make the method available for the research community, it is planed to upload the source code to a public repository such as GitHub (**http://github.com**). Note that we will not include any commercial components or libraries and therefore achieve compatible with the General Public License guidelines.

Name: **Martin Peifer**
Institution: Department of Translational Genomics, University of Cologne
Degree: Dr. rer. nat.
E-mail: mpeifer@uni-koeln.de
Day of Birth: 16th December 1975
Place of Birth: Trier, Germany
Nationality: German

Education

Feb. 2007: Ph.D. in Physics at the Faculty of Mathematics and Physics of the Albert-Ludwigs University Freiburg, Germany. Graduated with summa cum laude.

Feb. 2003: Diploma in theoretical physics at the Faculty of Mathematics and Physics of the Albert-Ludwigs University Freiburg, Germany.

Professional Details

Principle Investigator

since July 2013 Career Advancement Program of the Center for Molecular Medicine Cologne, University of Cologne.

Post-Doctoral Training

since April 2012: Postdoc at the Department of Translational Genomics, University of Cologne.

June 2008 – March 2012: Postdoc at the Max-Planck Institute for Neurological Research, Cologne.

Feb. 2007 – June 2008: Postdoc at the Institute of Chemistry of the Karl-Franzens-University Graz, Austria.

Third party funding

July 2012 – June 2015: Deutsche Krebshilfe: Comprehensive molecular and histopathological characterization of small-cell lung cancer. (Subproject 1); **450,000 Euro**.

Jan. 2014 – Dez. 2016: Center for Molecular Medicine Cologne: Systematic identification of significantly mutated gene groups in cancer genomes by co-expression patterns; **176,400 Euro**.

Jan. 2014 – Dez. 2016: German Ministry of Education and Research (BMBF), e:Med initiative: Systems-level modeling of cancer genome evolution; **300,000 Euro**.

Selected papers

1. Danila Seidel*, Thomas Zander*, Lukas C Heukamp*, **Martin Peifer***, et al. A genomics-based classification of human lung tumors. ***Science Translational Medicine*** 5: 209ra153 (2013). *** equal contribution**

2. **Martin Peifer**, Lynnette Fernandez-Cuesta, Martin L Sos, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. ***Nature Genetics*** 44:1104-1110 (2012).

3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. ***Nature*** 489:519-525 (2012).

4. Kwon-Sik Park, Luciano G Martelotto, **Martin Peifer**, et al. A crucial requirement for Hedgehog signaling in small cell lung cancer. ***Nature Medicine*** 17:1504-1508 (2011).

5. Juliann Chmielecki, **Martin Peifer**, Peilin Jia, et al. Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer. ***Nucleic Acids Research*** 38:6985-6996 (2010).

Yupeng Cun

University of Cologne
Department of Translational Genomics
Weyertal 115b
50931 Cologne, Germany

Tel.: +49 (0) 221 478 98782
email: ycun@uni-koeln.de

Personal information

Date of Birth: 24.09.1981
Citizenship: P.R.C.

Academics degrees

- 09.2010 - 09.2013, PhD student in Computational life science, University Bonn, Germany.
  Thesis title: Network-based Biomarker Discovery
  Advisor: Prof. Dr. Holger Fröhlich
- 07.2007, Master degree in Physics, Yunnan University, Kunming, China.
  Thesis title: Analysis of the fixation time of mutant allele at duplicate loci
  Advisor: Prof. Dr. Yun-Xin Fu
- 07.2004, Bachelor degree of Computer Science, Yunnan University, Kunming, China.

Academics work experiences

- *Since 09.2013:* Postdoc at department of translational genomics, University of Köln.
  Research topics: Computational cancer genomics.
  Advisor: Prof. Dr.Thomas Roman and Dr. Martin Peifer
- *09.2010 - 09.2013:* PhD Student at B-IT, University of Bonn.
  Research topics: Developing statistical learning models for prognostic biomarker discovery.
  Advisor: Prof. Dr. Holger Fröhlich
- *08.2009 - 07.2010:* Research Assistant at Beijing Institute of Genomics, CAS.
  Research topics: Population genomics models for speciation.
  Advisor: Prof. Dr. Chung-I Wu
- *08.2008 - 07.2009:* Staff scientist at CAS-MPG Partner Institute for Computational Biology.
  Research topics: Evolution and dynamics of protein family.
  Advisor: Dr. Frauke Gräter
- *07.2007 - 07.2008:* Research Assistant at Kunming Institute of Zoology, CAS.

Peer-reviewed journal publications

1. ***Yupeng Cun***, Holger Fröhlich (2013) netClass: An R-package for network based, integrative biomarker signature discovery. (Acceptd by ***Bioinformatics***)
2. ***Yupeng Cun***, Holger Fröhlich (2013) Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics, ***PLoS One***, dio:10.1371/journal.pone.0073074
3. ***Yupeng Cun***, Holger Fröhlich (2012) Prognostic Gene Signatures for Patient Stratification in Breast Cancer - Accuracy, Stability and Interpretability of Gene Selection Approaches Using Prior Knowledge on Protein-Protein Interactions. ***BMC Bioinformatics*** 13:69 doi:10.1186/1471-2105-13-69
4. ***Yupeng Cun*****,** Holger Fröhlich (2012) Biomarker Gene Signature Discovery Integrating Network Knowledge, ***Biology*** 1, no. 1: 5-17. doi:10.3390/biology101000517. doi:10.3390/biology1010005

J.J. Fernández-Melgarejo

CV

B *jj.fernandezmelgarejo@gmail.com*

Personal Data

First Name **José Juan**.

Family Name **Fernández Melgarejo**.

Date of Birth **26-01-1985**.

Place of Birth **Murcia**, *Spain*.

Marital status **Single**.

Business address **Departamento de Física, Campus de Espinardo**, *Universidad de Murcia*, 30100 Murcia, Spain.

Personal address **Mayor, 67, 2A**, *30830*, La Ñora, Murcia, Spain.

Mobile Phone **0034 616769811**.

Phone **0034 868888094**.

E-mail **jj.fernandezmelgarejo@um.es**.

Education

2009-2013 **PhD in Theoretical Physics, European Doctorate "*summa cum laude*"**, *University of Murcia*, Thesis: Gaugings and other aspects in Supergravity.
Supervisors: Emilio Torrente-Luján, Tomás Ortín

2008-2009 **MSc in Theoretical Physics**, *University of Valencia*, , Master Thesis: Dark Matter in the Universe and its direct detection.
Supervisor: J. Bernabéu

2003-2008 **Bachelor of Physics**, *University of Murcia*, Murcia, .
Student internship in Particle Physics and Cosmology group for 2 years. Collaboration with Chemical Physics Group.

Research and academic positions

2009-2013 **FPU Predoctoral fellowship**, *University of Murcia*, Funded by Spanish Ministry of Education.

2008 **Research technician**, *Department of Theoretical Physics, University of Valencia - IFIC*.

2008 **Assitant professor**, *Department of Physics, University of Murcia*.

Research Experience

2009-2013 **University of Murcia, IFT - UAM (Madrid)**, *Advisors: E. Torrente-Luján, T. Ortín*.
I have studied gauged supergravities and embedding tensor formalism. After that, I have studied $N = 2$ black holes and specially multi-center black holes.

2012 **Queen Mary University of London**, *Advisor: D. Berman*.
I have studied non-geometric fluxes and generalized complex geometry for 4 months. My stay in this department was for 4 months.

2011 **University of Groningen**, *Advisors: E. Bergshoeff, D. Roest*.
I have studied massive gravity with Prof. Bergshoeff and the implementation of double field theory (DFT) to explain the gaugings of several gauged supergravities that do not have a geometric origin. Mi stay in this university was for 4 months.

2008-2009 **University of Valencia, Department of Theoretical Physics**, *Advisor: J. Bernabéu*.
I have studied the problem of Dark Matter, the possible detection methods and the status of DAMA/LIBRA experiment.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mutations in regulators of the epigenome and their effects on the DNA methylome

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Christoph Plass, Division of Epigenomics and Cancer Riskfactors, DKFZ, Heidelberg (Co-PI: ICGC EOPCA)
Benedikt Brors, Division of Theoretical Bioinformatics, DKFZ, Heidelberg (PI: ICGC EOPCA)

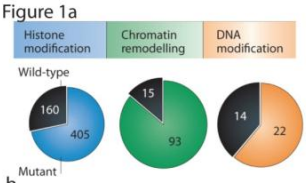
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Olga Bogatyrova and Lars Feuerbach, DKFZ, Heidelberg	Y. Assenov (DKFZ), T. Wang (WashU), J. Costello (UCSF)

Background and preliminary data

Alterations in the epigenetic regulation of genome activity, are as important to tumorigenesis as the genomic coding information itself [1,2]. High-resolution genome sequencing efforts have discovered a wealth of mutations in genes encoding epigenetic factors that are involved in DNA methylation and/or modulation of chromatin states either as writers, readers or editors of these modifications (e.g. *H3F3A*[3], *IDH1*[4]). Elucidation of these networks of epigenetic factors will provide a mechanistic understanding of the interplay between genetic and epigenetic alterations and will inform novel therapeutic strategies. For this work a bioinformatics group (Brors) and an epigenomics group (Plass) will engage in a collaborative effort.

Preliminary data: In preliminary data from our groups have created a list of about 700 genes/enzymes involved in the establishment of epigenomes. These genes are involved in DNA methylation, histone modifications or chromatin remodeling (**Figure 1a**). Up to 85% of these genes have been found mutated in cancer. The frequencies of mutations in this gene set differ between tumortypes. We have established a scoring system that allows us to rank these genes based on normalized mutation frequencies, functional impact, type of alterations, oncogene or tumorsuppressor function in order to define affected driver genes (**Figure 1b**). Our data analysis, based on 13 cancer types from TCGA data sets, is currently in preparation for submission. Our key-findings are: 1. The identification of cancer driver epigenetic genes and tissue specific epigenetic genes. 2. using integrative data analysis we have subdivided in TSGs and oncogenes. 3. We have created a list of hotspots mutations, based on the frequency and functional impact of mutations. 4. We have identified methylation subtypes in TCGA methylomes and correlated them with deregulated epigenetic genes. In particular our work on H3.3 mutations which we performed together with groups of Dr. Lichter and Pfister (both ICGC embers and at DKFZ) represent examples of such work including functional studies.

Figure 1a



b
List of most frequently mutated candidate drivers in the cohort

gene	fm-bias	clust-bias	mut-freq	CGC	intogen
TP53	< 1E-16	0.367	0.26	CGC Rec	Driver
TTN	3.221E-6	0.978	0.243		Driver
APC	< 1E-16	0.903	0.172	CGC Rec	Driver
GATA3	< 1E-16	0.966	0.082	CGC Rec	Driver
KMT2C	8.341E-15	0.971	0.071	CGC Rec	Driver
OBSCN	5.938E-6		0.069		Driver
NCOR2	0.015		0.059		
MAP3K1	< 1E-16	0.966	0.059		Driver
CDH1	< 1E-16	0.966	0.058	CGC Rec	Driver
DST	6.65E-6		0.052		Driver

Timelines & resources dedicated to project

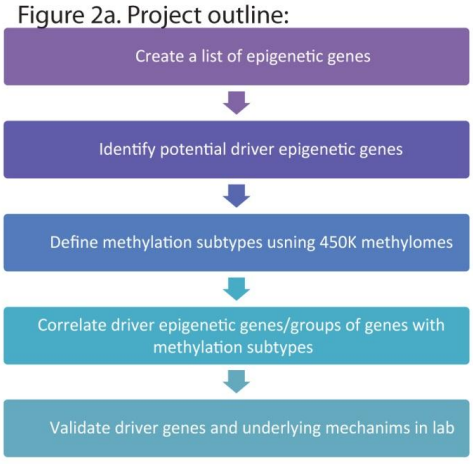
Compute the list of deregulated gene using integrative algorithm and identify potential driver genes (3 month)
Identify methylome subgroups and correlate with driver epigenetic genes/groups of genes (6 month)
Prediction of molecular function of mutation and validation in laboratory settings (12 month)
Visualization of data for ICGC community by integrating data to WashU epigenome browser (12 month)
Manuscript writing (3month)
Human Resources: collaboration between epigenomics group (Christoph Plass) and bioinformatics group (Benedict Brors). Additional expertise from the epigenomics group for functional experiments will be provided by Chris Oaks and Anders Lindroth.
Computer Resources: Our previous study on methylation and gene mutation data from TCGA shows that the available compute cluster architecture at DKFZ is sufficient for performing all analyses described above.

Research proposal

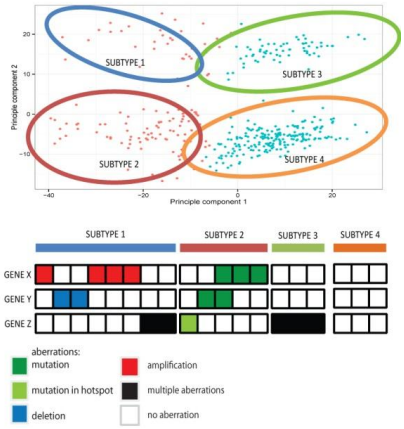
Based on our current knowledge in the field of cancer biology, we are proposing the **hypothesis** that mutations in enzymes that regulate epigenetic pathways have a chain reaction like effect on the epigenome of the tumor cell which is manifested in altered DNA methylation patterns of the tumor genome. To address this hypothesis we propose a systematic evaluation of the available cancer genome profiling data established by the ICGC consortia, in order to identify recurrently mutated regulators of epigenetic pathways. To support this approach we have established a list of about 700 epigenetic genes that are involved in either establishing an epigenetic mark (writers), modify these epigenetic marks (editors), or translate the epigenetic mark into a molecular signal (reader). Using ICGC genome-wide datasets on genetic and epigenetic alterations in cancers, we will systematically evaluate the genetic and epigenetic datasets using the following steps:

1. describe the distribution of mutations including gene mutations, copy number alterations, promoter methylation, in a pan-cancer study.
2. classify genes as potential oncogenic or those with tumor-suppressor function based on the location of mutations relative to functional domains and their frequency. Using this datalist and scoring system we will define driver epigenetic genes.
3. determine if mechanisms of deregulation of epigenetic genes/pathways occur in a non-random (e.g. KDMs) and tumor-type specific (e.g. H3.3 mutations in pediatric glioblastomas) manner.
4. correlate affected genes or groups of genes with tumor subtypes based on methylation profiling data (450K Illumina arrays) as seen for example for IDH mutations, K27M or G34V mutations in pediatric glioblastomas approach to defects in regulators of the epigenome will help to understand mechanisms leading to distinct epigenetic patterns.
5. use molecular assays to validate underlying mechanisms leading to epigenetic subgroups, such as cell culture experiments (knock – down and overexpression), use mutant cell lines, biochemical activity assays, mouse models and epigenomic profiling.

Figure 2a. Project outline:



b. PCA analysis of THCA methlyomes



References:

1. Plass, C. et al. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat Rev Genet* 14, 765-80 (2013).
2. Shen, H. & Laird, P.W. Interplay between the cancer genome and epigenome. *Cell* 153, 38-55 (2013).
3. Sturm, D. et al. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* 22, 425-37 (2012).
4. Figueroa, M.E. et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* 18, 553-67 (2010).

Legacy plans

- All algorithms, computational pipelines and scripts will be made available to the scientific community.
- Mutation data will be included into the **WashU epigenome browser**
- Methylation genome wide data will be made available through **WashU epigenome browser and UCSC/IGV visualization tools.**
- Association between driver epigenetic genes and genome wide methylation data will be available through convenient web interfaces
- Hotspots mutation and will be made available through the **cBio** browser

Prof. Dr. Christoph Plass

Head of Division Epigenomics and
Cancer Risk Factors
German Cancer Research Center (DKFZ)
E-Mail: c.plass@dkfz.de

General Information

Born July 30th 1961 in Bremen, Germany
Nationality: German
Current Position: Full Professor (W3)

Degree

1988-1993	University Lübeck, Ph.D.
1982-1987	University Berlin, Diploma

Scientific Career

2007-	German Cancer Research Center (DKFZ), Heidelberg, Germany, Department of Epigenomics and Cancer Risk Factors, Professor
2005-2007	The Ohio State University, Columbus, USA, Department of Medical Microbiology and Immunology, Division of Human Cancer Genetics, Professor
2002-2005	The Ohio State University, Columbus, USA Department of Medical Microbiology and Immunology, Division of Human Cancer Genetics, Associate Professor
1997-2002	The Ohio State University, Columbus, USA, Department of Medical Microbiology and Immunology, Division of Human Cancer Genetics, Assistant Professor
1996-1997	Roswell Park Cancer Institute, Buffalo, NY, Molecular and Cellular Biology Department, Cancer Research Scientist II
1993-1996	Roswell Park Cancer Institute, Buffalo, NY, Laboratory of Dr. Verne Chapman, Molecular and Cellular Biology Department, Postdoc

Selected Publications

Bender, S. ….Plass C, Cho YJ, Pfister S. Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. **Cancer Cell** 2013, 11, 660-672

Plass, C. et al. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. **Nat Rev Genet** 14, 765-80 (2013).

Weischenfeldt J, …, **Plass C**, et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. **Cancer Cell** 2013, 23, 159-170.

Schwartzentruber J, …, **Plass C**, et al. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. **Nature** 2012, 482, 226-231.

Sturm, D…….**Plass C**. Jabado N, Pfister S. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell 22, 425-37 (2012).

Raval A, … **Plass C.** Down-regulation of death associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. **Cell** 2007, 129, 879-890.

Costello JF. ……**Plass C**. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. **Nature Genetics**, (2000). 24):132-138

Dr. Benedikt Brors

Group Leader Computational Oncology
Div. Theoretical Bioinformatics
German Cancer Research Center, Heidelberg

E-Mail: b.brors@dkfz.de

Degree

1989–1995	Diploma in chemistry, University of Düsseldorf, Germany
1999	Doctoral degree (Dr. rer. nat) in biochemistry, University of Düsseldorf

Scientific Career

1995–1999	Pre-doctoral research assistant, Inst. of Biochemistry, University of Düsseldorf
1999–2000	Postdoctoral Researcher, Div. Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
2000–2003	Postodoctoral Researcher, Intelligent Bioinformatics Systems, DKFZ
2003–present	Group Leader Computational Oncology, Div. Theoretical Bioinformatics, DKFZ
2003–present	Lecturer in bioinformatics, Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg
2007	Offer of a post as Full Reader in Medical Bioinformatics, Queen's University, Belfast, UK (not realized)
2008	Offer of a post as W2 professor of biostatistics (non-tenured), University Medical Center Göttingen, Germany (not realized)
2009–present	Tenured research position, DKFZ
2013	Offer of a position as full professor of applied bioinformatics, DKFZ and National Center for Tumor Diseases, Heidelberg (under negotiations)

Selected Publications

Jones, D.T.W.,* **Hutter, B.*, Jäger, N.***, (...), **Brors, B.**, (...), Eils, R., Lichter, P., Pfister, S.M., 2013. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet.** 45, 927-932

Weischenfeldt, J.*, Simon, R.*, **Feuerbach, L.***, (...), Sültmann, H.#, Sauter, G.#, Plass, C.#, **Brors, B.#**, Yaspo, M.-L.#, Korbel, J.O.#, Schlomm, T.#, 2013. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. **Cancer Cell** 23, 159–170.

Richter, J., **Schlesner, M.***, (...), Hummel, M#., Klapper, W.#, Rosenstiel, P.#, Rosenwald, A.#, **Brors, B.#**, Siebert, R.#, 2012. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. **Nat. Genet.** 44, 1316–1320.

Jones, D.T.W.*, **Jäger, N.***, (...), **Brors, B.**, (...), Eils, R., Pfister, S.M., Lichter, P., 2012. Dissecting the genomic complexity underlying medulloblastoma. **Nature** 488, 100–105.

Oberthuer, A., Hero, B., Berthold, F., **Juraeva, D.**, (...), **Brors, B.**, Fischer, M., 2010. Prognostic impact of gene expression-based classification for neuroblastoma. **J. Clin. Oncol.** 28, 3506–3515.

(* contributed equally; # contributed equally as senior authors)

Relevant Projects

- PI (bioinformatics analysis) in ICGC-PedBrain
- PI (bioinformatics analysis) in ICGC – Early Onset Prostate Carcinoma
- PI (bioinformatics analysis) in ICGC – Molecular Mechanisms in Malignant Lymphoma
- Member of ICGC Bioinformatics Work Group and Mutation Consequences and Pathways WG

Olga Bogatyrova

Division Epigenomics and
Cancer Risk Factors
German Cancer Research Center (DKFZ)
E-Mail: o.bogatyrova@dkfz.de

General Information

Born October 27th 1987 in Cherkassy, Ukraine
Nationality: Ukrainian
Current Position: PhD Student

Scientific Career

2011-	Helmholtz International Graduate School for Cancer Research German Cancer Research Center (DKFZ), Heidelberg, Germany Division of Epigenomics and Cancer Risk Factors (Prof. Dr. Christoph Plass)
2009-2011	Taras Shevchenko Kyiv National University (KNU), Ukraine Biological Department, specialization: Molecular Biology (Dr. Tatyna Andreychuk) Collaboration project with Institute of Molecular Biology and Genetics NAS of Ukraine (IMBiG) (Prof. Dr. V.I. Kashuba), (Diploma with honors degree)
2005-2009	Taras Shevchenko Kyiv National University (KNU), Ukraine Biological Department, specialization: Biochemistry (Prof. Dr. Olga Matishevskaya) Collaboration project with Institute of Molecular Biology and Genetics NAS of Ukraine (IMBiG) (Prof. Dr. V.I. Kashuba), (Diploma with honors degree)
2005-2011	Taras Shevchenko Kyiv National University (KNU), Ukraine, Biological Department, Pedagogical studies: High School Teacher of Biology and Chemistry, (Diploma with honors degree)
2000-2005	Cherkassy Physics and Mathematics lyceum (PhyMLy), Ukraine Grammar School (Graduated with Honors)
1995-2000	Grammar School: Cherkassy college "Berehinya", Ukraine (Graduated with Honors)

Selected Publications

Plass, C.,.. **Bogatyrova O.**, Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. **Nat Genet review**, 2013.

Kostareli E, Holzinger D, **Bogatyrova O**, .., Hess J. HPV-related methylation signature correlates with survival in oropharyngeal squamous cell carcinomas, **J Clin Invest.** 2013 May 1. doi:pii: 67010. 10.1172/JCI67010

Goeppert B, Konermann C, Schmidt CR, **Bogatyrova O**,…, Weichenhan D. Global alterations of DNA methylation in cholangiocarcinoma targets the Wnt signaling pathway **Hepatology**, 2013

Weischenfeldt, J.,…**Bogatyrova O**., et al., Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. **Cancer Cell**, 2013. 23(2): p. 159-70.

E. Braga, W. Loginov, **O. Bogatyrova** et al**.** A novel meca3 region in human 3p21.3 harboring putative tumor suppressor genes and oncogenes**. Exp Oncol** 2011, 56, 3, 56-61.

Lars Feuerbach – Curriculum Vitae

Phone: +49 6221 42 3603 E-mail: l.feuerbach@dkfz.de

RESEARCH POSITIONS

10/2011-	Research Fellow – ICGC Early onset Prostate Cancer Computational Oncology, Theoretical Bioinformatics German Cancer Research Center, Heidelberg, Germany
10/2007-09/2011	Graduate Research Assistant Computational Biology and Applied Algorithms Department Max-Planck Institut für Informatik, Saarbrücken, Germany

EDUCATION

10/2007 –	PhD in Bioinformatics (Thesis submitted – 08/2013) Max-Planck Institut für Informatik, Saarbrücken, Germany
10/2005 – 09/2007	Master of Science (MSc) with Honor's Degree - Bioinformatics Center for Bioinformatics, University of Saarland, Germany
09/2002 – 07/2005	Bachelor of Science (BSc) - Bioinformatics Free University of Berlin, Germany

SELECTED PUBLICATIONS

Joachim Weischenfeldt*, Ronald Simon*, Lars Feuerbach*, Karin Schlangen*, et al.
Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer
Cancer Cell, 2013, 23(2):169-170

Lars Feuerbach, Konstantin Halachev, Yassen Assenov, Fabian Müller, Christoph Bock, Thomas Lengauer
Analyzing epigenome data in context of genome evolution and human diseases
Methods Mol. Biol. 2012,856:431-67

Malay Bhattacharyya, Lars Feuerbach, Tapas Bhadra, Thomas Lengauer, Sanghamitra Bandyopadhyay
MicroRNA Transcription Start Site Prediction with Multi-objective Feature Selection
Statistical Applications in Genetics and Molecular Biology, 2012, 11(1) 1–25

Lars Feuerbach, Rune B. Lyngsoe, Thomas Lengauer, Jotun Hein
Reconstructing the ancestral germline methylation state of young repeats
Molecular biology and evolution 2011;28(6):1777-84

Pavlo Lutsik, Lars Feuerbach, Julia Arand, Thomas Lengauer, Jörn Walter, et al.
BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing.
Nucleic Acids Research, May 11, 2011, 39(Web Server issue):W551-6

Dr. Yassen Assenov

Division Epigenomics and
Cancer Risk Factors
German Cancer Research Center (DKFZ)
E-Mail: y.assenov@dkfz.de

General Information

Born December 4th 1981
Nationality: Bulgarian
Current Position: Research Assistant

Degree

2007-2012 Ph.D in Bioinformatics

Scientific Career

2013-	German Cancer Research Center (DKFZ), Heidelberg, Germany, Department of Epigenomics and Cancer Risk Factors
2007-2012	Ph.D in Bioinformatics Saarland University & Max Planck Institute for Informatics, Saarbrücken, Germany
2004-2006	M.Sc. in Computer Science, GPA 1.1 (Best: 1.0) Saarland University, Department of Computer Science & International Max Planck Research School
2003-2004	Sokrates (exchange) student at the Catholic University Eichstätt-Ingolstadt
1999-2003	B.Sc. in Computer Science, GPA 5.96 (Best: 6.00) Sofia University “St. Kliment Ohridski”, Faculty of Mathematics and Informatics
1998-1999	St. Lawrence College, Athens, Greece
1992-1999	High School of Mathematics “Dr. Petar Beron”, Varna, Bulgaria; Profiles: Mathematics, Computer Science, English

Selected Publications

Oakes, CC, Claus, R, Gu, L, Assenov, Y, Hüllein, J, Zucknick, M, Bieg, M, Brocks D, Bogatyrova O, Schmidt C, Rassenti, L, Kipps, TJ, Mertens, D, Lichter, P, Döhner, H, Stilgenbauer, S, Byrd, JC, Zenz, T, Plass, C. Heterogeneity and evolution of DNA methylation are linked to genetic aberrations in chronic lymphocytic leukemia. Cancer Discovery, in press

Fernandez, AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo M, Tan AC, Galm O, Ferrer I, Sanchez-Cespedes M, Villanueva A, Carmona J, Sanchez-Mut JV, Berdasco M, Moreno V, Capella G, Monk D, Ballestar E, Ropero S, Martinez R, Sanchez-Carbayo M, Prosper F, Agirre X, Fraga MF, Graña O, Perez-Jurado L, Mora J, Puig S, Prat J, Badimon L, Puca AA, Meltzer SJ, Lengauer T, Bridgewater J, Bock C, Esteller M. A DNA methylation fingerprint of 1628 human samples. Genome Research, 22(2):407-419, 2012

Assenov, Y, Ramírez, F, Schelhorn, SE, Lengauer, T, Albrecht, M. Computing topological parameters of biological networks. Bioinformatics, 24(2):282-284, 2008

Salamat-Miller, N, Fang, J, Seidel, CW, Assenov, Y, Albrecht, M, Middaugh, CR. A network-based analysis of polyanion-binding proteins utilizing human protein arrays. The Journal of Biological Chemistry, 282(14):10153-10163, 2007

Prof. Ting Wang

Assistant Professor of Genetics and of
Computer Sciences and Engineering
Washington University School of Medicine
E-Mail:twang@genetics.wustl.edu
Website: http://wang.wustl.edu/

General Information

Born July 31, 1973
Nationality: USA
Current Position: Assistant Professor

Degree

2001-2006	Washington University in St. Louis, Ph.D. in Computational Biology
1999-2001	Washington University in St. Louis, M.S. in Computer Sciences
1993-1997	Peking University, B.S. in Biochemistry and Molecular Biology

Scientific Career

2009-	Assistant Professor, Department of Genetics, Washington University School of Medicine
2006-2009	Helen Hay Whitney Fellow, Laboratory of Dr. David Haussler, University of California at Santa Cruz
2001-2006	Pre-doctoral Researcher, Laboratory of Dr. Gary Stormo, Washington University School of Medicine

Selected Publications

Stevens M, .., **Wang T**. (2013) Estimating absolute methylation levels at single CpG resolution from methylation enrichment and restriction enzyme sequencing methods. **Genome Research** 2013 Sep;23(9):1541-53.

Zhang B, .., **Wang T**. (2013) Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. **Genome Research** 2013 Sep;23(9):1522-40.

Xie M, .., **Wang T**. (2013) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. **Nature Genetics** 2013 Jul;45(7):836-41.

Zhou X, .., **Wang T**. (2013) Exploring long-range genome interaction data using the WashU Epigenome Browser. **Nature Methods** 10(5): 375-376.

Zhou X., .., **Wang T**. (2012) Using the Wash U Epigenome Browser to examine genome-wide sequencing data. **Curr Protoc Bioinformatics.** 2012 Dec; Chapter 10:Unit10.10

Xiao S, .., **Wang T**, and Zhong S. (2012) Comparative epigenomic annotation of regulatory DNA. **Cell** 2012 Jun 8;149(6):1381-92.

Zhou X, .., **Wang T**. (2011) The human epigenome browser at Washington University. **Nat Methods** 2011, 8(12):989-990

Prof. Joseph F. Costello

Professor of Neurosurgery
Karen Osney Brownstein Endowed Chair
University of California San Francisco (UCSF)
E-Mail:jcostello@cc.ucsf.edu

General Information

Born November 30, 1965
Nationality: United States
Current Position: Full Professor

Degree

1990-1994 Loyola University Medical Center, Ph.D.
1982-1987 Marquette University, Bachelors of Science

Scientific Career

2000	Assistant Professor, Dept. of Neurological Surgery, University of California San Francisco
2005	Director, Epigenetics Division of the UCSF CCC Program in Cell Cycling and Signaling
2005	Associate Professor, Dept. of Neurological Surgery, University of California San Francisco
2005	Karen Osney Brownstein Endowed Chair in Molecular Neuro-Oncology
2008	Director, NIH Roadmap Epigenome Mapping Center
2010	Professor, Department of Neurological Surgery, UCSF
2012	BC Genome Sciences Center, Associate Member

Selected Publications

Johnson, B.E. **….Costello, J.F.** Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. **Science,** 2013 Dec 12. [Epub ahead of print]

Xie, M.,….***Costello, J.F.,** *Wang T. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. **Nature Genetics**. 2013 May 26..

Maeder M.L.,... **Costello, J.F.**, Wilkinson MF, Joung JK. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins.**Nature Biotechnology,** 2013 Dec;31(12):1137-42.

Maunakea, A.K.,...**Costello J.F**. Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters. **Nature**, 2010 Jul 8;466(7303):253-7. PMID: 20613842.

Harris, R.A.,...**Costello, J.F**. Comparison of Sequence-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. **Nature Biotechnology**, 2010 Oct;28(10):1097-105.

Costello JF. Stem Cells: Tips for priming potency (editorial). **Nature,** 2008 July 3, 454, 45-46.

Costello JF. ……Plass C. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. **Nature Genetics**, (2000). 24):132-138

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

PAN-CANCER TRANSPOSOME AND VIROME

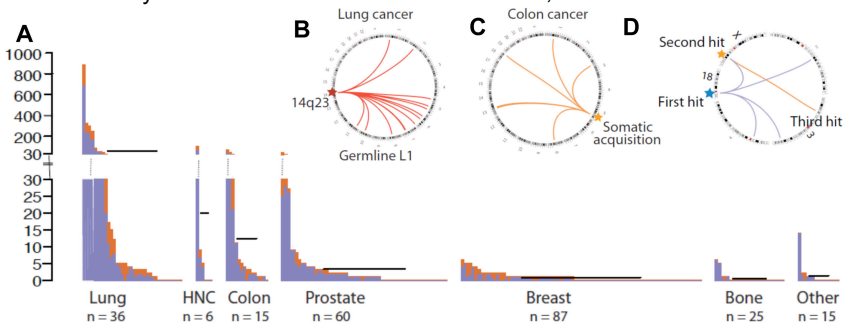
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Jan Korbel, EMBL Heidelberg: ICGC PedBrainTumor, MMML-Seq & early onset prostate cancer
Peter Campbell, Sanger Institute: ICGC breast, bone & chronic myeloid cancers

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Jose Tubio, Post-doctoral Fellow, Sanger Institute Jelena Tica, Pre-doctoral Fellow, EMBL	

Background and preliminary data

While most cancer genome studies so far have focused on the effects of point mutations and DNA rearrangements, there is ample evidence pointing to oncogenic roles of genomic insertions of DNA, including transposable element (TE) and viral insertions (1,2). Initial efforts to catalogue these have demonstrated viral insertions and somatic activity of TEs in several cancer entities and cell lines (3-5). Owing to the lack of comprehensive analyses across different cancer entities, it has thus far remained largely unclear to what extent genetic and/or environmental factors contribute to element insertion activities, and what their impact is on cancer evolution and progression.



We will assess the extent, impact and functional consequences of insertions of viral or TE origin with WGS pan-cancer data, generating both somatic and germline (6) insertion calls. We have developed powerful new approaches to uncover TEs, retrogenes (insertions of processed genic mRNAs), and viral insertions. In addition to identifying these insertions, our pipelines enable, for the first time, the detection of somatic TE-mediated translocations of unique DNA sequences, so called transductions, which based on our prior data are highly abundant in cancers (Fig. 1).

In our prior analyses we ascertained TEs in 290 WGS across 12 cancer types (7), identifying ~3,000 somatic L1 insertions, out of which 25% involved the transduction of unique DNA sequence (including functional elements such as exons). We further inferred retrogene insertions (e.g. *USP17* and *DGKB* retrogenes) in medulloblastoma, human Herpesvirus in lymphoma and ALL (unpublished data), and HPV integration in a cervical cancer cell-line (4).

Fig 1 (above). (A) L1 elements and transductions per sample. (B-C) Analysis of transductions enables the mapping of source loci, including secondary and tertiary hits caused by active L1s (D).

Timelines & resources dedicated to project

1. Element insertion calling somatic/germline (completed June 2014; CPU months per sample = 0.5); algorithms are already incorporated into the official (Sanger and Heidelberg) pan cancer pipelines.
2. Initial downstream data analyses to be initiated once 500 samples are complete (start: April 2014).
3. Analysis of entire set, and initial integration of RNA and DNA methylome data (to begin July 2014).
4. Assessment of transduction events to locate source elements (by detecting TE insertions involving unique sequence) by September 2014 (little compute resources required for this step).
5. Assembly of unmapped reads to infer non-integrated viruses (Oct. 2014; CPU months per sample = 0.5).
6. Analyse entire dataset from October 2014 onwards; provide data to others.
7. Present results at ICGC meeting in spring 2015, release source code, followed by paper submission.

Dedication to the project in terms of committed time: JoTu 100%, JeTi 50%, PC 5%, JK 10%.

Research proposal

We will identify somatic element (*i.e.* TE and virus-associated) insertions as well as non-integrating infecting viruses throughout the pan-cancer dataset, an analysis that will also encompass cataloguing transduction events – a relevant form of somatic DNA variation that in some cancer samples can exceed classical DNA rearrangements, and thus far has been overlooked. In addition to assessing the consequences of insertions and non-insertional infections on cancer genes, we will determine how insertion/infection correlates with genomic instability and tumor temporal evolution, and relate our findings to indicators for tumor progression (*e.g.*, clinical data).

Our work plan is as follows:

A. Investigate the relative representation of TEs and viral activity in different cancer types and subtypes. We will generate a comprehensive catalogue comprising the frequency and type of both somatic and germline TEs across cancers, including transductions of non-repetitive genomic material, and retroposed gene transcript insertion events. To this end, we have developed algorithms (TraFiC) (7,8) that identify ~25% more somatic element insertion events than previous pipelines (4) with an FDR<5%.

B. Investigate the frequency of viral infection in cancer. We will use our DELLY tool (8-9) to identify viral integrations, and additionally perform sequence assembly of unmapped reads to detect the presence of non-integrating viruses, with the aim to determine the abundance of viral infections in the WGS pan-cancer data.

C. Assess the functional impact of TEs and viruses. Analyses of the functional consequences of TEs and viral insertions will involve determining their effect on protein-coding gene structure, DNA methylation (TEs and viruses can be epigenetically silenced during or after their integration, thereby affecting the epigenetic status of adjacent regions (4)), and gene expression (using RNA-Seq data, an analysis that can be performed in collaboration with pan-cancer working group(s) focusing on gene expression data). We will further ascertain the clinical impact of insertions and non-insertional infections, *e.g.*, in collaboration with the clinical pan-cancer working group.

D. Characterize the temporal evolution of insertion events. Our preliminary analyses have shown that along tumour evolution, the activity of particular L1 element loci can be markedly altered, *i.e.* is significantly increased in more progressed tumours (7). Since the activation and deactivation of retrotransposition activities is correlated with DNA methylation changes (7), we will specifically correlate methylation profiles with TE activities in WGS-pan cancer sample triplets including tumours sampled prior to, and after, relapse – which, in addition to facilitating an understand of consequences of element insertions may help to establish new phenotypes linked to tumor evolution and progression,

E. Assessment of links between element insertion and genetic instability. We will further assess effects of insertions on genetic instability, since inserting elements may themselves act as a source of genetic instability by inducing ectopic recombination or through other molecular mechanisms (11). Using paired-end mapping (12-13) we will assess the frequency and locus specificity of element insertions and non-insertional infections across different cancer types.

REFERENCES:

(1) Miki et al. 1992. Cancer Res 52: 643-5; (2) Zur Hausen. 1991. Science 254: 1167-73; (3) Lee et al. 2012. Science 337: 967-71; (4) Landry et al. 2013. G3 3:1213-24; (5) Fujimoto et al. 2012. Nat Genet 44: 760-4; (6) Miki et al. 1996. Nature Genet 13: 245-7; (7) Tubio et al., submitted; (8) Murchison et al. Science, in press; (9) Rausch et al. 2012. Cell 148: 59-71; (10) Rausch et al. 2012. Bioinformatics 28: i333-i339; (11) Akagi et al. 2013. Genome Res, in press; (12) Korbel et al. 2007. Science 318: 420-6; (13) Campbell et al. 2008. Nat Genet 40: 722-9.

Legacy plans

1) VCF-format list with annotated TEs, retrogene, and viral sequences in cancer and germline genomes.

2) Jose Tubio will contribute visualization tool, integrating TEs, retrogenes, and viral sequence insertions for each cancer as well as across all cancer types (for pan-cancer working groups, and the cancer genomics community).

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

Since 2013	Fellow of the European Academy of Cancer Sciences.
2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci*** **USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

CURRICULUM VITAE

NAME	POSITION TITLE
Dr Peter J Campbell	Head of Cancer Genetics & Genomics, Wellcome Trust Sanger Institute

EDUCATION/TRAINING

FIELD OF STUDY	INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED
Mathematics and Statistics	University of Otago, New Zealand	BSc Hons (1st Class)	1994
Medicine	University of Otago, New Zealand	MB ChB (Distinction)	1995
Haematology	Royal Australasian College of Physicians	FRACP	2003
Haematology	Royal College of Pathologists of Australasia	FRCPA	2003
Haematology	University of Cambridge	PhD	2006

SELECTED PEER-REVIEWED PUBLICATIONS

Papaemmanuil E, Rapado I, …, Greaves M **and Campbell PJ**. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. **Nature Genetics** 2013 (in press).

Alexandrov LB, Nik-Zainal S, …, **Campbell PJ** and Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013, 500(7463), 415-21.

Nik-Zainal S, Van Loo P, … Futreal PA, Stratton MR, **and Campbell PJ**. The life history of 21 breast cancers. **Cell** 2012, 149(5), 994-1007.

Nik-Zainal S, Alexandrov LB, … Futreal PA, **Campbell PJ** and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. **Cell** 2012, 149(5), 979-993.

Papaemmanuil E, Cazzola M, …Futreal PA, Stratton MR, **and Campbell PJ**. Somatic *SF3B1* mutation in myelodysplasia with ring sideroblasts**. N Engl J Med** 2011, 365(15):1384-95.

Stephens PJ, Greenman CD, …, Futreal PA **and Campbell PJ**. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. **Cell** 2011, 144(1), 27-40.

Greenman CD, Pleasance ED, …Futreal PA, Stratton MR, **and Campbell PJ.** Estimation of rearrangement phylogeny for cancer genomes. **Genome Res**. 2012, 22(2), 346-61.

Campbell PJ, Yachida S,… Iacobuzio-Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. **Nature** 2010, 467(7319), 1109-13.

Pleasance ED, Stephens PJ, … Stratton MR, Futreal PA, **and Campbell PJ**. A small cell lung cancer genome with complex signatures of tobacco exposure. **Nature** 2010, 463(7278), 184-90.

Campbell PJ, Stephens PJ, , … Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics** 2008, 40(6), 722-9.

Campbell PJ, Pleasance ED, … Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. **Proc Natl Acad Sci USA** 2008, 105(35), 13081-6.

Campbell PJ, Scott LM, … Harrison CN, Green AR. Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. **Lancet** 2005, 366(9501), 1945-1953.

CURRICULUM VITAE – JOSE MC TUBIO, PhD

Highlights

- Dedicated to the study of the dynamics of Transposable Elements in eukaryotic genomes since 2001
- Author of 6 papers on the field of transposable elements and 7 papers on cancer genomics
- I have developed my research within the framework of the ICGC since 2010 (CLL and Bone Project)
- Currently, I have a Marie Curie Grant for the analysis of transposons and viruses of 1,000 cancers
- Availability: 100% of my time to the project on this proposal

Professional background and Education

Since 2013	Marie Curie Postdoctoral Fellow, Sanger Institute (UK). Advisor: Peter J Campbell. Project: Analysis of the transposable element and viral complement of 1,000 cancer genomes
2010-2012	Postdoc at Center for Genomic Regulation (Barcelona). Advisor: Xavier Estivill. Project: Analysis of the Structural Variation of the CLL (CLL Genome Project, ICGC)
2009	Ph.D. in Biology, University of Santiago de Compostela (Spain). Thesis theme: Evolutionary dynamics of transposable elements in insect genomes
2006-2009	Clinical Study Coordinator, Department of Haematology, University Hospital of Santiago de Compostela, Spain
2003-2005	Teaching Assistant in Genetics, University of Santiago, Spain

Selected Publications

Papaemmanuil et al. (Includes **Tubio J**). RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet* (2013), in press

Tarpey et al. (Includes **Tubio JM**). Frequent mutation of the major cartilage collagen COL2A1 in chondrosarcoma. *Nature Genetics* (2013), 45(8): 923-926

Quesada et al. (Includes **Tubio JMC**). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 in chronic lymphocytic leukemia. *Nature Genetics* (2011), 44(1): 47-52

Puente et al. (Includes **Tubío JM**). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* (2011), 475(7354): 101-105

Tubio JM, Estivill X. When catastrophe strikes a cell. *Nature* (2011), 470(7335): 476-477

Tubio JM et al. Evolutionary dynamics of the Ty3/gypsy LTR retrotransposons in the genome of Anopheles gambiae. *PLoS ONE* (2011), 6(1): e16328

Arensburger et al. (Includes **Tubio JM**). Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science* (2010), 330(6000): 86-88

Nene et al. (Includes **Tubio JM**). Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. *Science* (2007), 316(5832): 1718-1723

CURRICULUM VITAE – M.Sc. Jelena Tica

Jelena Tica
PhD student
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, 69 117 Heidelberg, Germany

Academic Education and Qualification

2011 – present	PhD student in Molecular Biology with specialization in Computational Biology, Joint PhD degree between EMBL and Ruprecht-Karls-Universität Heidelberg / University of Heidelberg, mentor: Dr. Jan O. Korbel
2008 – 2011	M.Sc. Molecular Biology at University of Zagreb, Faculty of Science, mentor: Prof. Dr. Kristian Vlahovicek
2005 – 2008	B.Sc. Molecular Biology at University of Zagreb, Faculty of Science, mentor: Prof. Dr. Srecko Gajovic

Other Professional Experience

2012 – 2013	Organization of the 3rd Heidelberg Forum for Young Life Scientists in Heidelberg (http://www.life-science-forum-hd.de/)
2011 – 2012	Organization of the 14th International EMBL PhD symposium: "Networks in Life Sciences" in Heidelberg (http://www.phdsymposium.embl.org/symp2012/)
2009	Laboratory work practice at Ruder Boskovic Institute (Zagreb, Croatia) in Molecular Microbiology Laboratory, supervised by Dr. Ksenija Zahradka
2007 – 2008	Laboratory work practice at School of Medicine, Croatian Institute for Brain Research in Neurogenetics and Developmental Biology Laboratory (Zagreb, Croatia), supervised by Prof. Dr. Srecko Gajovic

Publications

Schröder MS, Harnett D, Minke BA, Nair PS; **Committee Member Consortium** (2013) Organizing a PhD symposium--an inside view. Setting up a scientific meeting is challenging - a PhD symposium offers its own unique opportunities and pitfalls. *EMBO reports* 14:856 - 860

Gokcumen O, Tischler V, **Tica J**, ..., Korbel JO. (2013) Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc Natl Acad Sci U S A* 110 (39):15764-15769

Rausch T, Jones DT, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger N, Remke M, Shih D, Northcott PA, Pfaff E, **Tica J**, ..., Korbel JO. (2012) Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* 148:59-71

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Inference of timing, signatures, mechanisms and consequences of structural variation by digital karyotyping

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

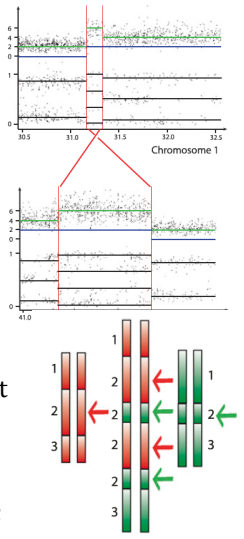
Dr Jan Korbel, EMBL: Affiliation to ICGC Ped-brain; ICGC-MMML; ICGC-early onset prostate cancer
Dr Peter Campbell, Sanger Institute: Affiliation to ICGC breast, bone & chronic myeloid cancers

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yilong Li, Sanger Institute (yl3@sanger.ac.uk) Joachim Weischenfeldt, EMBL (weischen@embl.de)	

Background and preliminary data

Objective: Using naïve structural variant calling approaches from short-read shotgun sequencing data can lead to misclassification of the temporal order and composition of the variant locus. We will classify and describe the mutational signatures and consequences of somatically acquired structural variants in 2000 cancer genomes using an exhaustive, pre-computed library of chromosomal structures arising from sequentially applied simple rearrangements. By deducing the temporal ordering of somatic structural rearrangements on each chromosome our approach can facilitate the categorization and further characterization of both simple and complex structural rearrangements (*e.g.*, breakage-fusion-bridge cycles, chromothripsis, and chromoplexy; McClintock, *Genetics*, 1941; Stephens et al. *Cell*, 2011; Rausch et al. *Cell* 2012; Baca et al. *Cell* 2013; Korbel & Campbell, *Cell*, 2013) in the WGS pan-cancer data.

Motivational example: To illustrate why a long-range reconstruction of digital karyotype improves accuracy of inferences about underlying structural rearrangements processes, consider the relatively straightforward example of an unbalanced translocation followed by a tandem duplication across the translocation breakpoint (see figure to the right). Here, the actual tandem duplication appears to be an interchromosomal translocation, since the two segments joined originated either side of the unbalanced translocation. An analysis of mutational signatures performed without understanding this would misattribute the second rearrangement as a translocation. Similarly, an analysis of gene consequences might fail to identify that the end-product of the second rearrangement is to increase copy number of genes within the duplication.



Preliminary conceptual analysis: It turns out that for the example shown, there is only one possible temporal ordering of rearrangements and only one possible digital karyotype that is consistent with the exact allele-specific copy numbers and rearrangement joins identified. We have built the theoretical foundation for reconstructing long-range digital karyotypes using a graph theory approach (Greenman et al, Genome Research 2012). This is essentially a deductive approach, hampered by missing data (copy number changes for which the corresponding rearrangement is not called) or inaccuracies in allelic ratios.

Timelines & resources dedicated to project

(1) Development of infrastructure for generating pre-computed library of rearrangements (January 2014)
(2) Pilot on 20 breast, 20 paediatric brain, 20 osteosarcoma and 20 prostate cancer genomes, aiming for reconstruction of major somatic clone for 80% of all chromosomes, in 80% of all good quality samples (April 2014).
(3) Benchmarking of our digital karyotyping approach using validation data for two cancer cell lines (May 2014).
(4) Run on chr21 and chr10 through entire 2,000 cancers (September 2014).
(5) Run on whole genomes from 2,000 cancers (November 2014)

International Cancer Genome Consortium

Research proposal
We will perform digital karyotyping in 2,000 cancer genomes to facilitate structural variant classification, temporal ordering of different rearrangement forms, and inference of mechanism and functional consequences of structural alterations. A detailed work plan is described below: **Phase 1:** We will build an exhaustive, pre-computed library of digital karyotypes that can arise from sequential application of simple rearrangements. The simple rearrangements to be considered will include deletion, tandem duplication, inversion, inverted duplication, breakage-fusion-bridge, balanced chromosomal translocation, unbalanced translocation, and arm or whole chromosome gain or loss. Essentially, starting from the reference genomic configuration, we will generate the directed graph of connections and allele-specific copy number for every possible combination of rearrangements. This will be stored in a data structure that will allow matching of observed genomic graphs to any potential digital karyotype and temporal order of rearrangements. We believe we will be able to build this library for up to 7 sequential rearrangements within a cluster, and possibly deeper. **Phase 2:** We will implement a statistical algorithm to match observed cancer genome data to entries in the precomputed library. Since the precomputed library is a discrete distribution, a maximum likelihood approach will be relatively straightforward to implement. The algorithm will be capable of making inferences in the presence of uncertainty, such as in estimating allele-specific copy number and when there are missing structural variants (copy number changes without matched rearrangement). **Phase 3:** We will test and validate phases 1 and 2 on two cancer cell lines with complicated karyotypes for which we have extensive sequencing data. This includes BAC libraries shotgun sequenced and finished to gold standard completion (Bignell et al, Genome Research 2007); whole genome massively parallel sequencing; spectral karyotyping and high-resolution, genome wide optical mapping using BioNano Genomics Irys technology. We will also undertake large insert mate-pair mapping on these cell lines (4-5kb insert size), and have the potential to test predictions of the phase 1 and 2 algorithms by targeted FISH. **Phase 4:** We will benchmark the algorithm on 20 breast, 20 paediatric brain, 20 osteosarcoma and 20 prostate cancer genomes, aiming for reconstruction of the major clone on at least 80% of all chromosomes in at least 80% of all good quality samples. This will enable testing of the ability of our algorithms to reconstruct digital karyotypes on a variety of real-world data sets in primary cancers. **Phase 5:** The algorithms will then be implemented across all 2,000 whole cancer genomes. **Output:** The following items will be the anticipated product of these approaches: (1) Long-range, phased digital karyotypes for the majority of chromosomes in the majority of samples. (2) Putative temporal order of sequential structural rearrangements (including the identification of interactions between structural rearrangement and point mutation processes). We note that these could be used as the starting point for many additional analyses, e.g. the classification of tumour types based on rearragement signatures, and establishment of combined signatures of point mutations and rearrangements. (3) Corrected classification of type of structural variant, which will facilitate detailed analyses of mutational signatures associated with structural variation (such as microhomology, chromatin distribution etc). (4) Predicted gene consequences associated with each rearrangement, isolated from the other rearrangements affecting that region. This will be useful for studies identifying driver mutations linked to structural variants. (5) Complete description of individual cancers' structural variants in a standardized VCF-based format to be developed for recording digital karyotype data. (6) Identification of clusters of structural variants that cannot be reconciled with sequential application of simple rearrangements. This will include patients with chromothripsis, chromoplexy and potentially new, complex rearrangement processes.
Legacy plans
(1) Development of a VCF-like format for recording long-range karyotype data (2) Partial digital karyotypes for over >80% of 2,000 samples (3) Computational framework for generating pre-computed karyotype library (4) Statistical algorithm for pattern matching (5) Pre-computed library of chromosomal structures associated with sequential structural variants (6) Non-circos visualisation tool

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL- EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci*** **USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

CURRICULUM VITAE

NAME
Dr Peter J Campbell

POSITION TITLE
Head of Cancer Genetics & Genomics,
Wellcome Trust Sanger Institute

EDUCATION/TRAINING

FIELD OF STUDY	INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED
Mathematics and Statistics	University of Otago, New Zealand	BSc Hons (1st Class)	1994
Medicine	University of Otago, New Zealand	MB ChB (Distinction)	1995
Haematology	Royal Australasian College of Physicians	FRACP	2003
Haematology	Royal College of Pathologists of Australasia	FRCPA	2003
Haematology	University of Cambridge	PhD	2006

SELECTED PEER-REVIEWED PUBLICATIONS

Papaemmanuil E, Rapado I, …, Greaves M **and Campbell PJ**. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. **Nature Genetics** 2013 (in press).

Alexandrov LB, Nik-Zainal S, …, **Campbell PJ** and Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013, 500(7463), 415-21.

Nik-Zainal S, Van Loo P, … Futreal PA, Stratton MR, **and Campbell PJ**. The life history of 21 breast cancers. **Cell** 2012, 149(5), 994-1007.

Nik-Zainal S, Alexandrov LB, … Futreal PA, **Campbell PJ** and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. **Cell** 2012, 149(5), 979-993.

Papaemmanuil… E, Cazzola M, Futreal PA, Stratton MR, **and Campbell PJ**. Somatic *SF3B1* mutation in myelodysplasia with ring sideroblasts**. N Engl J Med** 2011, 365(15):1384-95.

Stephens PJ, Greenman CD, …, Futreal PA **and Campbell PJ**. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. **Cell** 2011, 144(1), 27-40.

Greenman CD, Pleasance ED, …Futreal PA, Stratton MR, **and Campbell PJ.** Estimation of rearrangement phylogeny for cancer genomes. **Genome Res**. 2012, 22(2), 346-61.

Campbell PJ, Yachida S,… Iacobuzio -Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. **Nature** 2010, 467(7319), 1109-13.

Pleasance ED, Stephens PJ, … Stratton MR, Futreal PA, **and Campbell PJ**. A small cell lung cancer genome with complex signatures of tobacco exposure. **Nature** 2010, 463(7278), 184-90.

Campbell PJ, Stephens PJ, , … Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics** 2008, 40(6), 722-9.

Campbell PJ, Pleasance ED, … Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. **Proc Natl Acad Sci USA** 2008, 105(35), 13081-6.

Campbell PJ, Scott LM, … Harrison CN, Green AR. Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. **Lancet** 2005, 366(9501), 1945-1953.

Curriculum Vitae

Yilong Li

E-mail: yl3@sanger.ac.uk
Cancer genome project, Wellcome Trust Sanger Institute

Hinxton, CB10 1SA, United Kingdom

Research experience

2010-2011	Research student in cancer genetics and genomics, University of Helsinki, Finland
2010	Internship in genomewide methylome analysis, BGI, Shenzhen, China
2010	Summer research student in high-throughput microtubule microscopy, University of Toronto, Canada
2007 - 2010	Undergraduate research student in cancer genetics and genomics, University of Helsinki, Finland

Education

2011-present	Post-graduate studies in cancer genomics, Wellcome Trust Sanger Institute, United Kingdom
2010-2011	Master's degree in Genetic Bioinformatics, University of Helsinki, Finland
2005-2010	Bachelor's degree in genetics, University of Helsinki, Finland

Publications

Kaasinen E, Aavikko M, Vahteristo P, Patama T, **Li Y**, Saarinen S, Kilpivaara O, Pitkänen E, Knekt P, Laaksonen M, Artama M, Lehtonen R, Aaltonen LA, Pukkala E. Nationwide registry-based analysis of cancer clustering detects strong familial occurrence of Kaposi sarcoma. PLoS One 2013.

Aavikko M, Li SP, Saarinen S, Alhopuro P, Kaasinen E, Morgunova E, **Li Y**, Vesanen K, Smith MJ, Evans DG, Pöyhönen M, Kiuru A, Auvinen A, Aaltonen LA, Taipale J, Vahteristo P. Loss of SUFU function in familial multiple meningioma. Am J Hum Genet. 2012

Mäkinen N, Mehine M, Tolvanen J, Kaasinen E, **Li Y**, Lehtonen HJ, Gentile M, Yan J, Enge M, Taipale M, Aavikko M, Katainen R, Virolainen E, Böhling T, Koski TA, Launonen V, Sjöberg J, Taipale J, Vahteristo P, Aaltonen LA. MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. Science. 2011

Niittymäki I, Tuupanen S, **Li Y**, Järvinen H, Mecklin JP, Tomlinson IP, Houlston RS, Karhu A, Aaltonen LA. Systematic search for enhancer elements and somatic allelic imbalance at seven low-penetrance colorectal cancer predisposition loci. BMC Med Genet. 2011

Curriculum vitae
Joachim Weischenfeldt

PhD Joachim Weischenfeldt

European Molecular Biology Laboratory (EMBL)
Genome Biology Unit
Meyerhofstrasse 1, D-69117 Heidelberg
Phone: +49 (0) 151 43127450
E-mail: joachim.weischenfeldt@embl.de

Scientific vitae

2011-present	Postdoctoral fellow at European Molecular Biology Laboratory (EMBL) in the group of Dr Jan Korbel. Focus: Mechanisms of genomic structural variations in cancer.
2007 – 2010	Postdoctoral fellow at Rigshospitalet/Biotech Research & Innovation Centre (BRIC) in the group of Prof. Bo T. Pors. Focus: Genetics and transcriptional regulation in hematopoiesis and cancer.
2006	PhD in Molecular Biology, Genetics, University of Copenhagen

---

Selected Peer reviewed publications

- **Weischenfeldt J**, Simon R, Feuerbach L, Schlangen K, Weichenhan D, Minner S, Wuttig D, Warnatz HJ, Stehr H, Rausch T et al. 2013. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23(2): 159-170.
- **Weischenfeldt J**, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nature reviews Genetics*** 14(2): 125-138.
- **Weischenfeldt J**, Waage J, Tian G, Zhao J, Damgaard I, Jakobsen JS, Kristiansen K, Krogh A, Wang J, Porse BT. 2012. Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns.
- ***Genome biology*** 13(5): R35.
- Jones DT, Jäger N, Kool M, Zichner T, Hutter B, Sultan M, Cho YJ, Pugh TJ, Hovestadt V, Stütz AM, Rausch T, Warnatz HJ, Ryzhova M, Bender S, Sturm D, Pleier S, Cin H, Pfaff E, Sieber L, Wittmann A, Remke M, Witt H, Hutter S, Tzaridis T, **Weischenfeldt J** et al. 2012. Dissecting the genomic complexity underlying medulloblastoma. ***Nature*** 488(7409): 100-105.
- Rausch T, Jones David TW, Zapatka M, Stütz Adrian M, Zichner T, **Weischenfeldt J**, Jäger N, Remke M, Shih D, Northcott Paul A et al. 2012. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. ***Cell*** 148(1-2): 59-71.
- **Weischenfeldt J**, Damgaard I, Bryder D, Theilgaard-Mönch K, Thoren LA, Nielsen FC, Jacobsen SEW, Nerlov C, Porse BT. 2008. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. ***Genes & Development*** 22(10): 1381-1396.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Interface between germline and somatic genetic variation across multiple tumour types

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Dr Jan Korbel, EMBL Heidelberg, Germany; ICGCs: PedBrainTumor & MMML-Seq & Early Onset Prostate Cancer
Dr Douglas Easton, University of Cambridge, UK; ICGC Breast, ICGC UK Prostate (CR-UKPCN)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Sebastian Waszak, EMBL Heidelberg, Germany Jamie Allen, University of Cambridge, Cambridge, UK	Dr Gabor Marth, Boston College, Boston, MA, USA

Background and preliminary data

Susceptibility to cancer is mediated through a large number of germline genetic loci; these include rare variants in high penetrance genes (e.g. *BRCA1/2*) and more than 200 commoner genetic variants identified through genome-wide association studies (GWAS; www.genome.gov/26525384). The ongoing Oncoarray project, which involves genotyping >400,000 subjects and is due to complete in 2014, is likely to identify many more. Whereas most high-risk regions predispose to more than one cancer type, GWAS loci identified in different cancer entities frequently cluster in extended genomic areas, which suggests that they mediate the tissue-specific regulation of common gene targets (*e.g. TERT*, 8q24; www.nature.com/icogs).

Much less is known about the relationship between germline susceptibility and somatically altered genes. Some somatically altered genes (e.g. *MAP3K1*, *MYC*) are close to and plausible targets for known GWAS hits. In at least one case (*JAK2* and myeloprofilerative disease; Jones *et al.* 2009, *Nat Genet* 41:446-9) somatic mutations occur preferentially on a specific germline haplotype. Furthermore, there is evidence for the association of germline (rare) high-risk genetic variants with specific somatic mutational patterns (see *e.g.* Domingo *et al.* 2004, *J Med Genet*, 41:664-8; as well as our prior work on *TP53* in medulloblastoma in Rausch *et al.* 2012, *Cell*, 148:59-71, and in **Table 1** below), but it is not known whether this extends to commoner germline variants.

Extensive analyses of germline WGS data by the 1000 Genomes Project (1000GP; www.1000genomes.org), have fostered the development of algorithms for generating high quality genotypes and haplotypes (including SNPs, indels and structural variants, SVs) from short reads. These can now be applied to the WGS pan-cancer project. We will bring together PIs with lead roles in major germline genome (*i.e.*, the 1000GP), cancer GWAS (iCOGS and OncoArray), and somatic genome (ICGC) initiatives, to generate high-quality datasets (including SNPs, indels and SVs) of germline variants and haplotypes, and to link these variants to somatic DNA alteration patterns in multiple cancers. We will further use these data to examine more broadly the relationship between susceptibility genotypes and haplotypes, somatic mutations, and gene expression.

Timelines & resources dedicated to project

1) Preliminary analyses of GWAS loci (see proposals 1,2&4 below), after specifically genotyping these [March 2014].
2) Variant discovery with orthogonal 1000GP algorithms (read mapping [RM] & assembly-based [AB]) [completion: July 2014; CPU months per sample=7.5, if 3 RM and 3 AB callers are used (alternatively 2.5, if 1 RM and 1 AB caller are used)]. Algorithm usage will be coordinated/arranged with pan-cancer pipeline providers to avoid duplicate work.
3) Use genotyping algorithms to recall union set of all candidate germline genetic variants across all samples, using 1000GP pipelines [completion: August 2014; compute resources needed=0.5 CPU months per sample].
4) Construct haplotypes using 1000GP pipelines [completion: October 2014; CPU months per sample=1].
5) Verify quality of genotypes and haplotypes, by germline variant calling in two parent-offspring trios sequenced by the 1000GP for which we have independent validation data (including NA12878 PacBio WGS data) [October 2014].
6) Investigate low- to high-risk loci using the completed haplotypes; integrate with somatic genome callsets from other pan-cancer working groups [to be initiated October or November 2014; few compute resources needed for this step].
7) Submission of completed VCF files to imputation server (see legacy) [November 2014].
8) Presentation at ICGC meeting, completion of lookup tool (see legacy), manuscript submission [March 2015].
JK will be able to spend 60% of his time for the WGS pan-cancer initiative, and 25% on this project. DE will spend 20% of his time on this project. GM: 5%; SW: 100%; JA: 40%; Erik Garrison (GM lab): 20%. The look-up tool will be developed by a technical programmer in the Korbel group.

Research proposal

While significant advances have already been made in unraveling somatic DNA alteration patterns in cancer genomes, the interface of germline and somatic variation has so far been under-explored – despite prior findings pointing to biologically relevant links between germline and somatic variants (see *e.g.* references cited under "Background and preliminary data" and additional references provided below).

Significant resources will be devoted to the challenging task of generating high-quality germline haplotypes. We aim to approach the callset quality of the 1000GP final phase. Since some of our analyses can be performed independently of high-quality haplotypes (*i.e.* by genotyping known GWAS loci), we will be able to initiate integrative analyses of germline and somatic variation before completion of the haplotype set.

1. Linking known GWAS loci to nearby genes. We will define a set of cancer associated GWAS loci, augmented with additional loci identified through the Oncoarray project. For certain analyses, we will utilise where available loci with strong but not genome-wide significant evidence. Since many GWAS loci appear to confer susceptibility through regulation of neighbouring genes at some distance, we will define genes within 1Mb of known loci, augmented with genes that are the suspected targets of GWAS hits.

We will examine whether the frequency of somatic driver mutations in the target gene correlates with germline genotype or haplotype (*e.g.*, Jones *et al.* 2009, *Nat Genet* 41:446-9; Tuupanen *et al.* 2009, *Nat Genet* 41:885-90). To improve statistical power we will augment the dataset with available ICGC/TCGA exome sequencing data for which SNP-array genotyping data are also available. For genomic regions linked to several cancer types, we will be able to improve the power of this analysis by conducting a joint analysis across cancer entities (even if the associated lead SNP differs). We will conduct global analyses across all GWAS loci and suspected target genes, to examine the overall evidence for correlation between susceptibility alleles and mutations on the mutant or wild-type haplotype.

2. Linking known GWAS loci to gene expression data. Our germline haplotypes will provide high quality data for evaluating associations between germline genotype and tissue-specific expression (where possible allele-specific expression). To this end we will collaborate with working groups focussing on expression (*e.g.*, involving the Stegle, Brazma, or Huber groups). These analyses could be conducted globally, but our initial focus will be on *cis*-eQTL analyses at GWAS loci. For specific loci it may also be possible to conduct similar analyses for methylation data.

Table 1. The occurrence of somatic driver alterations in Sonic hedgehog-driven medulloblastoma differs between hereditary *TP53* mutation-linked *vs.* sporadic (typically *TP53*-wildtype) tumors (p=1 x 10^{-11}; Fisher's exact test).

	Shh-pathway mutations (in *PTCH1*, *SMO*, or *SUFU*)	*GLI2* amplification	*MYCN* amplification
hereditary	0/9 (0%)	5/9 (56%)	7/9 (78%)
sporadic	96/117 (82%)	2/117 (2%)	5/117 (4%)

3. Linking high-risk alleles to somatic events. We will further catalogue germline variants in known high-risk and moderate-risk genes (e.g. *BRCA1/2*, *TP53*, *ATM*, mismatch repair [MMR] genes; ~50 carriers expected in the pan-cancer WGS dataset). We will assess whether germline mutation carriers are reflected in specific somatic mutation signatures (*e.g.* mismatch repair mutation signatures and *BRCA1/2*-associated deletion signatures; see Nik-Zainal *et al.*, *Cell*, 2012, 149:979-93; Alexandrov *et al.*, 2013, *Nature*, 500:415-21) or differences in specific driver genes being mutated (see our prior work in Rausch *et al.* 2012, *Cell*, 148:59-71, and our preliminary data in **Table 1**).

Potentially, the analyses could be extended to examine germline genetic variants of uncertain significance. We will also catalogue the somatic mutation status (including LOH) of the wild-type allele.

4. Linking GWAS loci to mutational signatures. We will consider two types of analyses. For certain loci, we will examine evidence that tumours in susceptible individuals are enriched for specific types of biologically relevant somatic DNA alterations. Examples common across multiple tumour types are the *TERT* region, where tumours might be enriched for structural rearrangements initiating at the telomeres, and GWAS loci linked to DNA repair (e.g. *RAD51B*). We will also examine whether there are differences in somatic mutational spectra (as in 3) among individuals at the extremes of the risk distribution on the basis of their common germline SNP profile.

Legacy plans

1. High-quality germline variants and haplotypes for 2,000 cancer patients linked to 2,000 somatic genomes. These data will be shared with other pan-cancer working groups, and be available for addressing many additional questions (*e.g.*, investigating potential susceptibility loci below genome-wide significance based on somatic mutation patterns.).

2. Our data will further provide a starting point for identifying novel susceptibility loci, particularly rarer variants not genotyped in GWAS that may be enriched in cancer patients. Replication of such loci may be through replication genotyping or through imputation in larger GWAS datasets. Specifically we will make our germline haplotype data available for imputation through the imputation server being developed by Goncalo Abecasis, Jonathan Marchini and colleagues (see enclosed email communication with these PIs), which altogether will provide a panel of >30,000 genomes (contingent on data access regulations satisfying ICGC/TCGA criteria).

3. A look-up tool for correlating germline genotypes (both common and rare variants) with tumour expression levels, methylation status and somatic mutation status of nearby genes (this could be made available to ICGC/TCGA participants, or publicly, contingent on data access regulations).

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

Since 2013	Fellow of the European Academy of Cancer Sciences.
2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci* USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

Curriculum Vitae - Professor Douglas Frederick Easton

Professor of Genetic Epidemiology, Centre for Cancer Genetic Epidemiology
Department of Public Health and Primary Care and Department of Oncology, University of Cambridge, UK

Positions/Education

2011-	Director, Centre for Cancer Genetic Epidemiology, University of Cambridge
2003-	Professor of Genetic Epidemiology, University of Cambridge
2001-2011	Cancer Research UK Principal Research Fellow
2008-	Hon Co-Director, Strangeways Research Laboratory
1999 - 2003	Reader, Department of Public Health and Primary Care, University of Cambridge
1995 -1999	University Lecturer, Department of Community Medicine, University of Cambridge
1994	Visiting Professor, Dept of Medical Informatics, University of Utah
1982-1995	Research Fellow Staff, Scientist and Team Leader, Institute of Cancer Research
1980-1982	MRC Unit on the Development and Integration of Behaviour, University of Cambridge
1992	PhD, Genetic Epidemiology, University of London
1982	MA, Mathematics, University of Cambridge
1980	Diploma in Mathematical Statisitcs, University of Cambridge
1979	BA, Mathematics Tripos Class 1, University of Cambridge

Leadership of International Consortia

Co-ordinator of the Breast Cancer Association Consortium (BCAC)
Analysis co-ordinator for Prostate Cancer Associated Alterations in the Genome (PRACTICAL).
Scientific chair of Collaborative Oncological Genetics Study (COGS)
Chair of Oncoarray Steering Committee

Selected other experience

Cancer Research UK Science Committee 2012-
Deciphering Developmental Disorders Scientific Advisory Board 2011-
Genome Canada Science and Industry Advisory Committee 2010-2012
Research Assessment Exercise Cancer Sub-Panel 2008.
National Institute for Clinical Excellence (NICE) guideline development group on familial breast cancer.
Breast Cancer Campaign Scientific Executive Board 2004-2007.
MRC Molecular and Cellular Medicine Board 1999-2003.

Selected Publications

Michailidou K, …, **Easton DF**. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013; 45:353-361.

Turnbull C,…, **Easton DF**. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010; 42:504-507.

Eeles RA, …, **Easton DF**. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet 2009; 41:1116-1121..

Eeles RA, …, **Easton DF**. Identification of multiple novel prostate cancer susceptibility loci by a genome-wide association study. Nature Genet 2008; 40:316-321.

Antoniou AC, …, **Easton DF**. Common breast cancer predisposition alleles modify breast cancer risk in *BRCA1* and *BRCA2* mutation carriers. Am J Hum Genet 2008; 82:937-948.

Easton DF, …, Goldgar DE. A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the *BRCA1* and *BRCA2* Breast Cancer Predisposition Genes. Am J Hum Genet 2007; 81:873-883.

Easton DF, …, Ponder BAJ. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007; 447:1087-1093.

Antoniou AC…, **Easton DF**. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer 2004; 91:1580-1590.

Antoniou A, …, **Easton DF**. Average risks of breast and ovarian cancer associated with mutations in BRCA1 or BRCA2 detected in case series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 2003; 72:1117-1130.

Gayther S, …, **Easton DF** (1997) Variation of risks of breast and ovarian cancer associated with different germline mutations of the BRCA2 gene. Nat Genet 15: 103-105

Easton DF, …, Crockford GP. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. Am J Hum Genet 1993; 52:678-701.

CURRICULUM VITAE – Sebastian M. Waszak

Predoctoral Research Fellow
Institute of Bioengineering, School of Life Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
Email: sebastian.waszak@epfl.ch

Academic Education & Qualification

2010-2014	Ph.D., Bioinformatics, École Polytechnique Fédérale de Lausanne, Switzerland. Advisor: Prof. Bart Deplancke. Dean's prize for outstanding young researchers. Ph.D. thesis will be submitted by the end of January 2014.
2005-2010	Dipl.-Ing. (FH), Bioinformatics, Hochschule Weihenstephan-Triesdorf, Germany.
2009	Diplomarbeit (approx. M.Sc. thesis), Weizmann Institute of Science, Israel. Advisor: Prof. Doron Lancet. Feinberg Graduate School Fellowship.
2007-2008	Research Assistant, Ludwigs-Maximilians-Universität München, Germany. Advisors: Dr. Helmut Blum, Prof. Jens Michaelis.

Publications (*joint first authorships)

Waszak SM and Deplancke B (2013). Rounding up natural gene expression variation during development. ***Developmental Cell*** in press doi:10.1016/j.devcel.2013.12.007

Waszak SM, *et al* (2013). Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. ***Bioinformatics*** in press doi:10.1093/bioinformatics/btt667

Kilpinen H*, **Waszak SM***, Gschwind AR*, *et al* (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. ***Science*** 342:744-747.

Gubelmann C, **Waszak SM**, *et al* (2013). A yeast one-hybrid and microfluidics-based pipeline to map mammalian gene regulatory networks. ***Molecular Systems Biology*** 9:682.

Massouras A*, **Waszak SM***, *et al* (2012). Genomic variation and its impact on gene expression in Drosophila melanogaster. ***PLoS Genetics*** 8:e1003055.

Raghav S*, **Waszak SM***, *et al* (2012). Integrative genomics identifies the corepressor SMRT as a gatekeeper of adipogenesis through the transcription factors C/EBPβ and KAISO. ***Molecular Cell*** 46:335-350.

Olender T, **Waszak SM**, *et al* (2012). Personal receptor repertoires: olfaction as a model. ***BMC Genomics*** 13:414.

Schlattl A, Anders S, **Waszak SM**, *et al* (2011). Relating CNVs to transcriptome data at fine-resolution: assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Research*** 21:2004-2013.

Waszak SM, *et al* (2010). Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. ***PLoS Computational Biology*** 6:e1000988.

Kasowski M*, Grubert F*, Heffelfinger C, Hariharan M, Asabere A, **Waszak SM**, *et al* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-235.

Jamie Allen

Summary: I am a Bioinformatician with a wide and varied background in Biology and working in other industries. My current work involves the provision of bioinformatics support to genetic epidemiologists investigating breast, ovarian and prostate cancer. My primary focus at present is establishing and monitoring next generation sequencing pipelines for whole genomes, exomes and targeted resquencing projects. I also help out with various ad-hoc projects, often involving my experience with genetic regulatory systems.

Education

2005-2009 University of Newcastle PhD Systems Biology/Bioinformatics

I worked on in-silico methods of predicting gene co-expression in various organisms. The approach was purely sequence based, essentially the upstream sequences of genes were analysed for the presence of CIS Regulatory Element (CRE) motifs. Each gene would then have a "fingerprint" profile depending on the distribution of CRE motifs. These profiles were numerically transformed and then clustered using a custom written Self Organising Map (SOM) neural network. Those genes clustering closely were predicted to co-express.

2005 Glasgow University MRes Bioinformatics

Two main projects, the first to create Java DNA translation tool that could be accessed online via a front end HTML interface, a middle layer of CGI scripting connected to the back end Java program. The second project was the creation of a method of predicting contact residues in heterodimeric proteins, using Correlate Mutation Analysis (CMA) and a site class model of evolution.

1997-2001 University of Edinburgh BSc(*Hons.*) Neuroscience (2.1)

I undertook a wide range of courses during the 4 year degree. My final year project dissertation was on the neurobiological basis of consciousness.

Work experience

2011-now Bioinformatician, Cancer Genetic Epidemiology, Cambridge University.

2009-2011 Client Service Executive, BNYM, Edinburgh.

2003 Pharmacokineticist, Inveresk Research, Edinburgh (Tranent)

2002 New Business Administrator, Scottish Widows, Edinburgh.

Additional work (on a casual or short-term basis)

IT Assistant, Student Demonstrator, Care Assistant, Retail Assistant.

Gabor Tamas Marth

Technical University of Budapest, Budapest, Hungary	B.S.-M.S.	1983 - 1987	Electrical Engineering
Washington University, St. Louis MO	D.Sc.	1988 - 1994	Systems Science and Math
Washington University, St. Louis MO	Post-doc	1994-2000	Genome Informatics
National Center for Biotechnology Information, NLM, NIH, DHHS	Staff scientist	2000-2003	Genome variation research
Department of Biology, Boston College, Chestnut Hill MA	Assistant Professor	2003-2009	Genome variation research
Department of Biology, Boston College, Chestnut Hill MA	Associated Professor	2009 -present	Genome variation research

Personal statement

My research focuses on the development of DNA sequence analysis software. Over the past 15 years I have developed software to aid genome sequence completion (finishing), for single-nucleotide polymorphism discovery, for population genetic analysis of genomic variation data. I have participated in large consortia variant discovery projects. Most recently, as faculty in the Boston College Biology department, my group and I have developed software packages for base calling, read mapping, variant discovery, and data visualization in high-throughput, next-generation sequencing data. My current research is aimed at developing complete, automated pipelines for sequence processing, variant detection, and variant interpretation; adapt and extend our tools for cancer sequence analysis, and at developing informatics technologies to support population, medical, and personal genome sequencing of very large numbers of samples.

Other Experience and Professional Memberships relevant to this application

2010 – present Editorial Board Member, *Genome Research*
2010 – Steering Committee Member and Member of the Writing Group of the 1000 Genomes Project
2012 – present Member, External Evaluation Committee, T2D-GENES Project, NIDDK/NIH

Selected Peer-Reviewed Publications (in chronological order)

1. Huang W, **Marth G**. EagleView: A genome assembly viewer for next-generation sequencing technologies. ***Genome Research***. 2008;18:1538-43. Epub 2008 Jun 11
2. Quinlan AR, Stewart DA, Strömberg MP, **Marth GT**. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. ***Nature Methods***. 2008;5:179-81.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, **Marth G**, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. ***Bioinformatics***. 2009 Aug 15;25(16):2078-9. Epub 2009 Jun 8.
4. A comprehensive map of mobile element insertion polymorphisms in humans. Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, **Marth GT**; 1000 Genomes Project. ***PLoS Genet***. 2011 Aug;7(8):e1002236. Epub 2011 Aug 18.
5. The functional spectrum of low-frequency coding variation. **Marth GT**, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X, Chen Y, Challis D, Clarke L, Ball EV, Cibulskis K, Cooper DN, Fulton B, Hartl C, Koboldt D, Muzny D, Smith R, Sougnez C, Stewart C, Ward A, Yu J, Xue Y, Altshuler D, Bustamante CD, Clark AG, Daly M, Depristo M, Flicek P, Gabriel S, Mardis E, Palotie A, Gibbs R; the 1000 Genomes Project. ***Genome Biol***. 2011 Sep 14;12(9):R84.
6. A DOC2 protein identified by mutational profiling is essential for apicomplexan parasite exocytosis. Farrell A, Thirugnanam S, Lorestani A, Dvorin JD, Eidell KP, Ferguson DJ, Anderson-White BR, Duraisingh MT, **Marth GT**, Gubbels MJ. ***Science***. 2012 Jan 13;335(6065):218-21.
7. An integrated map of genetic variation from 1,092 human genomes. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, **Marth GT**, McVean GA. ***Nature***. 2012 Nov 1;491(7422):56-65.
8. Scribl: an HTML5 Canvas-based graphics library for visualizing genomic data over the web. Miller CA, Anthony J, Meyer MM, **Marth G**. ***Bioinformatics***. 2013 Feb 1;29(3):381-3. doi: 10.1093/bioinformatics/bts677. Epub 2012 Nov 19.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The ICGC PAN-CANCER Study on Genomic Commonalities in Clinically Defined Subgroups across Tumor Entities

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Peter Lichter
Head of Division of Molecular Genetics (B060)
German Cancer Research Center (DKFZ)
Im Neuenheimer Feld 280
69120 Heidelberg
Germany
Ph: +49 6221 42 4619
Fax: +49 6221 42 4639
Email: Peter.Lichter@dkfz-heidelberg.de

Andrew V. Biankin
Regius Professor of Surgery
Director, Wolfson Wohl Cancer Research Centre,
University of Glasgow
Garscube Estate, Switchback Road, Bearsden, Glasgow Scotland G61 1BD
United Kingdom
Ph: +44 141 330 5670 (direct)
Email: Andrew.Biankin@glasgow.ac.uk

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
This project relies on the interaction of several experts in the field of clinical genomics that do not classify as junior investigators (see below timelines and resources).	

Background and preliminary data

The long-term goal of the ICGC is to improve outcomes for patients with cancer. While defining the genomic events that characterize cancer increases our understanding of the underlying molecular pathology, defining the clinical characteristics associated with specific genomic subtypes may implicate functional consequences that will prioritize ongoing research. In addition, identifying genomic subclasses that co-segregate with clinical and pathological characteristics may provide insights that impact clinical decision-making. As the number of available cancer genomes increases, our power to detect associations between genomic subtypes and clinico-pathological features becomes feasible. Effectively analyzing these data to glean clinically meaningful results is a current major challenge that requires particular attention to maximize the benefits of sequencing cancer genomes. As a consequence, the **overall aim** of this proposal is to correlate clinical and pathological features with genomic subtypes (e.g.: mutated or amplified genes, aberrant pathways, subtypes based on structural variation and mutational signatures) identified in the proposed pan-cancer analysis.

Early efforts in specific cancers have identified relevant genomic subtypes that are characterized by distinct clinical features. Recent analysis of pancreatic cancer genomes by members of the applicant team defined that HER2 amplified pancreatic cancer was associated with a distinct pattern of metastatic spread. Initially, metastases occurred to lung and brain but not to liver. This knowledge has significant implications for initial diagnosis and prediction of recurrent disease. Importantly, it implies that the site of metastases is related to inherent tumor biology rather than the venous drainage from the pancreas to the liver. HER2 amplified breast cancer, and more recently gastric cancer, is also characterized by brain metastases. Cross cancer comparisons through the planned pan-cancer project will validate these findings and better define the characteristics of this genotype across different cancer types, and potentially uncover other novel insights.

Timelines & resources dedicated to project

The following experts in their field participate in this project:

- Stefan Pfister, Head of Division of Pediatric Neurooncology, DKFZ, Heidelberg, Germany
- Benedikt Brors, Theoretical Bioinformatics, DKFZ, Heidelberg, Germany
- Reiner Siebert, Director of the Institute of Human Genetics, University of Kiel, Germany
- Sean Grimmond, Chair of Medical Genomics, University of Glasgow, UK
- Rebecca Fitzgerald, Department of Oncology, University of Cambridge, UK
- Tom Hudson, President and Scientific Director of the Ontario Institute for Cancer Research (OICR), Canada
- Ros Eeles, The Institute Of Cancer Research, London

Each of the PIs will dedicate resources to the project and Rebecca Fitzgerald and Ros Eeles will seek additional funding from Cancer Research UK if this proposal is selected. The workload will be divided between PIs and in particular we will share out the liaising with groups participating in the pan-cancer analysis to ensure provision of high quality clinic-pathological data where-ever possible and conversion to a unified format. The specific aims for analysis will be divided up and close liaison will be achieved through regular conference calls and face to face meetings as required.

Milestone	**2014**				**2015**	
	I	II	III	IV	I	II
M1: Assembly of clinical datasets across cancer types	X	X				
M2: Quality control of pathological diagnosis	X	X	X			
M3: Harmonize categories within and across different tumor types		X	X	X		
M4: Correlation of harmonized clinical and pathological features	X	X	X	X	X	X
M5: Further development of tools for dissemination and data write up		X	X	X	X	X

Research proposal

To decipher possible common tumorigenic mechanisms of clinical relevance, we will study clinical parameters across cancer types that are within the ICGC pan-cancer consortium. For this purpose we will correlate clinical parameters with mutation calls (SNVs, CNVs, indels, rearrangements, SNV mutation frequencies) affected biochemical pathways and mutation signatures (according to Alexandrov et al. Nature 2013) and – where possible – also with transcriptome and methylome data.
We aim to:

1. **Curate available clinico-pathological data** to identify deficiencies and enrich these data through direct interaction with individual ICGC project teams and compare the overall cohort and specific cancer types with the broad nature of the clinical disease seen in the community to define the acquisition bias of sequencing studies.
2. **Quality control of pathological diagnosis** in selected cases by cancer-specific expert pathologists on the basis of the images stored in the ICGC database.
3. **Harmonize categories within and across different tumor types** that are clinically relevant and pathologically meaningful.
4. **Correlate harmonized clinical and pathological features and outcome with genomic subtypes across the PanCancer tumor panel**. We will correlate clinical parameters with mutation calls (SNVs, CNVs, indels, rearrangements, SNV mutation frequencies) affected biochemical pathways and mutation signatures (according to Alexandrov et al. Nature 2013) and – where possible – also with transcriptome and methylome data. Previously described associations between genotype and phenotype will also be considered. Focus will be given on parameters such as:
   a) **Age**, as most tumor entities occur predominantly at certain time windows of life span, and e.g. for DNA methylation an age-dependence is already known.
   b) **Gender**, as e.g. for many tumor entities strong gender predominance has been well established and novel data suggest influence of X chromosome hypermutations.
   c) **Tumor/metastasis localization** in the body, as the microenvironment greatly influences tumorigenesis.
   d) **TNM status**, as lymphoid dissemination and distant metastasis are influenced by common mechanisms (such as EMT/MET in carcinomas).
   e) **Tumor grade**, as the status of differentiation/de-differentiation may represent the progression of tumors (this will certainly require an adaption of current grading systems to match with a few common categories as a major task).
   f) **Environmental factors**, such as tobacco, in case these data will be available also for non-lung tumors.
   g) **Survival** such as OS and TFS, this will likely also be calculated for tumor-type adapted time windows.
   h) **Occurrence of secondary malignancies**/family history of cancer, in case these data will be available.
   i) **Therapy type**, in case these data will be available; this might include radiotherapy (yes/no/in combination), chemotherapy (yes/no/in combination), type of chemotherapy (alkylating, DNA intercalating, anti-metabolites, anti-microtubule agents, topoisomerase inhibitors, cytotoxic antibiotics) or molecular therapy (kinase inhibitors, epigenetic drugs).
   j) **Toxicity profiles**, if data will be available.

The study will be coordinated in close interaction between the two co-applicants with i) Andrew Biankin specifically focusing on the aims 1-3, utilizing an interacting group of pathologists in the UK and beyond (see timeline and resources) and ii) Peter Lichter focusing on aim 4 within an established strong network of clinical geneticists and bioinformaticians (see timeline and resources).

Legacy plans

Developed software tools will be embedded in a virtual machine enabling their use in the context of the ICGC Pan-Cancer cloud computing infrastructure. The tools and comprehensive documentation will be made publicly available upon publication of the results. We will also provide a full set of highly curated clinical parameters that will be of use for future studies aiming at analyzing genotype-phenotype associations on the pan-cancer data set. Relevant subgroupings as well as classifiers of prognosis or therapy response will be made available to clinical researchers. The prediction tools that we provide will in addition also use mutational signatures and other molecular alterations.

Peter Lichter

*(*1957)*

E-Mail: peter.lichter@dkfz-heidelberg.de

ResearcherID: I-3483-2013

Current Position

Head of Division Molecular Genetics (B060) since 1992
Full Professor at the Faculty of Medicine, University of Heidelberg since 2000

Research Topics

- Pathomechanisms of tumor development
- Tumor markers
- Molecular profiling of tumor cells
- Genome organization and gene function

Degree

Graduation (Biology) and PhD degree University of Heidelberg 1983-1986
Postdoc in the laboratory of Prof. D.C. Ward, Dept. of Genetics, School of Medicine, Yale University, New Haven, USA 1986-1990
Habilitation and venia legendi in Molecular Human Genetics, Faculty of Medicine, University of Heidelberg 1995

Previous appointments

Head of project group "Organization of complex genomes", DKFZ 1990-1992
Interim Director of the Management Board of the DKFZ 2003

Selected Publications

- Döhner H,..., **Lichter** P Genomic aberrations predict survival of patients with B-cell chronic lymphocytic leukemia. **New Engl. J. Med**. 343, 1910-1916 (2000)
- Rausch T, Jones DTW, Zapatka M, Stütz AM,..., **Lichter** P*, Pfister SM*, Korbel JO* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. **Cell** 148, 59-71 (2012) * shared senior authorship
- Jones DTW, Jäger N,..., **Lichter** P. Dissecting the genomic complexity underlying medulloblastoma. **Nature** (2012) (DOI: 10.1038/nature11284)
- Jones DT, Hutter B, Jäger N,..., Eils R, **Lichter** P, Pfister SM; International Cancer Genome Consortium PedBrain Tumor Project. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. **Nat Genet**.;45(8):927-32 (2013)
- Remke M, Hielscher T, ..., Taylor MD, **Lichter** P, Pfister SM. FSTL5 is a marker of poor prognosis in non-WNT/non-SHH medulloblastoma. **J Clin Oncol**.;29(29):3852-61 (2011)

Activities in the scientific community

- Member of Hinterzartener Kreis (DFG) 2004-2010
- Sci. Progr. Committee of the Eur. Soc. of Human Genetics 2004-2008
- Member of Wissenschaftsrat 2005-2011
- Member of Leopoldina since 2006
- Member of EMBO since 2008

Honors and awards

- "Karl-Freudenberg" Award of the Academy of Sciences, Heidelberg (1991)
- Award of the (German) Society for Human Genetics (1992)
- "Walther und Christine Richtzenhain" Award (1993)
- Deutscher Krebspreis (German Cancer Award, 2002)
- Award of Deutsche Krebshilfe (2003)
- Award of the European Society of Human Genetics (2012)

Andrew V. Biankin

CURRENT POSITIONS
Regius Professor of Surgery, University of Glasgow.
Director, Wolfson Wohl Cancer Research Centre, University of Glasgow.
Head, Pancreatic Cancer research, Garvan Institute of Medical Research,
Professor, Conjoint Appointee University of New South Wales

QUALIFICATIONS

1988	B. Med. Sc.	University of New South Wales
1992	M.B.,B.S. (HONS)	University of New South Wales
1999	F.R.A.C.S.	Royal Australasian College of Surgeons
2003	Ph. D.	University of New South Wales
2011	F.F.S (RCPA)	Royal College of Pathologists of Australasia
2012	F.R.C.S. (Glasg.)	Royal College of Physicians and Surgeons of Glasgow
2013	F.R.C.S. (Edin.)	Royal College of Surgeons of Edinburgh

HONOURS and AWARDS (selected)

2012	Cancer Institute NSW Wildfire Award
2010	Landon Foundation-AACR INNOVATOR Award
2008	Hirshberg Award for Pancreatic Cancer, American Pancreatic Association
2007	Cancer Institute NSW Premier's Award for Outstanding Cancer Research Fellow
2005	Cure Cancer Australia Young Researcher of the Year (Open Division)
2004	Excellence in Translational Research Award, Johns Hopkins University
2003	Garvan Institute of Medical Research, Thesis Prize

RESEARCH INTERESTS
Biankin's primary scientific focus is on the molecular pathology of pancreatic cancer, the development of early detection and novel therapeutic strategies based on molecular phenotyping and the delineation and implementation of biomarkers that facilitate clinical decision-making. He contributes to the International Cancer Genome Consortium through extensively characterising the genomic, transcriptomic and epigenomic aberrations in pancreatic cancer, and is extending this knowledge to a personalized model of cancer care, where molecular characteristics guide treatment decisions.

PREVIOUS APPOINTMENTS

2005 – 2014	Head, Pancreatic Cancer Research, The Kinghorn Cancer Centre, Cancer Research Program, Garvan Institute of Medical Research
2005 – 2013	Consultant HPB and Upper GI Surgeon, Sydney South West Area Health.
2005 – 2014	Chairman, Australian Pancreatic Cancer Network
2009 – 2014	Clinical Lead, Australian Pancreatic Cancer Genome Initiative (ICGC).
2006 – 2012	Chairman Bankstown Hospital Multidisciplinary Team (GI Oncology).

SELECTED RECENT PUBLICATIONS (from a total of 100)

1. Chang DK, Johns A, … Kench JG and **Biankin** AV. (2009) Margin clearance and outcome in resected pancreatic cancer. **J Clin Oncol** 27: 2855-2862
2. **Biankin** AV, Waddell N, … Pearson JV, McPherson JD, Gibbs RA, Grimmond SM. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. **Nature**. 2012;491:399-405
3. Chang DK, Colvin EK, Johns A, … Kench JG and **Biankin** AV. Histomolecular Phenotypes and Outcome in Adenocarcinoma of the Ampulla of Vater. **J Clin Oncol**. 2013 31:1348-56
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S, Behjati S, **Biankin** AV, … Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013 500:415-21.
5. Chou A, Waddell N, Cowley MJ, Gill AJ, Chang DK, Patch AM, Nones K, Wu J, Pinese M, Johns AL, Miller DK, Kassahn KS, … Grimmond SM, **Biankin** AV. Clinical and molecular characterization of HER2 amplified pancreatic cancer. **Genome Med**. 2013 Aug 31;5:78.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identification of non-coding cancer drivers in pan-cancer data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)

Jan Korbel, EMBL Heidelberg, Germany; ICGC PedBrainTumor, MMML-Seq, Early Onset Prostate Cancer
Mark Gerstein, Yale University, USA; TCGA Prostate

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Ekta Khurana, Yale University Vasilisa Rudneva, EMBL Heidelberg	Kevin White, Institute for Genomics and Systems Biology, University of Chicago, USA

Background and preliminary data

Introduction. Although it is known that non-coding regions are under purifying selection and variants in them have been linked to many diseases (*e.g.*, through GWAS; www.genome.gov/26525384), their role in cancer is generally less well understood. Some recent studies have highlighted the role of non-coding regulatory mutations in human melanoma by identification of potential driver mutations in the *TERT* gene promoter (Horn et al., *Science*, 2013; Huang et al., *Science*, 2013). Indeed, a recent study indicates that *TERT* promoter mutations could provide a biomarker for early detection and classification of tumors (Killela et al., *PNAS*, 2013). However, most previous cancer studies focused on variants in protein-coding genes and functional effects of non-coding mutations have been mostly ignored. This is primarily because of methodology (application of exome sequencing), and also because of a lack of standard methods to characterize and interpret non-coding variants. Unlike most cancer data produced previously, one of the prominent features of the WGS pan-cancer data is availability of whole-genome sequences, in many cases linked with transcriptome and methylome data. These data provide an unprecedented opportunity for analysis of non-coding regions to identify potential non-coding drivers across multiple tumor types.

Preliminary results. We have developed statistical models of open chromatin associated with gene-expression (Cheng et al., *Nucleic Acids Res.*, 2011; Cheng et al., *Genome Biol.,* 2011) and ncRNA-finder (Lu et al., *Genome Res.*, 2011). We have extensively analyzed patterns of sequence variants in non-coding regions along with the likely protein-coding target genes of these regions (Mu et al., *NAR*, 2011; Yip et al., *Genome Biology*, 2012; Gerstein et al., *Nature*, 2012; Khurana et al., *PLoS Comp Bio*, 2013). By contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, we developed a scheme and a software tool (FunSeq) for identification of candidate non-coding driver mutations (Khurana et al., *Science*, 2013). In this study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project data, with cancer genomics data. Using FunSeq, we were able to identify ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples. Our tool identifies potential driver events in various non-coding functional elements, including: transcription-factor (TF) binding sites, their higher resolution motifs, regions of active chromatin corresponding to enhancer elements and regions of open chromatin corresponding to DNase I hypersensitivity sites. We have followed this up with tumor-vs-normal gene expression studies on genes associated with the regulatory changes. We are also designing experiments to introduce non-coding mutations into Bacterial Artificial Chromosomes carrying the appropriate regions of the human genome; with these we will transform cell lines to test for oncogenic activity. As an alternative we will also attempt to investigate the effect of (site-specific) DNA regulatory region alterations introduced into cell line genomes using CRISPR technology (Cong et al., *Science* 2013). Positive results can be followed up with xenograft models or through introducing orthologous mutations into the mouse genome.

Timelines & resources dedicated to project

1) Initial application of Yale FunSeq pipeline as well as the recently developed EMBL non-coding driver pipeline on partial sample set (after calling of somatic mutations in 500 sample sets is finished): will begin April 2014.

2) Analysis of full somatic call set to identify candidate drivers with FunSeq pipeline and EMBL pipeline (to be completed August 2014: expected 30 CPU min. per sample).

3) Integration with transcriptome and methylome data (we foresee the possibility to interact with transcriptome working groups; *e.g.* Stegle/Brazma, or others) in this regard (will begin September 2014).

4) Joint analysis of entire data, with models about non-coding mutation recurrence, will begin October 2014.

5) Present results at ICGC meeting in spring 2015, release source code, followed by paper submission.

EK will spend 70% of her time on this project, VR: 70%, JK: 20%, MG: 5%, KW: 5%.

Research proposal

We plan to apply our approach on large-scale whole-genome cancer data sets produced by WGS pan-cancer effort to predict functional consequences of non-coding mutations. FunSeq in its current form will provide great insights and lists of potential non-coding driver events across all cancer types. Furthermore, these data provide a unique opportunity for various additional, novel analyses of non-coding somatic mutations, some of which are listed below:

(1) Identification of statistically significant recurrent intergenic mutations in specific cancer entities. A strong indicator of potential driver mutations is the recurrence of those mutations in multiple cancer samples. Thus, availability of thousands of tumor genomes would allow us to systematically interrogate the recurrence of mutations across diverse tumor types and examine cross-cancer similarities and within-cancer heterogeneities.

We will additionally develop statistical models to identify recurrent candidate driver mutations within a given cancer entity in the context of mutation rate heterogeneity. Somatic mutation rates tend to be non-uniform throughout the genome – with tissue/disease entity-related dependencies, *e.g.* gene mutation status (*e.g.* somatic hypermutation in lymphoma), chromatin structure, transcriptional activity, and replication timing (Schuster-Böckler and Lehner, *Nature* 2012; Lawrence *et al. Nature* 2013; Alexandrov et al. *Nature* 2013). Thus, observed mutation recurrences need to be carefully weighed against some background expectancy. To this end we will make use of genetic algorithms (*e.g.* identity genes with same chromatin and cell types) to generate baseline estimates for the expected mutation rate. We will further compare mutation rates with observed rates in the immediate genomic neighborhood of somatic mutational events, to generate statistical models facilitating the identification of driver events in pan-cancer data.

(2) Integration of sequence variants with transcriptome and epigenome: Currently our scheme relies mostly on DNA re-sequencing data from cancer samples. The availability of expression and epigenetic data from a large number of samples will undoubtedly allow novel integrative analyses and is likely to greatly increase our power for prediction of regulatory drivers. Target genes of regulatory elements are expected to show differential expression in tumor relative to normal samples. Availability of RNA-seq data will thus allow us to integrate gene expression to predict non-coding drivers. Our analyses will involve controlling for population structure to identify population-based confounding effects. We will also integrate chromatin and transcriptome data to predict target genes of distal regulatory elements (Yip et al., *Genome Biology*, 2012).

(3) Driver mutations in noncoding RNA: Small noncoding-RNAs (like miRNA, snoRNA etc) play an essential role in the machinery for gene regulation and have been implicated in cancer before. However, they cover a much smaller fraction of the genome than other non-coding elements like TF binding sites and hence the sample size needed to analyze signatures of selection in them is much higher. In the past, inherited and somatic variants from a very large sample size were not available to develop a systematic approach for analyzing the effects of mutations in noncoding-RNAs. TCGA/ICGC data will thus be extremely useful to uncover damaging mutations in these important regulatory elements. We also plan to analyze the larger lincRNAs (long intergenic noncoding-RNAs) for presence of potential driver events. Candidate non-coding driver mutations will further be evaluated with Phylogenetic Module Complexity Analysis (Claussnitzer et al. *Cell*, in press) to identify sites with potential cis-regulatory functionality.

(4) Integration with ENCODE & Cancer subgroup. Mark Gerstein and Kevin White are the co-chairs of the ENCODE & Cancer subgroup. They will feed practical experience from running FunSeq on the full somatic call set into ENCODE, trying to refine non-coding annotations to make them more suitable for cancer -- e.g. adjusting the size distribution of the annotated regions. They also will integrate the latest ENCODE results into the analysis, serving as a coordination point between the ENCODE & Cancer AWG and the WGS pan-can consortium.

Legacy plans

All of the software that we will develop for the project in terms of intersecting the somatic variants with functional annotation will be made freely available to project participants and the community. All the files produced will be small simple files that do not have privacy issues so we believe that we can distribute them readily in the community.

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci* USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

Mark Gerstein

Education

Harvard College, AB Physics '89
Cambridge University, PhD Chemistry '93
Stanford University, postdoc '93-'96, Bioinformatics (advisor M Levitt)

Positions

2006- **AL Williams Prof. Biomedical Informatics, Yale**
2002- co-director Yale Computational Biology and Bioinformatics Program
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

Honors

'89-'93 Herchel-Smith Scholarship for PhD at Cambridge
'93-'96 Damon Runyon-Walter Winchell post-doctoral Fellowship
'09 AAAS Fellow

Consortia

Analysis co-chair: NHGRI modENCODE Project AWG ('07-), Brainspan Project ('09-), 1000 Genomes Functional Interpretation Group ('12-), ENCODE & Cancer Group ('13-) exRNA consortium ('13-)

Publications (senior author on all papers listed below, which are selected from a total of >460; H-index=116)

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, M Gerstein (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342:1235587

E Khurana, Y Fu, J Chen, M Gerstein (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comp Bio* 9:e1002886.

M Gerstein, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." *Nature* 489:91

A Abyzov, J Mariani... (16 authors)... M Gerstein, FM Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells." *Nature* 492:438

B Pei, C Sisu... (10 authors)... J Harrow, M Gerstein (2012). "The GENCODE pseudogene resource." *Genome Biol* 13:R51.

C Cheng, R Alexander... (16 authors)... M Gerstein (2012). " Understanding transcriptional regulation by integrative analysis of transcription factor binding data." *Genome Res* 22:1658.

DG MacArthur, S Balasubramanian... (50 authors)... M Gerstein, C Tyler-Smith (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." *Science* 335:823.

A Abyzov, AE Urban, M Snyder, M Gerstein (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21:974

A Sboner, L Habegger... (9 authors)... MA Rubin, M Gerstein (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." *Genome Biol* 11:R104.

HY Lam, XJ Mu, AM Stütz, A Tanzer, PD Cayting, M Snyder, PM Kim, JO Korbel, M Gerstein (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." *Nat Biotech* 28:47.

RP Alexander, G Fang, J Rozowsky, M Snyder, M Gerstein (2010). "Annotating non-coding regions of the genome." *Nat Rev Genet* 11:559.

KK Yan, G Fang, N Bhardwaj, RP Alexander, M Gerstein (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107:9186.

N Bhardwaj, KK Yan, M Gerstein (2010). "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels." *PNAS* 107:6841

M Gerstein, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330:1775.

EKTA KHURANA
Program in Computational Biology and Bioinformatics
Molecular Biophysics and Biochemistry Dept.
Yale Univ., USA
Email: ektakhurana@gmail.com

POSITIONS

Associate Research Scientist (Lab PI: M Gerstein), Yale Univ.	2012 to Present
Postdoctoral Research Associate (advisor: M Gerstein), Yale Univ.	2008 to 2012

EDUCATION

Ph. D. (Chemistry, Research area: Computational Biology) Advisor: ML Klein; Univ. of Pennsylvania, USA	2002 to 2008
Masters of Science (Chemistry, Research area: Computational Biology) Advisor: B Jayaram; Indian Institute of Technology, Delhi, India	2000 to 2002
Bachelors of Science (Hons. in Chemistry) St. Stephen's College, Delhi Univ., India	1997 to 2000

HONORS AND AWARDS

EMBL Corporate Partnership Fellowship: Cancer Genomics conf. at EMBL, Germany (2013)
Travel Grant: CECAM Workshop on ion channels at Lyon, France (2007)
Chair's fund: Gordon conf. on computer aided drug design at New Hampshire, USA (2007)
Marie Curie fellowship: International School of Solid State Physics for summer school at Erice, Italy (2005)
Science Meritorious Award: Delhi University for academic excellence (1997-1998)

SELECTED PUBLICATIONS

(*Equal contribution, #Corresponding author, 25 articles with 11 as first/co-first and 7 as corresponding author)

E Khurana*, Y Fu*, V Colonna*, X Mu*, HM Kang, T Lappalainen…..1000 Genomes Project Consortium….. M Rubin, C Tyler-Smith, M Gerstein, "Integrative annotation of variants from 1092 humans: application to cancer genomics", *Science*, 342, 84 (2013)

E Khurana*, Y Fu*, J Chen, M Gerstein, "Interpretation of genomic variants using a unified biological network approach", *PLoS Computational Biology*, 9, e1002886 (2013)

M Gerstein*, A Kundaje*, M Hariharan*, S Landt*, K Yan*, C Cheng*, X Mu*, **E Khurana***, J Rozowsky*, R Alexander*, R Min*, P Alves*, A Abyzov, N Addleman, N Bhardwaj…40 authors…M Snyder, "Architecture of the human regulatory network derived from ENCODE data", *Nature*, 489, 91 (2012)

The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome", *Nature*, 489, 57 (2012)

Z Lu…… **E Khurana**…..M Gerstein, "Prediction and characterization of non-coding RNAs in C. elegans by integrating conservation, secondary structure and high throughput sequencing and array data", *Genome Research*, 21, 276 (2011)

E Khurana, H Lam, C Cheng, N Carriero, P Cayting, M Gerstein, "Segmental duplications in the human genome reveal details of pseudogene formation", *Nucleic Acids Research*, 38, 6997 (2010)

E Khurana#, MD Peraro#, R DeVane, S Vemparala, WF DeGrado#, ML Klein, "Molecular dynamics calculations suggest a conduction mechanism for the M2 proton channel from influenza A virus", *Proceedings of the National Academy of Sciences USA*, 106, 1069 (2009)

E Khurana#, R DeVane, A Kohlmeyer, ML Klein, "Probing peptide nanotube self-assembly at a liquid-liquid interface with coarse-grained molecular dynamics", *Nano Letters*, 8, 3626 (2008)

CURRICULUM VITAE – Vasilisa Rudneva

PhD student
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstraße 1
69117 Heidelberg Germany
Email: rudneva@embl.de

Academic Education & Working Experience

Since 2012	PhD student at EMBL Heidelberg, Germany
2011	Scientist at Medical Faculty, Dresden University of Technology, Germany
2011	Guest scientist at Bioquant, Heidelberg University, Germany
2011	Specialist in Bioengineering, Lomonosov Moscow State University, Moscow
2006-2011	Junior researcher at Laboratory of Lipid Systems Biology, A.N.Belozersky Institute of Physico-Chemical Biology, Moscow

Other Professional Experience

2013	Organizing committee, 15th EMBL PhD Symposium

Publications

Ivliev AE, **Rudneva VA**, Sergeeva MG (2010). Applicability of coexpression networks analysis to anticancer drug targets discovery. *Mol Biol (Mosk)*.44(2):366-74

Kevin P. White

Education

Yale University, New Haven, B.S./M.S., Biology 1993
Stanford University, Stanford, CA, Ph.D., Developmental Biology 1998
Stanford Genome Technology Ctr, Palto Alto, CA, Postdoc, Biochemistry & Genomics, 1998-2000

Professional Positions

2006-present	Director, Joint Institute for Genomics & Systems Biology, The University of Chicago and Argonne National Laboratory
2006-present	James and Karen Frank Family Professor, Human Genetics and Ecology & Evolution, The University of Chicago
2004-2006	Associate Prof. of Ecology & Evolutionary Biology (joint appointment), Yale University
2004-2006	Associate Professor of Genetics, Yale University School of Medicine
2001-2004	Assistant Professor of Genetics, Yale University School of Medicine

Publications Selected from 97 peer-reviewed publications

1. Michelle N. Arbeitman, Eileen E. M. Furlong, Farhad Imam, Eric Johnson, Brian H. Null, Bruce S. Baker, Mark A. Krasnow, Matthew P. Scott, Ronald W. Davis and Kevin P. White. Gene Expression During the Life Cycle of Drosophila melanogaster. **Science,** 297: 2270-2275, **2002.**
2. Giot L, Bader JS, Brouwer C, Chaudhuri, et al. A genome-scale protein interaction map of Drosophila melanogaster. **Science**, 302: 1727-36, **2003.**
3. Viktor Stolc*, Zareen Gauhar*, Christopher Mason*, Gabor Halasz, Marinus F. van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Barbano, Harmen J. Bussemaker, and Kevin P White. A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster. **Science**, 306:655-60, **2004.**
4. Scott Rifkin, David Houle, Junhyong Kim and Kevin P. White. A mutation accumulation assay reveals extensive capacity for rapid gene expression evolution. **Nature,** 438:220-3, **2005.**
5. Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed and Kevin P. White. "Expression profiling in primates reveals a rapid evolution of human transcription factors." **Nature,** 440:242-5, **2006**.
6. Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Nègre N, Ludwig M, Stricker T, Al-Ahmadie HA, Tretiakova M, Camp RL, Perera-Alberto M, Rimm DL, Xu T, Rzhetsky A, White KP. Analysis of Drosophila Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer. **Science,** 323:1218-22, **2009**.
7. Hua SJ, Kittler R, and White KP. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. **Cell.** 137:1259-71, **2009**.
8. modENCODE Consortium, et, al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. **Science**. 330:1787-97. **2010**
9. Nègre N*, Brown CD*, Ma L*, Bristow CA*, Miller S*, Kheradpour P, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony R, Chen J, Hwang L, Wagner U, Auburn R, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Bild NA, Grundstad AJ, Hanley D, Mannervik M, Venken K, Bellen H, White R, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. A cis-regulatory map for the Drosophila genome. **Nature**. 471:527-31. **2011**.
10. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. **Nature.**489:57-74. 2012.:
11. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumors, **Nature.** 490:61-70. 2012
12. Xiaochun Ni, Yong E. Zhang, Nicolas Negre, Sidi Chen, Manyuan Long and Kevin P. White. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. **PLoS. Biology**. 10(11):e1001420. 2012.
13. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. **Blood**. 121: 975-83. 2013.
14. Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP. A comprehensive nuclear receptor network for breast cancer cells. **Cell Rep.** 3:538-51. 2013.
15. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. **Cell**. Sep 26;155:70-80. 2013.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

A pan-cancer analysis of *Alu* retro-transposition and post-transposition mutagenesis

Name, institute & ICGC/TCGA affiliation of principal investigator

Prof. **Hong Xue**, Hong Kong University of Science & Technology; ICGC Technology Working Group

Names & institute of junior investigators	Name & institute of non-ICGC collaborator
Dr. Sue Tsang, and Mr. Xiaofan Ding, Hong Kong University of Science and Technology	Dr. Ting Fung Chan, The Chinese University of Hong Kong

Background and preliminary data

Alu retroelements are members of the family of Short Interspersed Nucleotide Elements, representing the most abundant and active retrotransposon in humans with over 1 million copies. Retrotransposition events are known to be disease-causing and a number of mechanisms have been proposed to account for this effect. These include (1) *Alu* insertion-mediated alterations to gene expression by either disruption of a coding region or splicing signal; and (2) non-allelic homologous recombination between *Alu* elements that contribute to genome instability. Moreover, apart from germline activity, *Alu* elements may also be active in somatic tissues and contribute to genome instability throughout the life of an individual. That *Alu* elements play an important role in cancer etiology is strongly supported, but the precise consequences of *Alu* retrotransposition, the intricate relationships between *Alu* elements and other genomic features, and their contribution to somatic cancer-related genetic variations are still unclear. Availability of the pan-cancer WGS dataset will allow us to examine these questions at the whole genome level, particularly the contribution of *Alu* elements to somatic mutations through either active retrotranspositions or acting as recombination hotspots throughout the genome.

Previously, our work demonstrated that the recombination processes furnish a mechanism for the enhancement of SNP frequencies at *Alu* element insertion sites[1]. Moreover, among the three classes of *Alu*-elements, *viz.* the youngest *Alu*-Y, the intermediate-age *Alu*-S and the oldest *Alu*-J, *Alu*-Ys are associated with the highest SNP frequencies. In addition, we have recently developed the genome-wide sequencing method AluScan which utilizes inter-*Alu* PCR to generate for massively parallel sequencing a complex pool of amplicons genomic regions flanked by *Alu* elements that are prone to genetic mutations[2]. By focusing only on such regions, our novel method is efficient in terms of both cost and DNA requirement, thereby allowing studies of large sample sets. Due to the repeat-rich nature of *Alu*-neighboring regions, we have developed a SVM-based algorithm to assist in removing misaligned reads after BWA alignment and prior to genotype calling to increase calling accuracy. This algorithm can be applied to any BWA aligned data that need to forego repeat masking in analysis. In a preliminary study, we used WGS data from three control-tumor pairs and found that frequencies of somatic SNVs, as in the case of germline SNPs, were enhanced within *Alu* elements and neighboring regions, thus increasing the probability of *Alu* mediation of cancer-related mutations[3]. Similar analysis was also extended to the profiles of somatic CNVs and indels in relation to that of *Alu* elements based on AluScan data.

Timelines & resources dedicated to project

12/2013 - 04/2014: Provide AluScan data for 200 control-tumor pairs;
01/2014 - 05/2014: Adapt the SVM algorithm to WGS data analysis under cloud computing;
06/2014 - 08/2014: Variant calling after SVM-based filtration;
09/2014 - 10/2014: Data analysis for *Alu* retro-transposition and somatic mutations;
11/2014 - 12/2014: Paper writing for publications and disseminations.
Data contribution: In addition to the 2000 WGS data to be produced collectively by ICGC team, the PI's group will provide 200 pairs of AluScan data for five cancer types.
Resources: A GPU cluster and cloud-based virtual machines will be provided by the PI's institute.

Research proposal

The present proposal aims to examine the role of *Alu* retroelements in cancer through the analysis of correlations between *Alu* profiles (including new somatic retrotranspositions as well as existing *Alu* elements in the reference genome and personal germ line genome) and those of various somatic mutation types including SNVs, indels, LOHs and CNVs in the vicinity of *Alus*. The availability of the pan-cancer data sets allows analysis at the level of the whole genome, and this will provide insight into the mechanisms of *Alu*-mediated mutations in unprecedented details.

The workflow for our proposal is as follows:

(1) In order to analyze the repeat-rich *Alu*-associated sequence regions, the commonly taken step of masking sequence repeat regions in the reference genome prior to alignment of sequence reads cannot be practically performed. This brings in the problem of high levels of misalignment in these repeat regions. To address this, we have developed an SVM-based algorithm that will identify probable misaligned reads and remove them from further analysis[4]. This algorithm has been applied to our AluScan data and shown to be successful in removing approximately 20% of reads after initial alignment. To handle WGS data, we need to modify the algorithm to increase input-output efficiency and allow data handling on multiple threads.

(2) After ensuring the accuracy of the aligned reads, the data will be subjected to the various variant calling pipelines to generate profiles of SNVs, indels, LOHs and CNVs. While variant calling for each sample can be achieved using standard pipelines, at present software for their processing at the population level is not well developed, especially for the examination of recurrent LOHs and CNVs. Previously, population level CNV analysis has been based on microarray data, and the GISTIC program has become integral to such analysis. We propose to adapt this program for use with WGS data to obtain profiles of recurrent LOHs and CNVs. Moreover, our analysis using microarray data has suggested that subsets of recurrent focal CNVs selected from total CNVs can be used to differentiate between control-subjects and cancer patients[3]. We will further explore this finding using the pan-cancer WGS data.

(3) Having obtained profiles of the various somatic mutation types, we will perform correlation analysis between the profiles of each mutation type and the *Alu* profiles. The *Alu* profiles will include information on *Alu* type, genomic loci, and *Alu* activity derived from retro-transposition data.

(4) We will perform a parallel study on data obtained from WGS and data generated with AluScan methods. Thus the findings will be confirmed by two independent methods, and also enable an in-depth comparison of the two methods.

References:

1. Siu-Kin Ng and Hong **Xue** (2006) Alu-Associated Enhancement of Single Nucleotide Polymorphisms in the Human Genome. ***Gene*** 368: 110-116

2. Lingling Mei, Xiaofan Ding, Shui-Ying Tsang, Frank W. Pun, Siu-Kin Ng, Cunyou Zhao, Dezhi Li, Weiqing Wan, Gilberto Ka Kit Leung, Ho-Keung Ng, Liwei Zhang and Hong **Xue** (2011) AluScan: a method for genome-wide scanning of sequence variations in the human genome. ***BMC Genomics*** 12:564

3. Jianfeng Yang, et al. and Hong **Xue** (2013) Alu element-centered somatic mutation hotspots in cancer genome. (In preparation)

4. Xiaofan Ding et al. and Hong **Xue** (2013) Application of machine learning to development of copy number variation-based prediction of cancer risk. (Submitted)

Legacy plans

The Support Vector Machine-based misalignment removal algorithm will be made available in executable code for both Linux cluster and cloud-based parallel computing and documented to make possible ready application by third parties.

Curriculum Vitae

Hong Xue

Education:

1978-1983	M.D., Second Military Medical University, China
1983-1986	M.Sc., Second Military Medical University, China
1986-1988	Postgraduate research, Second Military Medical University, China
1989-1993	PhD, University of Toronto, Canada

Experience:

2010-	Professor, Division of Life Science, HKUST, Hong Kong
2009-	Member, Technology Group, International Cancer Genome Consortium
2008-	Director of Biomedical Sciences, Ministry of Science and Technology International Collaboration Base, Nansha, Guangzhou, China
2008-2009	Director, Board of Directors, International Society of Computational Biology
2007-	Director, Applied Genomics Center, HKUST, Hong Kong
2006-	Director, HKH Bioinformatics Center, HKUST, Hong Kong
2003-2006	Member, International HapMap Consortium
2001-2009	Associate Professor, Department of Biochemistry, HKUST, Hong Kong
1995-2000	Assistant Professor, Department of Biochemistry, HKUST, Hong Kong
1993-1995	Post-doc Fellow, Department of Genetics, Glasgow University, UK

Representative publications:

1. Cunyou Zhao and Hong **Xue** (2012) A simple method for high-throughput quantification of genome-wide DNA methylation by fluorescence polarization. ***Epigenetics*** 7:335-339
2. Lingling Mei, Xiaofan Ding, Shui-Ying Tsang, Frank W. Pun, Siu-Kin Ng, Cunyou Zhao, Dezhi Li, Weiqing Wan, Gilberto Ka Kit Leung, Ho-Keung Ng, Liwei Zhang and Hong **Xue** (2011) AluScan: a method for genome-wide scanning of sequence variations in the human genome. ***BMC Genomics*** 12:564
3. Pun FW, Zhao C, Lo WS, Ng SK, Tsang SY, Nimgaonkar V, Chung WS, Ungvari GS, **Xue** H. Imprinting in the schizophrenia candidate gene *GABRB2* encoding $GABA_A$ receptor β_2 subunit. (2010) ***Mol Psychiatry*** 16: 557-568
4. The International Cancer Genome Consortium (2010) International network of cancer genome projects. ***Nature*** 464, 993-998.
5. Xiang Wan, Can Yang, Qiang Yang, Hong **Xue**, Nelson L.S. Tang and Weichuan Yu (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. ***American Journal of Human Genetics*** 87:325-340
6. Siu-Kin Ng, Wing-Sze Lo, Frank W. Pun, Cunyou Zhao, Zhiliang Yu, Jianhuan Chen, Ka-Lok Tong, Zhiwen Xu, Shui-Ying Tsang, Qiang Yang, Weichuan Yu, Vishwajit Nimgaonkar, Gerald Stöber, Mutsuo Harano, and Hong **Xue** (2010) A recombination hotspot in a schizophrenia-associated region of *GABRB2*. ***PLoS ONE*** 5(3): e9547
7. Qi Liu, Qian Xu, Vincent W. Zheng, Hong **Xue**, Zhiwei Cao and Qiang Yang (2010) Multi-task learning for cross-platform siRNA efficacy prediction: an in-silico study. ***BMC Bioinformatics*** 11:181
8. Zhao, C., Xu, Z., Wang, F., Chen, J., Ng, S.K., Wong, P.W. Yu Z, Pun FW, Ren L, Lo WS, Tsang S.Y. and **Xue**, H. (2009) Alternative-splicing in the exon-10 region of $GABA_A$ receptor β_2 subunit gene: relationships between novel isoforms and psychotic disorders. ***PLoS One*** 4, e6977
9. Can Yang, Wan, Xiang, Qiang Yang, Hong **Xue**, Nelson Tang and Weichuan Yu (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. ***Bioinformatics*** 25: 504-511
10. The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. ***Nature*** 449, 851-861.
11. Ng, SK and **Xue**, H (2006) Alu-Associated Enhancement of Single Nucleotide Polymorphisms in the Human Genome. ***Gene*** 368: 110-116.
12. Lo WS, Lau CF, Xuan Z, Chan CF, Feng GY, He L, Cao ZC, Liu H, Luan QM and **Xue** H (2004) Association of SNPs and haplotypes in $GABA_A$ receptor β_2 gene with schizophrenia. ***Molecular Psychiatry*** 9(6): 603-8.

Curriculm Vitae - Sue, Shui Ying Tsang

Academic Qualifications:

1993-1996 Rowett Research Institute / University of Aberdeen: PhD in Biochemistry
1989-1992 Fitzwilliam College, University of Cambridge: BA(Hons) in Natural Sciences

Working Experience:

2013-present: Research Associate in the Division of Life Science at HKUST.
2010-2013: Research Assistant Professor in the Division of Life Science at HKUST.
2008-2010: Research Assistant Professor in the Department of Biochemistry at HKUST.
2007-2008: Visiting Scholar in the Department of Biochemistry at HKUST.
2006-2007: Visiting Assistant Professor in the Department of Biochemistry at HKUST.
1997-2005: Research Associate in the Department of Biochemistry at HKUST.
1996-1997: Postdoctoral fellow in the Department of Biochemistry at CUHK

List of Publications:

S.Y. Tsang, S. Zhong, L. Mei, J. Chen, S.K. Ng, F.W. Pun, C. Zhao, B. Jing, R. Chark, J. Guo, Y. Tan, L. Li, C. Wang, S.H. Chew and H. Xue (2013) Social cognitive role of schizophrenia candidate gene *GABRB2*. PLoS One 8:e62322

C. Zhao, F. Wang, F.W. Pun, L. Mei, L. Ren, Z. Yu, S.K. Ng, J. Chen, **S.Y. Tsang** and H. Xue (2012) Epigenetic regulation on *GABRB2* isoforms expression: Developmental variations and disruptions in psychotic disorders. Schizophr Res. 134:260-266

L. Mei, X. Ding, **S.Y. Tsang**, F.W. Pun, S.K. Ng, C. Zhao, D. Li, W. Wan, G.K.K. Leung, H.K. Ng, L. Zhang and H. Xue (2011) AluScan: a method for genome-wide scanning of sequence variations in the human genome. BMC Genomics 12:564

F.W. Pun, C. Zhao, W.S. Lo, S.K. Ng, **S.Y. Tsang**, V. Nimgaonkar, W.S. Chung, G.S. Ungvari and H. Xue (2010) Imprinting in the schizophrenia candidate gene *GABRB2* encoding GABA(A) receptor beta(2) subunit. Mol Psychiatry 16:557-568

S.K. Ng, W.S. Lo, F.W. Pun, C. Zhao, Z. Yu, J. Chen, K.L. Tong, Z. Xu, **S.Y. Tsang**, Q. Yang, W. Yu, V. Nimgaonkar, G. Stöber, M. Harano and H. Xue (2010) A recombination hotspot in a schizophrenia-associated region of *GABRB2*. PLoS One 5: e9547

J. Chen, **S.Y. Tsang**, C.Y. Zhao, F.W. Pun, Z. Yu, L. Mei, W.S. Lo, S. Fang, H. Liu, G. Stöber and H. Xue (2009) GABRB2 in schizophrenia and bipolar disorder: disease association, gene expression and clinical correlations. Biochem Soc Trans. 37: 1415-1418

S.Y. Tsang, S.K. Ng, Z. Xu and H. Xue (2007) The evolution of $GABA_A$ receptor-like genes. Mol Biol Evol 24(2): 599-610

W. S. Lo, M. Harano, Z. Yu, J. Chen, F. W. Pun, K. L. Tong, C. Zhao, S. K. Ng, **S. Y. Tsang**, N. Uchimura, G. Stoeber and H. Xue (2007) *GABRB2* Association with Schizophrenia: Commonalities and Differences Between Ethnic Groups and Clinical Subtypes. Biological Psychiatry 61(5): 653-660

C. Zhao, Z. Xu, J. Chen, Z. Yu, K. L. Tong, W. S. Lo, F. W. Pun, S. K. Ng, **S. Y. Tsang**, and H. Xue (2006) Two isoforms of $GABA_A$ receptor β_2 subunit with different electrophysiological properties: differential expression and genotypical correlations in schizophrenia. Molecular Psychiatry 11(12): 1092-1105

S. Y. Tsang and H. Xue (2004) Development of effective therapeutics targeting the $GABA_A$ receptor: naturally occurring alternatives. Current Pharmaceut. Des. 10: 1035-44

Patents held:

Compositions and Methods for the Targeted Delivery of Agents to Treat Liver Cancer. **S. Y. Tsang** and J. T.F. Wong, United States Patent No. 7005139 (Date: 28 Feb 2006)

CURRICULUM VITAE

Name: Xiaofan Ding
Date of Birth: April 23, 1988
E-mail: xding@ust.hk; dingxiaofan1@gmail.com

Education:

Feb 2010 – present	**The Hong Kong University of Science and Technology**
	Division of Life Science
	PhD of Biochemistry
Sep 2005 – June 2009	**SUN YAT-SEN UNIVERSITY**
	Dept. of Life Science
	Bachelor in Biology Science degree

Personal information:

Projects next generation sequencing analyses & tumor genomics

Computer

Languages:	C++/Java/Perl/R/Octave/Linux bash
Analysis Tools:	Weka, R, Octave
Sequence analysis:	Bwa, bowite, soap2, samtools, GATK etc.
Microarray analysis:	Bioconductor, aroma package etc.
Others:	Ensembl, Bioperl, Modeller, MySQL etc.

Interests Biology, Computer skills, Probability graph model, Machine learning

Publications:

1. Song, L., Li W., Zhang H., Liao W., Dai T., Yu C., **Ding X.**, Zhang L. & Li J. Over-expression of AEG-1 significantly associates with tumour aggressiveness and poor prognosis in human non-small cell lung cancer. *The Journal of pathology* **219**, 317-326 (2009).
2. Li, M., Li, J., **Ding X.**, He, M. & Cheng, S.Y. microRNA and cancer. *The AAPS journal* **12**, 309-317 (2010).
3. **Ding, X.**, *et al*. Amino acid sequence analysis and identification of mutations under positive selection in hemagglutinin of 2009 influenza A (H1N1) isolates. *Virus genes* **41**, 329-340 (2010).
4. Mei, L., **Ding X.**, *et al*. AluScan: a method for genome-wide scanning of sequence and structure variations in the human genome. *BMC genomics* **12**, 564 (2011).

Name: Ting-Fung CHAN

Education:
1997 B.S., University of Wisconsin, Madison
2003 Ph.D., Washington University, Saint Louis

Previous position: 2004-2007 Postdoctoral fellow
University of California, San Francisco

Present positions: Assistant Professor, School of Life Sciences, and Dept. of Computer Sciences & Engineering, The Chinese University of Hong Kong
Assistant Professor, CUHK-BGI Innovation Institute of Trans-omics
Co-director, Hong Kong Bioinformatics Centre, CUHK

Relevant work:
High-throughput sequencing data analysis and software development in RNA bioinformatics; integrative analysis of multi-omic datasets for the study of genotype-phenotype relationships in complex traits.

Publications: (as either first or corresponding authors)

Section A: (*** denotes corresponding author**)

1. Law, P.T., Qin, H., Ching, A.K., Lai, K.P., Co, N.N., He, M., Lung, R.W., Chan, A.W., **Chan, T.F.***, and Wong, N*. (2013). Deep sequencing of small RNA transcriptome reveals novel non-coding RNAs in hepatocellular carcinoma. *J Hepatol* 58(6), 1165-73.
2. Li, J.W., Wan, R., Yu, C.S., Co, N.N., Wong, N., and **Chan, T.F.*** (2013). ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* 29(5), 649-51.
3. Li, S.K., Ng, P.K., Qin, H., Lau, J.K., Lau, J.P., Tsui, S.K., **Chan, T.F.***, and Lau, T.C*. (2013). Identification of small RNAs in *Mycobacterium smegmatis* using heterologous Hfq. *RNA* 19(1), 74-84.
4. Yu, C.S., Yim, A.K., Tsui, S.K., and **Chan, T.F.*** (2012). Complete genome sequence of *Bacillus subtilis* strain QB928, a strain widely used in *B. subtilis* genetic studies. *J Bacteriol* 194(22), 6308-9.
5. Lou, S. K., Ni, B., Lo, L.Y., Tsui, S.K., **Chan, T.F.***, and Leung, K.S. (2011). ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics* 27(3), 421-2.

Section B:

1. Bodian, D.L.†, **Chan, T.F.†**, Poon, A., Schwarze, U., Yang, K., Byers, P.H., Kwok, P.Y., and Klein, T.E. (2009). Mutation and polymorphism spectrum in osteogenesis imperfecta type II: implications for genotype-phenotype relationships, *Hum Mol Genet* 18(3), 463-471. (**†co-first authors**)
2. **Chan, T.F.**, Poon, A., Basu, A., Addleman, N., Chen, J.W., Phong, A., Byers, P.H., Klein, T.E., and Kwok, P.Y. (2008). Natural variation in four human collagen genes across an ethnically diverse population, *Genomics* 91, 307-314.
3. **Chan, T.F.**, Ha, C., Phong, A., Cai, D., Wan, E., Leung, L., Kwok, P.Y., and Xiao, M. (2006). A simple DNA stretching method for fluorescence imaging of single DNA molecules, *Nucleic Acids Res* 34, e113.
4. Zheng, X. F., **Chan, T. F.**, and Zhou, H. (2004). Genetic and genomic approaches to identify and study the targets of bioactive small molecules, *Chem Biol* 11, 609-618.
5. **Chan, T. F.**, Carvalho, J., Riles, L., and Zheng, X. F. (2000). A chemical genomics approach toward understanding the global functions of the target of rapamycin protein, *Proc Natl Acad Sci USA* 97, 13227-32.

Patent:

Polynucleotide Barcoding; US Ser No. 10/976,546; USPTO Patent No. 7,829,278; Issued: Nov 09, 2010

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Associating Survival with Germline Mutations in Genes of the Immune System in the Background of Distinct Whole-Genome Somatic Mutational Profiles in Solid Cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Partha P. Majumder

Rajiv Sarin

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Arindam Maitra Nidhan K. Biswas	

Background and preliminary data

Mutational changes which alter normal cell growth and survival pathways play a very important role in cellular transformation and tumor development. Recent findings indicate that the host immune system not only protects the host from development of cancer but also regulates tumor growth. The cancer immunosurveillance and regulation of tumor growth, also termed as cancer immunoediting, is based on three Es, which are three distinct stages of elimination, equilibrium and escape. When nascent transformed cells exist as single cells, they are readily eliminated by innate and adaptive immune responses. The equilibrium phase consists of continuous sculpting of tumor cells, which produces cells resistant to immune effector cells. In the escape phase, outgrowth of tumor cell variants that escaped from immune recognition results in expansion of tumor cells. Recent evidences indicate two broad mechanisms of tumor escape. One subset of escape is caused by T-cell inflamed phenotype involving innate immune activation while the other is effected by immune system exclusion or ignorance. Large cancer studies demonstrated that the number, type and location of tumor immune infiltrates in primary tumors, are prognostic for disease free survival and overall survival. Studies have shown that cytotoxic T cells, memory T cells and T_H1 cells are associated with prolonged survival for all cancer types; whereas the prognostic impact of other immune cells such as B cells, NK cells, MDSC, macrophages, and subset of T-helper populations, (T_H2, T_H17, Treg cells) may differ depending on the type of cancer, and on the cancer stage. T_H17 cells have been reported to be associated with poor prognosis in colorectal, lung and hepatocellular carcinoma, in contrast with, for example, esophageal and gastric cancers, for which the prognosis is better because of longer period of disease-free survival. High infiltration of Treg cells has been correlated with poor overall survival in breast cancer and hepatocellular carcinoma, while this phenomenon is associated with longer survival in patients with head and neck cancer. Similar observations are also available in respect of colorectal cancer and melanoma. Subsets of tumors which exhibit lack of T-cell infiltration have been found to have elevated expression of angiogenic factors. Immune evasion of tumor cells can also be caused by the presence of macrophages and other myeloid cells in the tumor microenvironment. Recent studies on whole-exome sequencing have shown that recurrent somatic alterations of genes of the immune system (HLA-A in lung cancer and HLA-B in cervical cancer), indicative of a strong role of the immune system in the development of cancer.

Discovering genetic mechanisms underlying these two phenotypes are of considerable importance. Three possible approaches are:

(a) Investigate heterogeneity of somatic mutation profiles within tumors caused by clonal differences focusing on genes belonging to specific pathways, e.g., RAS, STAT3, NOTCH and ß-catenin, that not only are oncogenic but also impinge on the immune system;

(b) Investigate inter-patient heterogeneity of germline variants in genes that regulate the immune system, including

genes of the innate immune system that may alter thresholds of immunomodulation and sensitivity of signalling pathways influencing T-cell activation. Recent examples of this type are CCR5 polymorphisms correlated to favourable clinical outcome with high dose IL2 treatment and IRF5 polymorphism associated with improved clinical response to tumor-infiltrating lymphocyte adoptive therapy in melanoma.

(c) Investigate differences in environmental contributions among patients, e.g., differential exposure to pathogens resulting in alteration of activation status and frequency of T-cells.

Recruitment of the host immune system for treatment of cancer, i.e. immunotherapy, has the promise to provide the biggest breakthrough in treatment of cancer patients, which is critically dependent of the identification of the nature and causes of heterogeneity outlined above.

Timelines & resources dedicated to project

We shall test the stated hypotheses in a sequential manner. Therefore, we expect to derive inferences sequentially, which we also plan to validate either using additional data on the same cancer type or using data on a different cancer type.

We will require data (WXS and WGS) on solid cancers. For each cancer type, we will require data on a minimum of 50 patients, including data on duration of post-treatment (chemotherapy, surgery, etc.) disease-free survival. We are not tied to any specific cancer type, and we can carry out our analyses sequentially as data become available.

Research proposal

We shall use the available Pan-Cancer DNA sequence data on solid cancers to address the issues stated in (a) and (b) of the section above titled "Background and preliminary data."

Data

We shall use the following resources of the Pan-Cancer Initiative, pertaining to solid cancers only:

(i) Variant calls of samples with WGS and survival data.
(ii) BAM files of WXS for tumour and normal samples for which survival data are available.
(iii) Data on variant calls will be generated by us from the BAM files noted in (ii).

Hypotheses

(i) Somatic mutations in specific oncogenic pathways which impinge on the immune system affect post-treatment disease-free survival.

(ii) Specific subsets of germline variants in genes of the immune system will modulate survival even within a homogeneous molecular tumor subgroup.

(iii) Overall survival is determined not only by the total mutational burden in genes of the immune system, but by the burden in genes involved in different types of immune response. Further, these correlates of survival are variable across cancer types.

Methodologies

(i) We shall create a patient-specific dataset of somatic driver mutations in all genes and germline mutations in genes known to be involved in immune response, separately for all available solid cancers, along with each patient's survival information.
(ii) We shall estimate the association between the total number of somatic mutations in specific oncogenic pathways (e.g., RAS, STAT3, NOTCH and ß-catenin) that also regulate expression of genes of the immune system with survival. We shall also estimate the joint impact of mutations in two of more pathways on survival.
(iii) Using data on germline mutations in genes of the immune system, we shall classify the patients into molecular subgroups, associate types of immune response with these subgroups by mining relevant information from the literature, and test differences in mean survival of patients among these subgroups.

Possible inferences

These analyses would help to identify sub-group of patients in each cancer subtype who are more likely to benefit from immunotherapy.

Legacy plans

Statistical methodology developed for these analyses will be shared openly. Further, computer programs develop to implement these methods will be provided freely in the public domain.

Curriculum Vitae: Partha P. Majumder, PhD

Education & Appointments:

- Obtained undergraduate (B.Stat. [Hons.]), graduate (M.Stat) and doctoral (Ph.D.) degrees from the Indian Statistical Institute, Kolkata (India).
- Post-doctoral work at the Center for Demographic & Population Genetics, Houston, USA.
- Visiting Assistant Professor of Human Genetics & Biostatistics at the University of Pittsburgh from 1987 to 1989.
- Currently the founding Director of National Institute of Biomedical Genomics in Kalyani, India. (On deputation from the Indian Statistical Institute.)

Awards and Honours:

- Recipient of the TWAS Biology Prize for 2009, the New Millenium Science Gold Medal of the Indian Science Congress Association in 2000, the G.D. Birla Award for Scientific Research in 2002, Shri Om Prakash Bhasin Award in Biotechnology in 2001, and the Ranbaxy Research Award in Applied Medical Sciences in 2000.`
- Elected to the Fellowship of the Indian National Science Academy, the Indian Academy of Sciences and the National Academy of Sciences, India.
- Council Member of the Human Genome Organization.

SELECTED PUBLICATIONS (* Corresponding Author)

Disease Genomics & Genetic Epidemiology

India Project Team of the International Cancer Genome Consortium [Corresponding Author: MAJUMDER PP] (2013) Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. *Nature Communications* 4:2873 doi: 10.1038/ncomms3837.

MAJUMDER PP*, Sarkar-Roy N, Staats H, Ramamurthy T, Maiti S, Chowdhury G, Whisnant CC, Narayanasamy K, Wagener DK (2013) Genomic correlates of variability in immune response to an oral cholera vaccine. *European Journal of Human Genetics* 21: 1000-1006.

Datta S., Chowdhury A., Ghosh M., Das K., Jha P., Colah R., Mukerji M, MAJUMDER PP* (2012) A Genome-Wide Search for Non-*UGT1A1* Markers Associated with Unconjugated Bilirubin Level Reveals Significant Association with a Polymorphic Marker Near a Gene of the Nucleoporin Family. *Annals of Human Genetics* 76: 33–41

Chakrabarti S, Ghanekar Y, Kaur K, Kaur I, Mandal AK, Rao KN, Parikh RS, Thomas R, MAJUMDER PP (2010) A Polymorphism in the *CYP1B1* Promoter is Functionally Associated with Primary Congenital Glaucoma. *Human Molecular Genetics* **19: 4083-4090.**

Human Population Genetics and Genome Diversity

Sarkar S, Biswas NK, Dey B, Mukhopadhyay D, MAJUMDER PP* (2010). A Large, Systematic Molecular-Genetic Study of G6PD in Indian Populations Identifies a New Non-Synonymous Variant and Supports Recent Positive Selection. *Infection, Genetics and Evolution* 10:1228-36.

MAJUMDER PP (2010) The Human Genetic History of South Asia. *Current Biology* 20: R184–R187.

Mukherjee S, Sarkar-Roy N, Wagener D, MAJUMDER PP* (2009) Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. *Proceedings of the National Academy of Sciences, USA* 106: 7073-7078.

MAJUMDER PP (2008) Genomic inferences on peopling of south Asia. *Current Opinion in Genetics & Development* 18:280–284

Statistical & Computational Genomics

Sarkar-Roy N, Mondal D, Bhattacharya P, MAJUMDER PP* (2011) A Novel Statistical Algorithm for Enhancing the Utility of HapMap Data to Design Genomic Association Studies in non-HapMap Populations. *International Journal of Data Mining and Bioinformatics* 5: 706-716.

Sarkar Roy N, Farheen S, Roy N, Sengupta S, MAJUMDER PP* (2008) Portability of Tag SNPs Across Isolated Population Groups: An Example from India. *Annals of Human Genetics* 72: 82–89.

Biswas N, Dey B, MAJUMDER PP* (2007) Using HapMap data: a cautionary note. *European Journal of Human Genetics*. 15: 246-249.

Basu, A., Chaudhuri, P, MAJUMDER PP* (2005) Identification of polymorphic motifs using probabilistic search algorithms. *Genome Research* 15: 67-77.

Curriculum Vitae: Rajiv Sarin, MD

Prof. Rajiv Sarin trained in Clinical Oncology & Radiation Oncology at the Tata Memorial Hospital (TMH) & the Royal Marsden Hospital London and Cancer Genetics at University of Utah, Salt Lake City. Prior to the current assignment as the Director of the institute since 2005, he was the lead investigator in Breast Cancer & Brain Tumour Radiotherapy & Cancer Genetics at TMH. Starting from the 1st registry of Hereditary Cancer in the country in 1996, he developed a model Cancer Genetics Unit comprising of a Cancer Genetics Clinic in 2003 & Genetics Lab (SARIN Lab) in 2007. Prof. Sarin is the lead investigator in the International Cancer Genome Consortium (ICGC) India project for Oral Cancers and also the ICMR Centre for Advanced Research in Cancer Genetics & Genomics. He serves as the member of two International working groups of the ICGC.

Prof. Sarin is a member of various expert committees & task forces for cancer research and cancer management & he drafted the base paper for research in the Indian National Cancer Control Programme. He is the Executive Editor of the Journal of Cancer Research & Therapeutics which is the leading Indexed Cancer Journal from developing World. He is also serves on the Editorial board of leading cancer journals like Lancet Oncology, Molecular Oncology, Radiotherapy Oncology, Cancer Biomarkers, Mammology etc. Has over 75 peer reviewed publications and editorials in leading journals on various issues of cancer, health care, Cancer genetics, radiobiology etc.

SELECTED PUBLICATIONS

1. India Project Team of the International Cancer Genome Consortium, Maitra A,Biswas NK, Amin K, et al. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun. 2013 Dec 2;4:2873.

   2. Mahantshetty, U., Jamema, S., Engineer, R., Deshpande, D., **Sarin, R.**, Fogliata, A., Nicolini, G., Clivio, A., Vanetti, E., Shrivastava, S. and Cozzi, L. Whole abdomen radiation therapy in ovarian cancers: a comparison between fixed beam and volumetric arc based intensity modulation. *Radiation Oncology.* 2010. 5, 106.

2. **Sarin R.** Ultra-targeted APBI using TARGIT-a cautionary note. *Nature Reviews. Clinical Oncology.* 2010. 7, 675-676.

3. Munshi, A., Dutta, D., Kakkar, S., Budrukkar, A., Jalali, R., **Sarin, R.**, Gupta, S., Parmar, V. and Badwe, R. Comparison of early quality of life in patients treated with radiotherapy following mastectomy or breast conservation therapy: a prospective study. *Radiotherapy and Oncology : Journal of the European Society For Therapeutic Radiology and Oncology.* 2010. 97, 288-293.

4. Hudson, T. J., … **Sarin, R.**, et al. International network of cancer genome projects. *Nature.* 2010. 464, 993-998.

5. **Sarin R.** From 3D to 5D radiotherapy: a blitzkrieg of DTH! *Journal of Cancer Research and Therapeutics.* 2009. 5, 223-224.

6. Kinhikar, R. A., Jamema, S. V., Pai, R., Zubin, M., Gupta, T., Dhote, D. S., Deshpande, D. D., Shrivastava, S. K. and **Sarin, R.** Dosimetric validation of first helical tomotherapy Hi-Art II machine in India. *Journal of Medical Physics / Association of Medical Physicists of India.* 2009. 34, 23-30.

7. Dwarakanath, B. S., Singh, D., Banerji, A. K., **Sarin, R.**, Venkataramana, N. K., Jalali, R., Vishwanath, P. N., Mohanti, B. K., Tripathi, R. P., Kalia, V. K. and Jain, V. Clinical studies for improving radiotherapy with 2-deoxy-D-glucose: present status and future prospects. *Journal of Cancer Research and Therapeutics.* 2009. 5 Suppl 1, S21-6.

8. Swamidas, V. J., Mahantshetty, U., Vineeta, G., Engineer, R., Deshpande, D. D., **Sarin, R.** and Shrivastava, S. K. Treatment planning of epithelial ovarian cancers using helical tomotherapy. *Journal of Applied Clinical Medical Physics / American College of Medical Physics.* 2009. 10, 3003.

9. Jalali, R., Mallick, I., Dutta, D., Goswami, S., Gupta, T., Munshi, A., Deshpande, D. and **Sarin, R.** Factors influencing neurocognitive outcomes in young patients with benign and low-grade brain tumors treated with stereotactic conformal radiotherapy. *International Journal of Radiation Oncology, Biology, Physics.* 2010. 77, 974-979.

10. Jalali, R., Dutta, D., Srinivas, C., Munshi, A., Limaye, U., Goel, A., Deshpande, D. and **Sarin, R.** Micromultileaf collimator-based stereotactic radiosurgery for selected arteriovenous malformations: technique and preliminary experience. *Journal of Cancer Research and Therapeutics.* 2009. 5, 186-191.

Curriculum Vitae: Arindam Maitra, PhD

Dr. Arindam Maitra has completed his Ph.D. on genetic epidemiology of HIV-1 from All India Institute of Medical Sciences, New Delhi, India in 1999. He is an Associate Professor and the Project Coordinator of the International Cancer Genome Consortium – India Project on gingivobuccal cancer in National Institute of Biomedical Genomics (NIBMG), Kalyani, India. He also leads the Platform Technologies in NIBMG. During his career, he has served in various capacities such as the Group Leader of Genomics in The Centre for Genomic Applications (TCGA), New Delhi and Section Head of Functional Genomics in Thrombosis Research Institute (TRI), Bangalore, India. Dr. Maitra has also acted as an honorary consultant on forensic DNA analysis to Government of India. He has been the Principal Investigator of a five-year Department of Biotechnology (GOI) program grant on "Identification of Genetic Risk Factors of Premature Cardiovascular Disease in the Asian Indian Population" and Co-investigator of Department of Biotechnology (GOI) research study entitled "Genome wide linkage study in families affected with early onset coronary artery disease in the Indian population."

Dr. Maitra has been a recipient of National Eligibility Test (NET) fellowship from Council for Scientific and Industrial Research (CSIR), Govt. of India for pursuing his Ph.D. and Takashi Kurimura Award for the best paper in HIV research conferred in the First International Conference on AIDS India 2000. He has multiple publications, GenBank submissions and a patent on real time PCR based detection and quantitation of HIV-1 to his credit. The following are some of his recent and important publications:

1. India Project Team of the International Cancer Genome Consortium, Maitra A,Biswas NK, Amin K, et al. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun. 2013 Dec 2;4:2873.
2. Shanker J, Maitra A, Arvind P, Nair J, et al. Role of vitamin D levels and vitamin D receptor polymorphisms in relation to coronary artery disease: the Indian atherosclerosis research study. Coron Artery Dis. 2011 Aug;22(5):324-32.
3. Maitra A, Shanker J, Dash D, Arvind P, Kakkar VV. Understanding the expression of Toll-like receptors in Asian Indians predisposed to coronary artery disease. Arch Med Sci. 2011 Oct;7(5):781-7.
4. Maitra A, Shanker J, Dash D, Sannappa PR. Polymorphisms in the pituitary growth hormone gene and its receptor associated with coronary artery disease in a predisposed cohort from India. J Genet. 2010 Dec;89(4):437-47.
5. Shanker J, Perumal G, Maitra A, Rao VS. Genotype-phenotype relationship of F7 R353Q polymorphism and plasma factor VII coagulant activity in Asian Indian families predisposed to coronary artery disease. J Genet. 2009 Dec;88(3):291-7.
6. Maitra A, Shanker J, Dash D, John S. Polymorphisms in the IL6 gene in Asian Indian families with premature coronary artery disease--the Indian Atherosclerosis Research Study. Thromb Haemost. 2008 May;99(5):944-50.
7. Kapoor G, Maitra A, Somlata, Brahmachari V. Application of SNaPshot for analysis of thiopurine methyltransferase gene polymorphism. Indian J Med Res. 2009 May;129(5):500-5.

Curriculum Vitae: Nidhan K. Biswas, PhD

Current Designation: Young Biotechnologist Fellow

Institute: National Institute of Biomedical genomics (NIBMG), Kalyani, India

Date of Birth: 29/10/1979

Education (Post-graduation onwards & Professional Career)

- M.Sc Biochemistry, Calcutta University, 2003.
- Research assistant, Human Genetics Unit, Indian Statistical Institute (ISI), 2003-2004.
- PhD.- Calcutta University on Human Evolutionary Genetics, ISI, 2004-2011.
  PhD supervisor – Prof. Partha P. Majumder, Director – NIBMG.
- Senior Technical Specialist, ICGC-India Project, NIBMG, 2010-2011.
- Senior Data Analyst, ICGC-India Project, NIBMG, 2011-2013.

Professional Experiences and Training relevant to the project
I acted as the group leader of statistical and bioinformatics analysis wing at ICGC-India oral cancer project. I have 10 plus years of research experience in human genomics field. I have expertise in the following area: dry-lab (creation and maintenance of whole genome and exome DNA and RNA sequencing data analysis pipelines, CNV analysis, development of statistical tools and databases handling) and wet-lab (large scale sequencing, genotyping).

Publications:

- India Project Team of ICGC (2013). Mutational landscape of gingivobuccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. **Nature Communications**. 4:2873 doi: 10.1038/ncomms3837.

  Member of Manuscript writing group & statistical-and-bioinformatics analysis group. Group leader – Statistical and bioinformatics analysis.

- Sarkar S, Biswas NK, Dey B, Mukhopadhyay D, Majumder PP (2010). A large, systematic molecular-genetic study of G6PD in Indian populations identifies a new non-synonymous variant and supports recent positive selection. **Infection Genetics Evolution.** 10(8):1228-36.

- Dutta S, Majumder M, Biswas NK, Sikdar N, Roy B (2007) "Increased risk of oral cancer in relation to common Indian mitochondrial polymorphisms and autosomal GSTP1 locus". **Cancer** . 110, Issue 9, , 1991-1999.

- Biswas NK, Dey B, Majumder PP.(2006) " Using HAPMAP data : a cautionary note". **European Journal of Human Genetics**. 15, 246–249.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrated mutation analysis of enhancer elements in pan-cancer genomes.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hiroyuki Aburatani, Genome Science Div. RCAST, The University of Tokyo

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Kenji Tatsuno, RCAST, The University of Tokyo **Hiroki Ueda, RCAST, The University of Tokyo**	

Background and preliminary data

Enormous contributions of ICGC/TCGA have identified driver genomic alterations in many types of cancers. However, in some cases that had very few mutations, we could not explain the driver event in tumorigenesis, which could be partly because current analysis focused mainly on protein coding sequence. Our ongoing analysis in liver cancer project identified extremely frequent mutation in *TERT* gene promoter region, which has been also reported in other tumors, e.g. melanoma, glioma and bladder cancer. In addition, Hepatitis B virus integration in TERT promoter region and focal amplification at *TERT* region were mutually exclusively found in the cases without TERT promoter mutation. These findings suggested that alterations in non-coding regions also could be a driver event in carcinogenesis. Currently we are analyzing interactions between alterations in promoter or enhancer regions and gene expression using WGS and RNA sequence data of liver cancer.

In addition, epigenetic regulators, such as *ARID1A* and *PBRM1*, are frequently mutated in various cancers, although their exact roles in carcinogenesis remain elusive. Our preliminary analysis indicated that, in the cases with mutation in chromatin remodeler genes, the chromatin accessibility is altered in enhancer elements.

One challenge in enhancer mutation analysis is that location of enhancer elements is mostly tissue specific and that we need to examine multiple enhancer elements per one gene. However, recent progress in epigenome mapping project will provide us with information on regulatory elements.

We plan to compare mutation status in enhancer elements in pan-cancer data and will assess their function on our cancer specimens. We hope we will be able to identify cancer specific enhancer variations using pan-cancer date set.

Finally, genetic variation in those enhancer elements could be associated with cancer susceptibility, as proposed in other common diseases.

Timelines & resources dedicated to project

- **Identification of enhancer elements in each tissue type from public and in-house epigenomic studies**
- **Mutation analysis of non-coding regions for liver cancer data and pan-cancer data set will provide gross number of mutation frequency in enhancer elements per gene.**
- Allelic expression status from RNA sequence data of our liver cancer specimens to estimate the effect of mutations in regulatory elements on transcriptional regulation.
- **Integration of genetic alteration and gene expression data to identify cancer specific alterations in enhancer regions.**
- **Test the function of enhancer mutations by reporter assay.**

Research proposal

We would like to propose a research plan for TCGA/ICGC pan-cancer analysis on the following themes: Functional consequences of non-coding mutation or Integration of genome and transcriptome.

Mutations in non-coding regions such as promoter or enhancer regions are known to affect gene expression. Moreover, epigenetic alterations also affect gene expression by modulating chromatin structure. Our research proposal is to perform a mutation analysis of regulatory elements on several cancer types. By comparing genomic alterations, particularly those in the regulatory elements, and gene expression data in several cancers, we will identify cancer common and cancer specific enhancer or promoter alterations involved in tumorigenesis or cancer progression, and also identify a novel therapeutic target for cancer.

1) Selection of regulatory elements:

To select candidate regulatory regions for each cancer type, we will use active epigenomic mark data, such as H3K27Ac and chromatin accessibility, from NIH Roadmap Project and International Human Epigenome Consortium. We will also generate epigenomic data from our clinical specimens.

We will identify significantly mutated enhancer elements, particularly those interrupting the transcriptional factor binding motifs.

2) Regulatory elements in cancer driver genes

Known cancer driver genes, such as *MYC* or *TERT*, have a significant effect in carcinogenesis, so alteration in their regulation could affect the cell differentiation state. Therefore, regulatory elements for those genes must be thoroughly analyzed.

3) Mutational analysis in regulatory elements:

Since the location of the regulatory elements differ among various cancer types, we need to establish a way to count the number of somatic mutations in non-genic regions per gene.

4) Allelic expression status from RNA sequence data:

To evaluate the cis-effect of non-coding mutation, we will evaluate the allelic expression status, which would require accurate estimation of tumor content in cancer tissues analyzed. We may need to examine cancer cell lines to precisely evaluate the effect of enhancer element mutations. To confirm the interaction between the enhancer and its promoter, chromatin interaction assay data, such as ChIA-PET and Hi-C. RNA-seq data will help identify active enhancer elements by the presence of enhancer RNA.

5) Integrative analysis:

To predict the effect of somatic alterations on transcriptional regulation, an integrative analysis of mutations in non-coding regions, epigenetic alterations, and gene expression profiles of cancers would be required.

6) Functional validation of regulatory mutations in cancer cell lines, e.g. reporter assays.

7) Search for germline variants in regulatory elements that are associated with tumor susceptibility.

Legacy plans

Virus integration analysis pipeline, including viral genome capture, has been developed.

We are developing a software, called Karkinos, an integrated genome alteration analysis pipeline, in which enables sensitive and accurate INDEL detection, SNV detection, and CNV analysis.

Hiroyuki Aburatani

Present position
Professor, Genome Science Laboratory
Research Center for Advanced Science and Technology
The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8904

Education

The University of Tokyo, Tokyo, Japan, M.D. 1980 Medicine
The University of Tokyo, Tokyo, Japan, Ph.D. 1988 Medicine

Professional experience

1980-1982	Internship in Internal Medicine, The University of Tokyo
1982-1983	Clinical fellow, Tokyo Metropolitan Komagome Hospital, Tokyo
1983-1988	Clinical Research Fellow, 3rd Dept of Internal med, The University of Tokyo
1988-1991	Assist professor, 3rd Dept. of Internal med, The University of Tokyo
1988-1994	Visiting Scientist, Center for Cancer Research, MIT, Cambridge, MA, USA
1995-1999	Assist professor, 3rd Dept. of Internal med, The University of Tokyo
1999-2001	Associate Professor, Genome Science Lab, Research Center for Advanced Science and Technology, The University of Tokyo
2001-present	Professor, Genome Science Lab, Research Center for Advanced Science and Technology, The University of Tokyo

Research Interest

My research interest is functional genomics, where I apply cutting-edge genomic technologies in biomedical research. I was originally trained as physician-scientist, and started the current laboratory in 1999 at Research Center for Advanced Science and Technology in The University of Tokyo. I have been working on various research topics in genomics and have published more than 300 papers, e.g. SNP array analysis for human copy number variation map (Nature 2006), ChIP-chip for cohesin/CTCF colocalization (Nature 2008), and next generation sequencing for liver cancer whole genome sequencing (Nature genet 2011) and glioma progression (Science 2013). I am interested in epigenomic regulation of gene expression and have been involved in ICGC and IHEC. Current challenge is integration of genomic and epigenomic information to understand the cellular state.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 31st December, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Pan-cancer analysis of the impact of breakage-fusion-bridge cycles in genome instability

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Satoru Miyano, Human Genome Center, The Institute of Medical Science, The University of Tokyo, (ICGC PI)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Atsushi Niida, Yuichi Shiraishi (Human Genome Center, The Institute of Medical Science, The University of Tokyo)	Hisashi Tanaka (Department of Molecular Genetics, Cleveland Clinic Lerner Research Institute)

Background and preliminary data

Gene amplification is a clinically important event in cancer genome evolution, as gene amplification causes aggressive tumor phenotypes, such as tumor progression and therapy resistance. Continuous DNA breaks and rearrangements through breakage-fusion-bridge cycles (BFB cycles) have been considered as a major mechanisms underlying gene amplification and genome instability (Tanaka and Yao, Nat Rev Cancer 2009). However, to what extent BFB cycles contribute to genome instability across human cancers remains to be determined. This is an important issue, because tumors with the signature of BFB cycles could have a potential to drive tumor aggressiveness through gene amplification. In this regard, pan-cancer analysis of 2000 whole genome sequence (WGS) data across all tumor types is a great opportunity to determine the impact of BFB cycles in cancer genome instability.

BFB cycles leave a unique breakpoint signature, fold-back inversion, in cancer genomes. BFB cycles are initiated by the formation of chromosomes with two centromeres either through telomere-telomere fusion or aberrant DNA repair events. During mitosis, each centromere goes to the opposite pole and eventually breaks the chromosome. When the broken chromosome replicates, two sister chromatids fuses at the break, leaving fold-back inversions at the breakpoint. In WGS data, fold-back inversions were identified by read-pairs aligning close together but in inverted orientation. Using the signature, fold-back inversions were reported to be very common and represent one-sixth of somatic breakpoints in pancreatic cancer, but not as common in breast cancer (Campbell, *et al*., Nature 2010). However these results were obtained from small numbers of tumors and cell lines (13 pancreatic and 24 breast cancers). A large-scale study focusing on primary tumors is necessary to fully determine the impact of fold-back inversions in cancer breakpoints.

Two possible outcomes are expected from pan-cancer analysis. First, fold-back inversions are very common in certain types of tumors (e.g. pancreatic tumors) but not other types. This hypothesis is supported by the previous small-scale analyses, and would suggest the cell-origin specific factors that promote BFB cycles. Second scenario is that fold back inversion occurs across all tumor types, but within each tumor type, fold back inversions are limited to certain cases. This hypothesis is derived from our previous genome-wide study that targeted fold-back inversions (DNA palindromes) for a set of breast tumors (Guenthoer *et al*., Genome Res 2012). In the study, DNA palindromes were abundant in about a half of primary breast tumors, suggesting that certain genetic backgrounds could be responsible for the generation of fold-back inversions.

To distinguish these possibilities, we will conduct systematic analysis of WGS, exome, RNAseq and clinical data. To do this, we employ our computer resources and expertise in cancer bioinformatics (Miyano, Niida and Shiraishi) and expertise in BFB cycles (Tanaka). Rigorous data analysis will be possible at the Institute of Medical Science, University of Tokyo. Notably, our center owns a large supercomputer system dedicated to medical research. Using the system, we have built many sequencing pipelines and bioinformatics methodologies that were successfully applied to cancer genomics data (Sato *et al*., Nat Genet 2013; Yoshida *et al*., Nat Genet 2013; Kon *et al*., Nat Genet 2013). We believe that our analysis will bring novel insights into the

mechanisms of cancer genome evolution, and contribute to development of novel diagnoses and therapies.

Timelines & resources dedicated to project

We will finish constructing a pipeline calling BFB signatures until Mar 2014. We will then start pan-cancer analysis and finish it by the end of 2014.

Research proposal

We will first determine the number of fold-back inversions in each tumor. Based on the result, we will address whether the tumors with frequent fold-back inversions (BFB signatures) are seen in certain types of tumors but not other types (hypothesis 1) or (2) whether BFB signatures are seen across all tumor types, but within each tumor type, fold back inversions are limited to certain cases (hypothesis 2). For tumors with BFB signatures, genetic background promoting BFB signatures will be defined. To do this, we will determine the association between the BFB signatures and other features, such as somatic mutations in exome data and tumor transcriptome in RNA-seq data. Furthermore, the association between clinical phenotypes and BFB signatures will be examined. With our large supercomputer system dedicated to medical research, we have a computing ability to complete these analyses in a given period of time (by the end of 2014).

1. Comprehensive detection of fold-back inversions as a signature of BFB cycles

Our pipeline screen for fold-back inversions applies in principle the criteria from Campbell et al., in which they identified a number of fold-back inversions in metastatic pancreatic tumors (Campbell, *et al*., Nature 2010). In the study, fold-back inversions were defined by read-pairs aligning close together but in inverted orientation. In WGS data, we will identify structural variants that indicate inverted duplications of genomic segments. We will further filter such structural variants by (1) the distance of two breakpoints (<20-kb) and (2) the demarcation of copy number change. The short distance between two breakpoints represents sister chromatid fusion events during BFB cycles. The demarcation distinguishes inverted duplications from small inversion polymorphisms. We will score fold-back inversions for each tumor and identify tumors with BFB signatures.

2. Genetic bakgroundss (mutations and gene expression changes) causing BFB signatures

Genetic backgrounds promoting amplification has been a long standing question (Tanaka and Yao, Nat Rev Cancer 2009). Other than the loss of TP53 function (Livingston et al., Cell 1992; Yin et al., Cell 1992), not much is known about such genetic backgrounds, mainly because systematic analyses haven't done. To address this issue, we will measure association between BFB signatures and genetic backgrounds (mutation status and expression level of each gene). We plan to do these analyses systematically in an unbiased fashion, but we also have a focused set of genes for intensive analyses, especially genes responsible for genome maintenance (genes involved in TP53 pathway, DNA replication, DNA repair and checkpoint genes). The association between TP53 status and BFB signatures would confirm the previous experimental data. Systematic analyses will bring new insights for the mechanisms of BFB cycles and gene amplification. Mutations will be extracted from VCF files and gene expression profiles will be obtained from mRNA-seq data.

3. Tumor aggressiveness and BFB signatures

BFB cycles are a major underlying mechanism of gene amplification and gene amplification drives tumor aggressiveness; however, the clinical significance of BFB signatures remains to be determined. As sample-wise association, we measure correlation between the frequency of fold-back inversions in each sample and their clinical information. We anticipate poor prognosis for the tumors with BFB signatures.

Potential problems, alternative approaches and future plans

A potential problem resides in the fact that this is the first large-scale study focusing on BFB signatures. Therefore, we may encounter issues when we distinguish tumors with BFB signatures (BFB-high tumors) from tumors without BFB signatures (BFB-low tumors). The frequency of fold-back inversion may show a continuum rather than clear distinction between high and low-type samples, although our preliminary analysis in breast tumors suggest otherwise. If that is the case, we will determine an appropriate cut-off value to distinguish BFB-high tumors from BFB-low tumors.

Successful completion of the study will provide a pipeline in which we further address the mechanisms of genome instability. We will address locus-wise association, namely association of genomic regions harboring BFB-induced amplicon and presence of sequence features, in order to obtain new insight into BFB mechanism. The sequence features to be studied will be prepared from ENCODE (e.g. epigenetic status, 3D chromatin structure, ncRNA etc), HapMap (e.g. haplotype block) and other public databases. A candidate amplified gene

is the *ERBB2* gene which is amplified in 15-20% of breast tumors and is associated with poor prognosis. We will also measure such association with other types of genomic aberration process (e.g. chromothripsis) by extending our collaboration into other groups.

Legacy plans

A pipeline calling BFB signatures will be implemented on our supercomputer and be made available for supercomputer users. Source codes will also be made available together with documentations in the format being ready for instillation on other systems. All results of pan-cancer analysis will be deposited on an SQlite database, which will be delivered publicly.

Satoru Miyano, PhD (ICGC PI)
Professor, Laboratory of DNA Information Analysis & Laboratory of Sequence Analysis
Human Genome Center, The Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan; email: miyano@ims.u-tokyo.ac.jp; Phone: +81-354495615

Degrees **BS (1977),MS (1979),PhD(1984) Dept. Mathematics, Kyushu University, Japan**

Professional Record

1979-1985: Assistant Professor, Faculty of Science, Kyushu University
1985-1987: Alexander von Humboldt Research Fellow University, U. Paderborn, Germany
1987-1993: Associate Professor, Faculty of Science, Kyushu University
1993-1996: Professor, Faculty of Science, Kyushu University
1996- : Professor, The Institute of Medical Science, The University of Tokyo

Awards

2013: Fellow of the International Society for Computational Biology

Selected Peer-Reviewed Publications

1. **Kayano M, Imoto S, Yamaguchi R,** Miyano S**. Multi-omics approach for estimating metabolic networks using low-order partial correlations.** *J Comput Biol***. 20(8):571-582, 2013.**
2. Kon A, Shih LY, Minamino M, Sanada M, Shiraishi Y, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, **Miyano S**, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
3. Sakaguchi H, Okuno Y, Muramatsu H, Yoshida K, Shiraishi Y, Takahashi M, Kon A, Sanada M, Chiba K, Tanaka H, Makishima H, Wang X, Xu Y, Doisaki S, Hama A, Nakanishi K, Takahashi Y, Yoshida N, Maciejewski JP, **Miyano S**, Ogawa S, Kojima S. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. ***Nature Genetics***. 45(8):937-941, 2013.
4. **Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M,** Miyano S**, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma.** *Nature Genetics***. 45(8):860-867, 2013.**
5. **Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S,** Miyano S**. An empirical Bayesian framework for somatic mutation detection from cancer.** *Nucleic Acids Res***. 41(7): e89, 2013.**
6. **Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Hai Nguyen H, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi S, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N,** Miyano S**, Nakagama H, Nakamura Y, Tsunoda T, Shibata T, Nakagawa H. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators.** *Nature Genetics***. 44(7):760-764, 2012.**
7. **Niida A, Imoto S, Shimamura T,** Miyano S**. Statistical model-based testing to evaluate the recurrence of genomic aberrations.** *Bioinformatics***. 28(12):i115-i120, 2012.**
8. **Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H,** Miyano S**, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia.** *Nature***. 478(7367): 64-69, 2011.**
9. **Shimamura T, Imoto S, Shimada Y, Hosono Y, Niida A, Nagasaki M, Yamaguchi R, Takahashi T,** Miyano S**. A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition.** *PLoS One***. 6(6): e20804, 2011.**
10. **Shiraishi Y, Okada-Hatakeyama M,** Miyano S**. A rank-based statistical test for measuring synergistic effects between two gene sets.** *Bioinformatics***. 27 (17): 2399-2405, 2011.**
11. **Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe G, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M,** Miyano S**, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing.** *Nature Genetics***. 42: 931–936, 2010.**
12. **Niida A, Imoto S, Yamaguchi R, Nagasaki M,** Miyano S**. Gene set-based module discovery decodes cis-regulatory codes governing diverse gene expression across human multiple tissues.** *PLoS One***. 5(6):e10910, 2010.**

Atsushi Niida, PhD (ICGC PI's Laboratory Member)
Project Assistant Professor, Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
E-mail: aniida@ims.u-tokyo.ac.jp

EDUCATIONS/TRAINING

2002-2007 Department of Biophysics and Biochemistry,
Graduate School of Science, The University of Tokyo.
2007 PhD

POSITIONS

2007-2008 Postdoctoral fellow, Department of Molecular and Genetic Information,
Institute of Molecular and Cellular Biosciences, The University of Tokyo.
2008-2011 Postdoctoral fellow, Laboratory of DNA Information Analysis,
Human Genome Center, Institute of Medical Science, The University of Tokyo.
2011-Present Project Assistant professor, Laboratory of DNA Information Analysis,
Human Genome Center, Institute of Medical Science, The University of Tokyo.

SELECTED PUBLICATIONS

1. Niida A, **Imoto S, Shimamura T, Miyano S. Statistical model-based testing to evaluate the recurrence of genomic aberrations.** *Bioinformatics*, **2012, 28:i115-i120**
2. Niida A, **Imoto S, Yamaguchi R, Nagasaki M, Fujita A, Shimamura T, Miyano S. Model-free unsupervised gene set screening based on information enrichment in expression profiles.** *Bioinformatics*, **2010, 26:3090-7**
3. Niida A, **Imoto S, Yamaguchi R, Nagasaki M, Miyano S. Gene set-based module discovery decodes cis-regulatory codes governing diverse gene expression across human multiple tissues.** *PLoS One*, **2010, 5:e10910**
4. Niida A, **Imoto S, Nagasaki M, Yamaguchi R, Miyano S. A novel meta-analysis approach of cancer transcriptomes reveals prevailing transcriptional networks in cancer cells.** *Genome Inform*, **2010, 22:121-31**
5. Niida A, **Smith AD, Imoto S, Aburatani H, Zhang MQ, Akiyama T. Gene set-based module discovery in the breast cancer transcriptome.** *BMC Bioinformatics*, **2009, 10:71**
6. Niida A, **Smith AD, Imoto S, Tsutsumi S, Aburatani H, Zhang MQ, Akiyama T. Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells.** *BMC Bioinformatics*, **2008, 9:404**

Yuichi Shiraishi, PhD (ICGC PI's Laboratory Member)
Project Assistant Professor, Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
E-mail: yshira@hgc.jp

Academic qualifications

- BSc (Engineering), Mathematical Engineering Course, Department of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, 2003.
- MSc (Information Science and Technology), Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 2005.
- PhD (Statistical Science), Department of Statistical Science, The Graduate University for Advanced Studies, 2008.

Professional Records

- Postdoctoral researcher at Cellular Systems Modeling Team, RIKEN Research Center for Allergy and Immunology, Japan (Apr. 2008–Sep. 2010).
- Postdoctoral researcher at Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan (Oct. 2010–May. 2012).
- Project Assistant Professor at Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan (Jun. 2012–).

Selected Publications

1. Kon A, Shih LY, Minamino M, Sanada M, **Shiraishi Y**, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, Miyano S, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
2. Sato Y*, Yoshizato T*, **Shiraishi Y***, Maekawa S*, Okuno Y*, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, Miyano S, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma. ***Nature Genetics***. 45(8):860-867, 2013. (* equally contributed).
3. **Shiraishi Y**, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, Miyano S. An empirical Bayesian framework for somatic mutation detection from cancer. ***Nucleic Acids Res***. 41(7): e89, 2013.
4. **Yoshida K*, Sanada M*,** Shiraishi Y***, Nowak D*, Nagata Y*, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H, Miyano S, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia.** *Nature***. 478(7367): 64-69, 2011. (*equally contributed).**
5. **Shiraishi Y**, Okada-Hatakeyama M, Miyano S. A rank-based statistical test for measuring synergistic effects between two gene sets. ***Bioinformatics***. 27 (17): 2399-2405, 2011.
6. Shiraishi Y**, Kimura S, Okada M. Inferring cluster-based networks from differently stimulated multiple time-course gene expression data.** *Bioinformatics***. 26(8):1073-81, 2010.**

Hisashi Tanaka MD. PhD (Non-ICGC/TCGA Collaborator)

Assistant Professor,
Department of Molecular Medicine
Cleveland Clinic Lerner College of Medicine of Case Western Reserve University
9500 Euclid Ave. NE20, Cleveland OH 44195, USA

Degrees **1988 MD Kyoto University School of Medicine, Japan**
1997 PhD Kyoto University Graduate School of Medicine, Japan (Surgical Oncology)

Positions

1988-1989	Resident in Surgery, Kyoto University Hospital, Kyoto, Japan.
1989-1993	Medical staff in Surgery, Fukui Red Cross Hospital, Fukui, Japan.
1994-1997	Graduate Student, Department of Surgical Oncology, Kyoto University Graduate School of Medicine, Japan. Research Trainee, Aichi Cancer Center Research Institute, Japan.
1998-2003	Postdoctoral Fellow, Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA.
2003-2006	Staff Scientist, Fred Hutchinson Cancer Research Center, Seattle, WA.
2006-present	Assistant Professor, Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland OH

Selected Peer-Reviewed Publications

1. Tanaka, H., **Shibagaki, I., Shimada, Y., Wagata, T., Imamura, M. and Ishizaki, K., (1996). Characterization of the p53 gene mutations in esophageal squamous cell carcinoma cell lines: Increased frequency and different spectrum of mutations from primary tumors.** *Int. J. Cancer*, **65, 372-376.**
2. Tanaka, H., **Shimada, Y., Imamura, M., Shibagaki, I. and Ishizaki, K. (1997). Multiple types of aberrations in the p16 (INK4a) and p15 (INK4b) genes in 30 esophageal squamous cell carcinoma cell lines.** *Int. J.Cancer*, **70, 437-442.**
3. Tanaka, H., **Shimada, Y., Harada, H., Shinoda, M., Hatooka, S., Imamura, M. and Ishizaki, K. (1998). Methylation of the 5′ CpG island of the FHIT gene is closely associated with transcriptional inactivation in esophageal squamous cell carcinomas.** *Cancer Res*., **58: 3429-3434.**
4. **Harada, H.,** Tanaka, H., **Shimada, Y., Shinoda, M., Imamura, M. and Ishizaki, K. (1999). Lymph node metastasis is associated with allelic loss on chromosome 13q12-13 in esophageal squamous cell carcinoma.** *Cancer Res.* **59, 3724-9.**
5. **Tanaka. H.**, Shimada. Y., Harada. H., Shinoda. M., Hatooka. S., Imamura. M. and Ishizaki K. (2000). Polymorphic variation of the ARP gene on 3p21 in Japanese esophageal cancer patients. ***Oncol Rep***. 7(3):591-3.
6. **Tanaka. H.**, Tapscott, S.J., Trask, B.J. and Yao, M.C. (2002). Short Inverted repeats initiate gene amplification through the formation of large DNA palindrome in mammalian cells. ***Proc. Natl. Acad. Sci. USA*** 99, 8772-7.
7. **Tanaka, H.**, Bergstrom, D. A., Yao, M. C. and Tapscott, S. J. (2005). Widespread and non-random distribution of DNA palindromes provides a structural platform for subsequent gene amplification. ***Nat. Genet.*** 37 320-7.
8. **Zhao, Y., Marotta, M., Eichler, E.E., Eng, C. and** Tanaka, H. **(2009). Linkage disequilibrium between two high-frequency deletion polymorphisms: implications for association studies involving the glutathione-S transferase (GST) genes.** *PLoS Genet*. **5, e1000472. PMCID: PMC2672168**
9. **Tanaka, H.*** and Yao, M. C. (2009). Palindromic gene amplification – an evolutionary conserved role for DNA inverted repeats in the genome. ***Nat. Rev. Cancer***, 9, 216-224. PMID: 19212324 (* corresponding author)
10. **Diede, S.J., Guenthoer, J., Geng, L.N., Mahoney, S.E., Marotta, M., Olson, J.M.,** Tanaka, H., **Tapscott, S.J. (2010) DNA methylation of developmental genes in pediatric medulloblastomas identified by Denaturation Analysis of Methylation Differences.** *Proc. Natl. Acad. Sci. USA*, **doi: 10.1073/pnas.0907606106. PMCID: PMC2806770**
11. **Guenthoer J, Diede S.J,** Tanaka, H., **Chai X, Hsu L, Tapscott S.J, Porter P.L. (2011). Assessment of palindromes as platforms for DNA amplification in breast cancer.** *Genome Res*, **PMID: 21752925**
12. **Marotta, M., Piontokivska, H., and** Tanaka, H. **(2012) Molecular trajectories leading to the alternative fates of duplicated genes.** *PLoS ONE* **7(6): e38958.**
13. **Marotta, M., Chen, X., Inoshita, A., Stephens, R., Budd, T.G., Crowe, J., Lyones, J., Kondratova, A., Tubbs, R. and** Tanaka, H. **(2012) A common copy number breakpoint of *ERBB2* amplification in breast cancer co-localizes with a complex block of segmental duplications.** *Breast Cancer Res* **14, R150.**
14. **Marotta, M., Chen, X., Watanabe, T., Faber. P.W., Diede, S.J., Kondratova, A., Tapscott, S.J. Tubbs, R., Stephens, R. and** Tanaka, H. **(2013) Homology-dependent end-capping as a primary step of sister chromatid fusion for the Breakage-Fusion-Bridge cycles.** *Nucleic Acids Research* **doi: 10.1093/nar/gkt762**

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

A novel statistical method for detecting somatic genomic mutations causing splicing aberrations and its application to pan cancer genomic and transcriptome sequencing data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Satoru Miyano, Human Genome Center, The Institute of Medical Science, The University of Tokyo

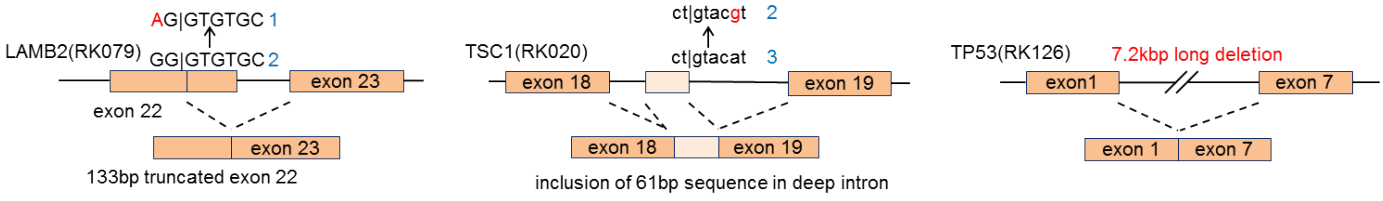
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yuichi Shiraishi, Human Genome Center, The Institute of Medical Science, The University of Tokyo	

Background and preliminary data

Defects in mRNA splicing are an important cause of cancer development, and many studies over the last 20 years have reported a number of cancer-specific alternative splicing caused by somatic mutations (Venables, Cancer Research, 2004; David et al., Genes Dev., 2010). Furthermore, recent advances in high-throughput sequencing technologies have enabled us to perform genome wide screenings of genomic mutations causing splicing aberrations (GMSAs) by comparing genome and transcriptome sequencing data. However, most current studies focus on those disrupting essential splice-sites (splice acceptor and donor sites located at the first and last two bases of an intron sequence), and other types of somatic mutations that can cause splicing aberrations have been largely ignored.

As a preliminary study, we have performed comparative analysis of 22 whole genome and whole transcriptome sequencing of HBV-related liver cancers (Shiraishi et al., in submission), and identified a number of examples that not only essential splice-site mutations, but also silent mutations in coding regions, deep intronic mutations and structural changes can cause splicing aberrations such as splice-site slipping, exon skipping and pseudo-exon inclusion (see Figures below). Notably they often occur in tumor suppressor genes such as AXIN1, RB1, TSC1, TP53. Therefore, in our view, a number of various GMSAs contributing cancer development will be found by genome wide screening of GMSAs across a broad spectrum of human cancers.

In this study, after developing a novel statistical method for sensitively and accurately detecting GMSAs, we collect a comprehensive catalogue of GMSAs from pan-cancer sequencing data to investigate the frequencies of oncogenic GMSAs and to understand novel mechanisms of splicing machinery.



Timelines & resources dedicated to project

Timelines

- Develop a novel statistical method for detecting GMSAs and evaluate it using whole genome and whole transcriptome sequencing data from liver cancers (by June. 2014).
- Perform the above method to obtain a comprehensive catalogue of somatic GMSAs and give interpretations on it (by August. 2014).
- Construct a classifier for predicting whether the given genomic mutations splicing aberrations using the above catalogue of somatic GMSAs as a learning set based on machine learning theory (by December 2014).

Resources dedicated to project

- Matched tumor and normal whole genome / exome sequencing and RNA sequencing pairs.

Research proposal

1. Development of new statistical method for sensitive detection of genomic mutations causing splicing aberrations

The approach we have used in the preliminary study was somewhat ad-hoc (e.g., a number of threshold values determined without thorough theoretical background). Thus, we will first develop a new approach based on rigorous statistical theory for detecting GMSAs using whole genome / exome and whole transcriptome sequencing data.

One possible approach, which we adopted in the preliminary study, is to independently characterizing genomic changes such as somatic point substitutions, indels and structural variations from genome sequencing data, and cancer specific splicing aberrations from transcriptome sequencing data. However, by complementary use of genome and transcriptome sequencing data, we would rescue more combinations of genomic mutations and associated transcriptional aberrations that narrowly miss the criteria for being called by single analysis. Thus, to improve the sensitivity, we will consider a statistical framework for simultaneously dealing with genome and transcriptome sequencing data.

For evaluating accuracies of our methods, we will use more than 250 whole genome and whole transcriptome sequencing data collected from liver cancers in Japanese ICGC project. In this data set, matched-control whole transciptome sequencing data is available for most samples whereas it is not because of sequencing cost and poor sample qualities in many studies, which makes it very difficult to characterize cancer-specific splicing aberrations. Therefore, this liver cancer data set will be a good resource for investigating the differences in the sensitivities of our approach in cases when matched whole transcriptome sequencing data is (1) available for most samples, (2) available only for several samples, and (3) not available.

Although we have concentrated on genomic mutations disrupting splicing acceptor / donor motifs, we will also try to find splicing aberrations caused by disruption of other splicing-related motif sequences, such as exonic splicing enhancer / silencer, intronic splicing enhancer / silencer, and splicing branch points.

Actually, it is difficult to prove the direct causality between genomic mutations and the corresponding splicing aberrations. Therefore, for each type of GMSAs (e.g., somatic mutations creating sudden splicing motifs in intronic regions causing peuso-exon inclusions, long deletions causing exon skippings, and so on) we will give theoretical evaluations on false discovery rate for their actual causalities using statistical approaches such as permutation tests and Bayesian statistical theory.

2. Acquisition of catalogue of genomic mutations causing splicing aberrations

First, we will identify GMSAs in each cancer type using the newly developed statistical method to acquire the catalogue of GMSAs. Frequencies and positional distributions of GMSAs, especially those in intronic regions, will be important guideline for clinical sequencing design.

Investigating the effects on cancer pathogenesis of each GMSA is an important issue. For this purpose, first, we will seek for recurrent splicing aberrations in a specific gene in some cancer types that can lead new therapeutics or novel understandings of cancer biology. Furthermore, we will examine whether detected GMSAs significantly enriches known or putative (significantly mutated) cancer driver genes, copy number status and so on.

3. Elucidation of genomic features contributing splicing machinery based on machine learning theory

The above catalogue of GMSAs will be a great source for understanding the mechanisms of splicing functions. In this study, we will construct a machine-learning-based classifier predicting whether given genomic mutations drive splicing aberrations utilizing the catalogue of GMSAs as training data set. By investigating the genomic features (e.g., ENCODE data such as nucleosome positioning, histone modifications, ChIP-Seq data, and so on) contributing to classification accuracy, we try to elucidate the general mechanism of splicing machinery. Furthermore, the obtained classifier will be of great importance in further cancer genomics or clinical sequencing studies especially when RNAs of the patients are not available.

Legacy plans

The current algorithm for detecting genomic mutations causing splicing aberrations from whole genome and whole transcriptome sequencing data is described in the reference below. The implementation of new statistical method that will be developed in this study will be publically available. In this study, we plan to use our internal method for detecting somatic mutations (https://github.com/friend1ws/EBCall, Shiraishi et al., Nucleic Acids Research, 2013), and we will further brush up the sensitivity and accuracy of this software and perform optimization so that this software can efficiently handle a number of whole genome sequencing data.

Furthermore, the catalogue of GMSAs will be publically available to the research community.

Reference

Yuichi Shiraishi et al., “Integrated analysis of whole genome and transcriptome sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in 22 HBV-related liver cancers”, in submission.

Satoru Miyano, PhD (ICGC PI)

Professor, Laboratory of DNA Information Analysis & Laboratory of Sequence Analysis
Human Genome Center, The Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan; email: miyano@ims.u-tokyo.ac.jp; Phone: +81-354495615

Degrees BS (1977),MS (1979),PhD(1984) Dept. Mathematics, Kyushu University, Japan

Professional Record

1979-1985: Assistant Professor, Faculty of Science, Kyushu University
1985-1987: Alexander von Humboldt Research Fellow University, U. Paderborn, Germany
1987-1993: Associate Professor, Faculty of Science, Kyushu University
1993-1996: Professor, Faculty of Science, Kyushu University
1996- : Professor, The Institute of Medical Science, The University of Tokyo

Awards

2013: Fellow of the International Society for Computational Biology

Selected Peer-Reviewed Publications

1. Kayano M, Imoto S, Yamaguchi R, **Miyano S**. Multi-omics approach for estimating metabolic networks using low-order partial correlations. ***J Comput Biol***. 20(8):571-582, 2013.
2. Kon A, Shih LY, Minamino M, Sanada M, Shiraishi Y, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, **Miyano S**, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
3. Sakaguchi H, Okuno Y, Muramatsu H, Yoshida K, Shiraishi Y, Takahashi M, Kon A, Sanada M, Chiba K, Tanaka H, Makishima H, Wang X, Xu Y, Doisaki S, Hama A, Nakanishi K, Takahashi Y, Yoshida N, Maciejewski JP, **Miyano S**, Ogawa S, Kojima S. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. ***Nature Genetics***. 45(8):937-941, 2013.
4. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, **Miyano S**, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma. ***Nature Genetics***. 45(8):860-867, 2013.
5. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, **Miyano S**. An empirical Bayesian framework for somatic mutation detection from cancer. ***Nucleic Acids Res***. 41(7): e89, 2013.
6. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Hai Nguyen H, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi S, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, **Miyano S**, Nakagama H, Nakamura Y, Tsunoda T, Shibata T, Nakagawa H. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. ***Nature Genetics***. 44(7):760-764, 2012.
7. Niida A, Imoto S, Shimamura T, **Miyano S**. Statistical model-based testing to evaluate the recurrence of genomic aberrations. ***Bioinformatics***. 28(12):i115-i120, 2012.
8. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H, **Miyano S**, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia. ***Nature***. 478(7367): 64-69, 2011.
9. Shimamura T, Imoto S, Shimada Y, Hosono Y, Niida A, Nagasaki M, Yamaguchi R, Takahashi T, **Miyano S**. A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition. ***PLoS One***. 6(6): e20804, 2011.
10. Shiraishi Y, Okada-Hatakeyama M, **Miyano S**. A rank-based statistical test for measuring synergistic effects between two gene sets. ***Bioinformatics***. 27 (17): 2399-2405, 2011.
11. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe G, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, **Miyano S**, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. ***Nature Genetics***. 42: 931–936, 2010.

Yuichi Shiraishi, PhD (ICGC PI's Laboratory Member)
Project Assistant Professor, Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
E-mail: yshira@hgc.jp

Academic qualifications

- BSc (Engineering), Mathematical Engineering Course, Department of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, 2003.
- MSc (Information Science and Technology), Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 2005.
- PhD (Statistical Science), Department of Statistical Science, The Graduate University for Advanced Studies, 2008.

Professional Records

- Postdoctoral researcher at Cellular Systems Modeling Team, RIKEN Research Center for Allergy and Immunology, Japan (Apr. 2008–Sep. 2010).
- Postdoctoral researcher at Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan (Oct. 2010–May. 2012).
- Project Assistant Professor at Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan (Jun. 2012–).

Selected Publications

1. Kon A, Shih LY, Minamino M, Sanada M, **Shiraishi Y**, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, Miyano S, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
2. Sato Y*, Yoshizato T*, **Shiraishi Y***, Maekawa S*, Okuno Y*, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, Miyano S, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma. ***Nature Genetics***. 45(8):860-867, 2013. (* equally contributed).
3. **Shiraishi Y**, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, Miyano S. An empirical Bayesian framework for somatic mutation detection from cancer. ***Nucleic Acids Res***. 41(7): e89, 2013.
4. Yoshida K*, Sanada M*, **Shiraishi Y***, Nowak D*, Nagata Y*, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H, Miyano S, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia. ***Nature***. 478(7367): 64-69, 2011. (*equally contributed).
5. **Shiraishi Y**, Okada-Hatakeyama M, Miyano S. A rank-based statistical test for measuring synergistic effects between two gene sets. ***Bioinformatics***. 27 (17): 2399-2405, 2011.
6. **Shiraishi Y**, Kimura S, Okada M. Inferring cluster-based networks from differently stimulated multiple time-course gene expression data. ***Bioinformatics***. 26(8):1073-81, 2010.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Acquisition of the catalogue of somatic ITDs

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Satoru Miyano, Human Genome Center, The Institute of Medical Science, The University of Tokyo, (ICGC PI)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yuichi Shiraishi (Human Genome Center, The Institute of Medical Science, The University of Tokyo)	Seishi Ogawa (Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University)

Background and preliminary data

It is widely known that an internal tandem duplication (ITD) that involves several tens to several hundreds of nucleotides represents a common type of somatic alteration and plays an important role in cancer pathogenesis. In particular, ITDs involving *FLT3* (FMS-like tyrosine kinase 3) are among the most frequent genetic lesions found in ~30% of patients with acute myeloid leukemia (AML), leading to aberrant activation of the kinase and negatively affecting their survival. However, despite their potential importance in cancer development, ITDs have been poorly focused in cancer genome sequencing studies because most current approaches are not sensitive enough to identify indels of more than several tens of base pairs (bp) or structural variations within small-scale regions (less than 1000 bp), rendering ITDs being “blind spots” for most existing analytical methods.

We have recently developed “Genomon ITDetector” for sensitively and accurately detecting ITDs genome-wide, and demonstrated that it can successfully detect not only common ITDs involving *FLT3*, but also a number of ITDs affecting putative driver genes such as *KIT*, *CEBPA*, *MLL*, *WT1* and *NRAS* in acute myeloid leukemia samples (Chiba et al., in submission).

Although the ITDs have been studied in hematology, only a few studies have been made on solid tumors. Therefore, as a preliminary study, we have performed this software on cohorts of several cancer types (colon and rectal carcinomas, lung adenocarcinomas, and so on), finding ITDs in putative cancer genes such as *APC* and *PIK3CA*. Notably, the ITD affecting *PIK3CA* was in-frame and duplicating nucleotides of a kinase domain suggesting a gain-of-function property of this mutation. Therefore, it is highly expectable that a number of cancer driver ITDs will be detected in pan cancer analysis. Based on this background, we need to collect a comprehensive catalogue of ITDs from pan-cancer sequencing data to investigate the frequencies of oncogenic ITDs and their possibilities as therapeutic target.

Timelines & resources dedicated to project

- Improve the algorithm evaluating the accuracy and sensitivity using real sequencing data in more detail (by Feb. 2014).
- Perform ITD detection on whole genome, whole exome and RNA sequencing data to obtain a comprehensive catalogue of somatic ITDs (by May. 2014).
- Investigate functions on cancer pathogenesis of each somatic ITD by checking their transcriptional consequences, relationships with copy number alterations and distributions of somatic substitutions. (by Aug. 2014).
- Investigate the generative mechanism of somatic ITDs by checking the surrounding sequences and relationships with mutational signatures (Nik-Zainal et al., Cell, 201, Alexandrov et al., Nature, 2013) (by Nov. 2014).

Research proposal

We will work on following three topics.

1. Generating a catalog of somatic ITDs.

We will perform ITD detection for all whole genome, exome and whole transcriptome sequencing data of whole cancer types, to obtain a comprehensive catalogue of somatic ITDs with the detailed evaluation of the sensitivity and accuracy. First, we will examine genes and pathways with significant frequencies for each cancer types. Then, we seek the possibility of somatic ITDs as drug targets. It would be wonderful if we could detect highly recurrent somatic ITDs in some cancer types that can lead new therapeutics or novel understandings of cancer biology like *FLT3* in AML. Currently, we are not sure whether we could encounter such an ideal situation. However, the preliminary analysis found at least several ITDs affecting genes with kinase activities, which will be very important in personalized clinical sequencing even if their frequencies are not so high. In addition, we will investigate the relationships between the number of somatic point substitutions and ITDs for whole cancer types, to check whether occurrence ratios of somatic ITDs vary depending on cancer types. These results will be a good guideline for deciding how to incorporate somatic ITD detection in future cancer genome sequencing studies and clinical sequencing.

2. Prediction of functional effect of each somatic ITD

Investigating functions on cancer pathogenesis of each somatic ITD is an important issue. Since, our preliminary analysis implies that most somatic ITDs are sporadic, one useful approach for predicting the deleteriousness of each somatic ITD is to correlate with other information such as (1) whether the lengths of the duplicated sequences are multiples of three (in-frame) or not, (2) transcriptional consequences of somatic ITDs (mutant transcripts are expressing or not), (3) co-occurrence with copy number alterations, (4) the distributions of somatic point substitutions of affected genes (e.g., duplicating hot-spot mutation sites or not). After performing detailed comparative analysis between the catalog of somatic ITDs and the above features, we will try to develop a framework for accurately classifying the somatic ITDs into gain-of-function, loss-of-function and passenger.

3. Investigating the mechanism of ITD generation

One of the merits of performing ITD detection using whole genome sequencing data is that we can collect a number of somatic ITDs, possibly enabling us to identify characteristic patterns of somatic ITDs that reflect the DNA damage and repair processes to which their specimens are exposed. We will try to classify the detected ITDs based on their surrounding sequences or genomic functions, and examine the frequencies and variations of characteristic patterns among pan cancer samples. Furthermore, we will investigate the relationships between the somatic ITDs patterns and known mutational signatures of somatic substitutions (Alexandrov et al., Nature, 2013).

Legacy plans

The software we have developed for detecting internal tandem duplication is already available at https://github.com/ken0-1n/Genomon-ITDetector. In the course of this pan-cancer analysis project, we will further brush up the sensitivity and accuracy of this software and perform optimization so that this software can efficiently handle a number of whole genome sequencing data.

Furthermore, the catalogue of somatic ITDs with inferred deleteriousness will be delivered to the research community.

Satoru Miyano, PhD (ICGC PI)

Professor, Laboratory of DNA Information Analysis & Laboratory of Sequence Analysis
Human Genome Center, The Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan; email: miyano@ims.u-tokyo.ac.jp; Phone: +81-354495615

Degrees BS (1977),MS (1979),PhD(1984) Dept. Mathematics, Kyushu University, Japan

Professional Record

1979-1985: Assistant Professor, Faculty of Science, Kyushu University
1985-1987: Alexander von Humboldt Research Fellow University, U. Paderborn, Germany
1987-1993: Associate Professor, Faculty of Science, Kyushu University
1993-1996: Professor, Faculty of Science, Kyushu University
1996- : Professor, The Institute of Medical Science, The University of Tokyo

Awards

2013: Fellow of the International Society for Computational Biology

Selected Peer-Reviewed Publications

1. Kayano M, Imoto S, Yamaguchi R, **Miyano S**. Multi-omics approach for estimating metabolic networks using low-order partial correlations. ***J Comput Biol***. 20(8):571-582, 2013.
2. Kon A, Shih LY, Minamino M, Sanada M, Shiraishi Y, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, **Miyano S**, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
3. Sakaguchi H, Okuno Y, Muramatsu H, Yoshida K, Shiraishi Y, Takahashi M, Kon A, Sanada M, Chiba K, Tanaka H, Makishima H, Wang X, Xu Y, Doisaki S, Hama A, Nakanishi K, Takahashi Y, Yoshida N, Maciejewski JP, **Miyano S**, Ogawa S, Kojima S. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. ***Nature Genetics***. 45(8):937-941, 2013.
4. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, **Miyano S**, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma. ***Nature Genetics***. 45(8):860-867, 2013.
5. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, **Miyano S**. An empirical Bayesian framework for somatic mutation detection from cancer. ***Nucleic Acids Res***. 41(7): e89, 2013.
6. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Hai Nguyen H, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi S, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, **Miyano S**, Nakagama H, Nakamura Y, Tsunoda T, Shibata T, Nakagawa H. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. ***Nature Genetics***. 44(7):760-764, 2012.
7. Niida A, Imoto S, Shimamura T, **Miyano S**. Statistical model-based testing to evaluate the recurrence of genomic aberrations. ***Bioinformatics***. 28(12):i115-i120, 2012.
8. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H, **Miyano S**, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia. ***Nature***. 478(7367): 64-69, 2011.
9. Shimamura T, Imoto S, Shimada Y, Hosono Y, Niida A, Nagasaki M, Yamaguchi R, Takahashi T, **Miyano S**. A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition. ***PLoS One***. 6(6): e20804, 2011.
10. Shiraishi Y, Okada-Hatakeyama M, **Miyano S**. A rank-based statistical test for measuring synergistic effects between two gene sets. ***Bioinformatics***. 27 (17): 2399-2405, 2011.
11. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe G, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, **Miyano S**, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. ***Nature Genetics***. 42: 931–936, 2010.

Yuichi Shiraishi, PhD (ICGC PI's Laboratory Member)
Project Assistant Professor, Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
E-mail: yshira@hgc.jp

Academic qualifications

- BSc (Engineering), Mathematical Engineering Course, Department of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, 2003.
- MSc (Information Science and Technology), Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 2005.
- PhD (Statistical Science), Department of Statistical Science, The Graduate University for Advanced Studies, 2008.

Professional Records

- Postdoctoral researcher at Cellular Systems Modeling Team, RIKEN Research Center for Allergy and Immunology, Japan (Apr. 2008–Sep. 2010).
- Postdoctoral researcher at Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan (Oct. 2010–May. 2012).
- Project Assistant Professor at Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan (Jun. 2012–).

Selected Publications

1. Kon A, Shih LY, Minamino M, Sanada M, **Shiraishi Y**, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, Miyano S, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
2. Sato Y*, Yoshizato T*, **Shiraishi Y***, Maekawa S*, Okuno Y*, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, Miyano S, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma. ***Nature Genetics***. 45(8):860-867, 2013. (* equally contributed).
3. **Shiraishi Y**, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, Miyano S. An empirical Bayesian framework for somatic mutation detection from cancer. ***Nucleic Acids Res***. 41(7): e89, 2013.
4. Yoshida K*, Sanada M*, **Shiraishi Y***, Nowak D*, Nagata Y*, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H, Miyano S, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia. ***Nature***. 478(7367): 64-69, 2011. (*equally contributed).
5. **Shiraishi Y**, Okada-Hatakeyama M, Miyano S. A rank-based statistical test for measuring synergistic effects between two gene sets. ***Bioinformatics***. 27 (17): 2399-2405, 2011.
6. **Shiraishi Y**, Kimura S, Okada M. Inferring cluster-based networks from differently stimulated multiple time-course gene expression data. ***Bioinformatics***. 26(8):1073-81, 2010.

	Name	Seishi Ogawa		
	Date of Birth	Aug. 12, 1962	Age	51
	Institution (University, College, etc.), Academic Unit (School, Faculty, etc.) & Position	Professor, Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University		
	Academic Degree	M.D., Ph.D.		
	Field of Specialization	Hematology/Oncology and Molecular Genetics		

Research Careers and Experience

2013 – Present : Professor, Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University

2008 – 2013 : Associate Professor, Cancer Genomics Project, University of Tokyo

Genetic analysis of human cancers especially focused on myelodysplastic syndromes and other myeloid neoplasms, using advanced genomics including massively parallel sequencing.

Discovery of RNA splicing factor mutations and other novel genetic alterations in myelodysplasia.

2006 – 2008 : Associate Professor, The 21st century COE program, Graduate School of Medicine, University of Tokyo

Genetic analysis of human cancers especially, using SNP array-based copy number analysis. Studies on *CBL* mutations in MDS, *A20* mutations in lymphomas, *ALK* mutations in neuroblastoma.

2002 – 2006 : Associate Professor, Department of Regeneration Medicine for Hematopoiesis, Graduate School of Medicine, University of Tokyo

Genome-wide association study on graft-versu-host disease.

Development of tools for genome-wide copy number analysis using Affymetrix SNP genotyping microarray (CNAG) (Nannya et al., Cancer Res. 2005.)

1996 –2002 : Assistant Professor of Medicine, University of Tokyo

Deletion mapping and identification of tumor suppressor genes on chromosome 6q in lymphoid neoplasms

1995 –1996 : Research fellow of Japan Society for the Promotion of Science

Studies on activated Evi-1 oncogene in myeloid neoplasms.

1994 –1995 : Clinical Associate, University of Tokyo

Studies on inactivation of the *CDKN2A(p16)* gene in hematopoietic neoplasms.

1994 –1995 : M.D., Ph.D. degree at the Graduate School of Medicine, University of Tokyo

Title of the thesis: "The C-terminal SH3 domain of the mouse c-Crk protein negatively regulates tyrosine-phosphorylation of Crk associated p130 in rat 3Y1 cells."

1988 –1989 : Postgraduate clinical training in internal medicine

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 31st December, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Genomic approach to cancer immunoediting in human through pan-cancer analysis

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Satoru Miyano, Human Genome Center, The Institute of Medical Science, The University of Tokyo, (ICGC PI)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Rui Yamaguchi, Seiya Imoto (Human Genome Center, The Institute of Medical Science, The University of Tokyo)	Shinichi Mizuno (Division of Cancer Research, Center for Advanced Medical Innovation, Kyushu University)

Background and preliminary data

In tumor development, cancer immunoediting is considered to be a significant process in which the immune system controls the growth of nascent tumor cells and also shapes their immunogenic phenotypes. Recent murine studies revealed that T-cell recognition of tumor-specific antigens derived from its missense mutations is critical for the process of immunoediting; however, the presence and degree of cancer immunoediting in human remain controversial. Although it is important to evaluate the roles of tumor-immune interactions in cancer development, it seems difficult to assess the immunoediting in human, since human tumors are fully developed and are thought to be edited already at the time of resection. In this project, we will apply a novel genomic approach to estimate the cancer immunoediting in human.

In the process of immunoediting, tumor cells that possess mutated proteins presented by HLA Class I molecules will be eradicated by T-cells, however, tumor cells with mutated proteins which do not bind to HLAs will survive and progress to the malignant stage. Thus, through the process of tumor progression, missense mutations that do not bind to HLAs will be prone to accumulate in genome. In other word, the number of missense mutations whose mutated proteins can bind to HLA molecules will be reduced. Therefore, the process of T-cell mediated cancer immunoediting will leave traces in genome as a set of selected mutations. In the case where non-specific immune system, such as natural killer cells and/or macrophages, are dominant, traces of immunoediting in genome will be less clear. These traces of cancer immunoediting in genome will be figured out by the analysis of mutated peptides for binding affinity to HLAs, namely, by the analysis of binding affinity to the tumor-own HLAs compared with binding affinity to the unrelated HLAs.

We performed a preliminary analysis, and we could find out a trace of cancer immunoediting in genome by using 19 cases of our exome sequencing data of renal cell carcinoma (RCC) (Sato et al; Nat Genet 2013). After HLA genotyping of tumors, we submitted a library of peptide sequences corresponding to missense mutations of each tumor to NetMHCpan for prediction of binding affinity of peptides to HLA Class I molecules. In the group of RCC with HLA-A*24:02, the number of missense mutations predicted to bind to A*24:02 was calculated to be less than that of mutations in other group not having A*24:02, that is, without pressure of immune selection through A*24:02. Thus, we could estimate cancer immunoediting by genomic analysis. We plan to extend this approach to various types of tumor through the project of pan-cancer analysis.

To elucidate the presence and degree of immunoediting in tumors, we will conduct systematic analysis of WGS, exome and RNA-seq data. To do this, we will employ our computer resources and expertise in cancer bioinformatics (Miyano, Yamaguchi and Imoto) and expertise in cancer immunology (Mizuno). Rigorous data analysis will be possible at the Institute of Medical Science, the University of Tokyo. Notably, our center owns a large supercomputer system dedicated to medical research. Using the system, we have built many sequencing pipelines and bioinformatics methodologies that were successfully applied to cancer genomics data (Sato et al., Nat Genet 2013; Yoshida et al., Nat Genet 2013; Kon et al., Nat Genet 2013). We believe that our analysis will clarify the roles of cancer immunoediting in human and the mechanisms of cancer genome evolution “sculpted” by the immune system, and will contribute to the development of novel immunotherapies.

Timelines & resources dedicated to project

We will finish constructing a pipeline estimating traces of cancer immunoediting in genome until April 2014. We will then start pan-cancer analysis and finish it by the end of 2014.

Research proposal

We will first determine HLA genotypes of tumors and predict HLA-binding affinities of mutated proteins derived from missense mutations. Based on these analyses, we will start to evaluate the presence and degree of cancer immunoediting in tumors with high immunogenesity such as renal cell carcinoma and malignant melanoma. Then, we will move on to the analysis of major types of cancers showing less immunogenesity in order to elucidate the differences in immune response among tumor types by genomic analysis. With our large supercomputer system dedicated to medical research, we have a computing ability to complete these analyses by the end of 2014.

1. HLA genotyping of tumors

WGS and exome date from pan-cancer analysis will be subjected to HLA genotyping, and paired-end reads will be aligned to the HLA locus of UCSC hg19. We will extract informative reads harboring SNVs on both forward and reverse reads and generate a set of heterozygous HLA genes. The defined HLA gene sequences are subjected to HLA allele determination by searching the IMGT/HLA database. We will deal with pairs of matched tumor and normal data, in order to exclude tumors affecting deletion of HLA genes in advance.

2. Prediction of HLA binding affinity of mutated peptides derived from missense mutations

For prediction of HLA binding of peptides, we will prepare a library of peptide sequences corresponding to missense mutations flanked by up to 10 amino acids on either side in each case. We will submit each library of peptide sequences to NetMHCpan 2.8 for prediction of binding affinity to HLA Class I molecules. We will focus on the HLA subtypes present more than 5% in subject group of tumors for statistical analysis. In addition, to distinguish self-reactive peptide sequences, we also count affinity values of wild type peptides as reference.

3. Assessment of the presence and degree of cancer immunoediting

In the process of cancer immunoediting, T-cell mediated immunoediting will leave traces in genome with the reduction of the number of missense mutations whose mutated peptides can bind to HLA molecules. Therefore, cancer immunoediting in genome will be unraveled by high-throughput analysis of mutated peptides for binding affinity to the pre-designed sets of HLAs. In this study, we will undertake the prediction of binding affinity of mutated peptides, (1) to the tumor-own HLAs, and (2) to the HLAs unrelated to the tumor as to set the neutral values of without pressure of immune selection. Then, missense mutations whose peptides can bind to each HLA are selected by the setting values of binding affinity, and tumors will be sorted into groups according to their HLA subtypes. Finally, we will compare the number of extracted mutations between defined groups and the difference between groups will be statistically evaluated.

4. Types of tumors and cancer immunoediting

Immunodeficiency has been reported to increase cancer risk in patients with AIDS and in transplant recipients under immunosuppression, however, the spectrum of tumor types of immunosuppressed patients is quite different from that of the general population, suggesting the notion that the process of cancer immunoediting may vary among tumor types. This notion will be quite important when we provide immunotherapies to cancer patients. To elucidate the differences in immune response among tumor types, we will make an analysis of major types of cancers showing less immunogenesity, after the analysis of tumors with high immunogenesity such as renal cell carcinoma and malignant melanoma.

Potential problems, alternative approaches and future plans

In this project, each step of the analysis is well established, however, potential problem could be on the accuracy of prediction of peptide binding to HLAs. We are preparing an alternative plan for estimating immunoediting by analysis of frameshift mutation (fm). A coding sequence from fm to premature stop codon will make a novel peptides chain. Since a part of a novel protein can be a target of T-cells, it is postulated that the longer the peptides chain stretches, the higher the incidence of T-cell recognition will be. Therefore, cancer immunoediting will be prone to eliminate fm's with generating longer peptides and tend to preserve fm's with producing shorter peptides. The computational analysis of the distribution of lengths of peptides from fm's to premature stop codon will represent an outcome of cancer immunoediting in human.

Overall, this project will provide us with novel insights about cancer immunoediting in human, and will be useful to shape the strategy for cancer immunotherapies as well as to clarify the mechanism of cancer genome evolution ("sculpting") by immune selection.

Legacy plans

A pipeline estimating traces of immunoediting in genome will be implemented on our supercomputer and be made available for supercomputer users. The pipeline structure will also be made available together with documentations in the format being ready for instillation on other systems. All results of pan-cancer analysis will be deposited on an SQlite database, which will be delivered publicly.

Satoru Miyano, PhD (ICGC PI)

Professor, Laboratory of DNA Information Analysis & Laboratory of Sequence Analysis
Human Genome Center, The Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan; email: miyano@ims.u-tokyo.ac.jp; Phone: +81-354495615

Degrees **BS (1977),MS (1979),PhD(1984) Dept. Mathematics, Kyushu University, Japan**

Professional Record

1979-1985: Assistant Professor, Faculty of Science, Kyushu University
1985-1987: Alexander von Humboldt Research Fellow University, U. Paderborn, Germany
1987-1993: Associate Professor, Faculty of Science, Kyushu University
1993-1996: Professor, Faculty of Science, Kyushu University
1996- : Professor, The Institute of Medical Science, The University of Tokyo

Awards

2013: Fellow of the International Society for Computational Biology

Selected Peer-Reviewed Publications

1. **Kayano M, Imoto S, Yamaguchi R,** Miyano S**. Multi-omics approach for estimating metabolic networks using low-order partial correlations.** *J Comput Biol***. 20(8):571-582, 2013.**
2. Kon A, Shih LY, Minamino M, Sanada M, Shiraishi Y, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, **Miyano S**, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genetics***. 45(10):1232-7, 2013.
3. Sakaguchi H, Okuno Y, Muramatsu H, Yoshida K, Shiraishi Y, Takahashi M, Kon A, Sanada M, Chiba K, Tanaka H, Makishima H, Wang X, Xu Y, Doisaki S, Hama A, Nakanishi K, Takahashi Y, Yoshida N, Maciejewski JP, **Miyano S**, Ogawa S, Kojima S. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. ***Nature Genetics***. 45(8):937-941, 2013.
4. **Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M,** Miyano S**, Homma Y, Ogawa S. Integrated molecular analysis of clear-cell renal cell carcinoma.** *Nature Genetics***. 45(8):860-867, 2013.**
5. **Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S,** Miyano S**. An empirical Bayesian framework for somatic mutation detection from cancer.** *Nucleic Acids Res***. 41(7): e89, 2013.**
6. **Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Hai Nguyen H, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi S, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N,** Miyano S**, Nakagama H, Nakamura Y, Tsunoda T, Shibata T, Nakagawa H. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators.** *Nature Genetics***. 44(7):760-764, 2012.**
7. **Niida A, Imoto S, Shimamura T,** Miyano S**. Statistical model-based testing to evaluate the recurrence of genomic aberrations.** *Bioinformatics***. 28(12):i115-i120, 2012.**
8. **Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H,** Miyano S**, Ogawa S. Frequent pathway mutations of splicing machinery in myelodysplasia.** *Nature***. 478(7367): 64-69, 2011.**
9. **Shimamura T, Imoto S, Shimada Y, Hosono Y, Niida A, Nagasaki M, Yamaguchi R, Takahashi T,** Miyano S**. A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition.** *PLoS One***. 6(6): e20804, 2011.**
10. **Shiraishi Y, Okada-Hatakeyama M,** Miyano S**. A rank-based statistical test for measuring synergistic effects between two gene sets.** *Bioinformatics***. 27 (17): 2399-2405, 2011.**
11. **Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe G, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M,** Miyano S**, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing.** *Nature Genetics***. 42: 931–936, 2010.**
12. **Niida A, Imoto S, Yamaguchi R, Nagasaki M,** Miyano S**. Gene set-based module discovery decodes cis-regulatory codes governing diverse gene expression across human multiple tissues.** *PLoS One***. 5(6):e10910, 2010.**

Seiya Imoto, PhD (ICGC PI's Laboratory Member)
Associate Professor, Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
E-mail: imoto@ims.u-tokyo.ac.jp

EDUCATIONS/TRAINING

1992-1996 Department of Mathematics, Undergraduate School, Kyushu University
1996-2001 Graduate School of Mathematics, Kyushu University
2001 Ph.D

POSITIONS

1999-2001 Research Fellow
Japan Society of the Promotion of Science for Young Scientists (DC2, Statistical Science)
2001-2001 Post Doctoral Fellow
Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo
2001-2007 Assistant Professor
Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo
2007-Present Associate Professor
Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo

SELECTED PUBLICATIONS

1. T. Yoshimaru, M. Komatsu, T. Matsuo, Y-A. Chen, Y. Murakami, K. Mizuguchi, E. Mizohata, T. Inoue, M. Akiyama, R. Yamaguchi, **S. Imoto**, S. Miyano, Y. Miyoshi, M. Sasah, Y. Nakamura, T. Katagiri (2013) Targeting the BIG3-PHB2 interaction to overcome tamoxifen resistance in breast cancer cells, ***Nature Communications***, 4, 2443.
2. M. Kayano, **S. Imoto**, R. Yamaguchi, S. Miyano (2013) Multi-omics approach for estimating metabolic networks using low-order partial correlations, ***Journal of Computational Biology***, 20(8), 571-582.
3. M. Affara, D. Sanders, H. Araki, Y. Tamada, B.J. Dunmore, S. Humphreys, **S. Imoto**, C. Savoie, S. Miyano, S. Kuhara, D. Jeffries, C. Print, D.S. Charnock-Jones (2013) Vasohibin-1 is identified as a master-regulator of endothelial cell apoptosis using gene network analysis. ***BMC Genomics***, 14(1): 23.
4. K. Ogami, R. Yamaguchi, **S. Imoto**, Y. Tamada, H. Araki, C. Print, S. Miyano (2012) Computational gene network analysis reveals TNF-induced angiogenesis. ***BMC Systems Biology***, 6(Suppl 2): S12.
5. A. Niida, **S. Imoto**, T. Shimamura, S. Miyano (2012) Statistical model-based testing to evaluate the recurrence of genomic aberrations, ***Bioinformatics***, 28, i115-i120.
6. S. Kawano, T. Shimamura, A. Niida, **S. Imoto**, R. Yamaguchi, M. Nagasaki, R. Yoshida, C. Print, S. Miyano (2012) Identifying gene pathways associated with cancer characteristics via sparse statistical methods, ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, 9(4):966-972.
7. Y. Tamada, **S. Imoto**, S. Miyano (2011) Parallel algorithm for learning optimal Bayesian network structure, ***J. Machine Learning Research***, 12, 2437-2459.
8. A. Niida, **S. Imoto**, R. Yamaguchi, M. Nagasaki, A. Fujita, T. Shimamura, S. Miyano (2010) Model-free unsupervised gene set screening based on information enrichment in expression profiles, ***Bioinformatics***, 26, 3090-3097.
9. K. Kojima, E. Perrier, **S. Imoto**, S. Miyano (2010) Optimal search on clustered structural constraint for learning Bayesian network structure, ***J. Machine Learning Research***, 11, 285-310.

RUI YAMAGUCHI, PhD (ICGC PI's Laboratory Member)
Lecturer, Laboratory of Sequence Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
E-mail: ruiy@ims.u-tokyo.ac.jp

Academic qualifications

1998 BS School of Sciences, Kyushu University
2000 MS Graduate School of Sciences, Kyushu University
2003 PhD Graduate School of Sciences, Kyushu University

Professional Records

2003-2003 Post Doctoral Fellow
Institute of Statistical Mathematics
2003-2006 Post Doctoral Fellow
Department of Mathematics, Faculty of Science, Kyushu University
2006-2007 Project Lecturer
Laboratory of Biostatistics, Human Genome Center,
The Institute of Medical Science, The University of Tokyo
2007-2009 Project Lecturer
Laboratory of DNA Information Analysis, Human Genome Center,
The Institute of Medical Science, The University of Tokyo
2009-Present Lecturer
Laboratory of Sequence Analysis, Human Genome Center,
The Institute of Medical Science, The University of Tokyo

Selected Publications

1. **Yoshimaru T, Komatsu M, Matsuo T, Chen Y-A, Murakami Y, Mizuguchi K, Mizohata E, Inoue T, Akiyama M,** Yamaguchi R**, Imoto S, Miyano S, Miyoshi Y, Sasa M, Nakamura Y, Katagiri T. Targeting BIG3-PHB interaction to overcome tamoxifen resistance in breast cancer cells.** *Nat Commun* **4:2443, 2013.**
2. **Kayano M, Imoto S,** Yamaguchi R**, Miyano S. Multi-omics approach for estimating metabolic networks using low-order partial correlations.** *J Comput Biol* **20(8):571-582, 2013.**
3. **Ogami K,** Yamaguchi R**, Imoto S, Tamada Y, Araki H, Print C, Miyano S. Computational gene network analysis reveals TNF-induced angiogenesis,** *BMC Systems Biology*, **6 Suppl 2:S12, 2012.**
4. **Yamauchi M,** Yamaguchi R**, Nakata A, Kohno T, Nagasaki M, Shimamura T, Imoto S, Saito A, Ueno K, Hatanaka Y, Yoshida R, Higuchi T, Nomura M, Beer DG, Yokota J, Miyano S, Gotoh N. Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma,** *PLoS ONE* **7(9): e43923, 2012.**
5. **Yamamoto M,** Yamaguchi R**, Munakata K, Takashima K, Nishiyama M, Hioki K, Ohnishi Y, Nagasaki M, Imoto S, Miyano S, Ishige A, Watanabe K. A microarray analysis of gnotobiotic mice indicating that microbial exposure during the neonatal period plays an essential role in immune system development,** *BMC Genomics*, **13:335, 2012.**
6. **Kawano S, Shimamura T, Niida A, Imoto S,** Yamaguchi R**, Nagasaki M, Yoshida R, Print C, Miyano S. Identifying gene pathways associated with cancer characteristics via sparse statistical methods,** *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9(4):966-972, 2012.**
7. **Tamada Y,** Yamaguchi R**, Imoto S, Hirose O, Yoshida R, Nagasaki M, Miyano S. SiGN-SSM: open source parallel software for estimating gene networks with state space models.** *Bioinformatics*, **27(8):1172-1173, 2011.**
8. **Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe A, Boroevich KA, Nagasaki M,** Yamaguchi R**, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing.** *Nat Genet*, **42(11):931-936, 2010.**
9. **Niida A, Imoto S,** Yamaguchi R**, Nagasaki M, Fujita A, Shimamura T, Miyano S. Model-free unsupervised gene set screening based on information enrichment in expression profiles.** *Bioinformatics*, **26(24):3090-3097, 2010.**

Shinichi Mizuno, MD, PhD (Non-ICGC/TCGA Collaborator)

Associate Professor, Division of Cancer Research, Center for Advanced Medical Innovation, Kyushu University

3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan; e-mail: mizuno@camiku.kyushu-u.ac.jp; Phone: +81926424760

Degrees

1989	**MD**	**Kyushu University, Fukuoka, Japan (Internal Medicine)**
2001	**PhD**	**Kyushu University, Fukuoka, Japan (Molecular biology)**

Positions

1989-1991	Clinical Fellow, First Dept. of Internal Medicine, Kyushu University
1992-1995	Research Fellow, First Dept. of Internal Medicine, Kyushu University
1996-2001	Research Fellow, Institute of Genetic Information, Kyushu University
2001-2008	Postdoctoral Fellow, Department of Cancer Immunology & AIDS, Dana-Farber Cancer Institute, Boston MA
2009-2012	Assistant Professor, Division of Hematology and Oncology, Kurume University School of Medicine
2012-present	Associate Professor, Division of Cancer Research, Center for Advanced Medical Innovation, Kyushu University

Awards

2001: Outstanding Young Investigator Award (Japanese Society of Hematology)
2002: Uehara Memorial Foundation Award of postdoctoral Fellow

Selected Peer-Reviewed Publications

1. **Nakaya T, Ishiguro KI, Belzil C, Rietsch AM, Yu Q,** Mizuno SI**, Bronson RT, Geng Y, Nguyen MD, Akashi K, Sicinski P, Nakatani Y. p600 Plays Essential Roles in Fetal Development.** *PLoS One*. **8, e66269, 2013**
2. **Kalaszczynska I, Geng Y, Iino T,** Mizuno S**, Choi Y, Kondratiuk I, Silver DP, Wolgemuth DJ, Akashi K, Sicinski P. Cyclin A is redundant in fibroblasts but essential in hematopoietic and embryonic stem cells.** *Cell*. **138, 352-65, 2009**
3. **Arinobu Y,** Mizuno S**, Chong Y, Shigematsu H, Iino T, Iwasaki H, Graf T, Mayfield R, Chan S, Kastner P, Akashi K. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages.** *Cell Stem Cell*. **1, 416-27, 2007**
4. **Iwasaki H,** Mizuno S**, Arinobu Y, Ozawa H, Mori Y, Shigematsu H, Takatsu K, Tenen DG, Akashi K.The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages.** *Genes Dev*. **20, 3010-21, 2006**
5. **Opferman JT, Iwasaki H, Ong CC, Suh H,** Mizuno S**, Akashi K, Korsmeyer SJ. Obligate role of anti-apoptotic MCL-1 in the survival of hematopoietic stem cells.** *Science*. **307, 1101-4, 2005**
6. **Iwasaki H,** Mizuno S**, Mayfield R, Shigematsu H, Arinobu Y, Seed B, Gurish MF, Takatsu K, Akashi K. Identification of eosinophil lineage-committed progenitors in the murine bone marrow.** *J Exp Med*. **201, 1891-7, 2005**
7. **Arinobu Y, Iwasaki H, Gurish MF,** Mizuno S**, Shigematsu H, Ozawa H, Tenen DG, Austen KF, Akashi K.Developmental checkpoints of the basophil/mast cell lineages in adult murine hematopoiesis.** *Proc Natl Acad Sci USA*. **102, 18105-10, 2005**
8. **Iwasaki H,** Mizuno S**, Wells RA, Cantor AB, Watanabe S, Akashi K. GATA-1 converts lymphoid and myelomonocytic progenitors into the megakaryocyte/erythrocyte lineages.** *Immunity*. **19, 451-462, 2003**
9. Mizuno S**, Chijiwa T, Okamura T, Akashi K, Fukumaki Y, Niho Y, Sasaki H. Expression of DNA methyltransferases DNMT1, 3A and 3B in normal hematopoiesis and in acute and chronic myelogenous leukemia.** *Blood*. **91, 1172-1179, 2001**
10. **Akashi K,** Mizuno S**. Epstein-Barr virus-infected natural killer cell leukemia.** *Leuk Lymphoma*. **40, 57-66, 2000**
11. Mizuno S**, Okamura T, Iwasaki H, Ohno Y, Akashi K, Inaba S, Niho Y. Hypercoagulable state following transfusions of granulocytes obtained from granulocyte colony- stimulating factor-stimulating donors.** *Int J Hematol*. **72, 115-117, 2000**
12. Mizuno S**, Akashi K, Ohshima K, Iwasaki H, Miyamoto T, Uchida N, Shibuya T, Harada M, Kikuchi M, Niho Y. Interferon-gamma prevents apoptosis in Epstein-Barr virus-infected natural killer cell leukemia in an autocrine fashion.** *Blood*. **93, 3439-3540, 1999**
13. **Takenaka K,** Mizuno S**, Harada M, Nagafuji K, Miyamoto T, Iwasaki H, Fujisaki T, Kubota A, Ohno Y, Arima F, Shigematsu H, Gondo H, Okamura T, Okamura S, Inaba S, Niho Y. Generation of human natural killer cells from peripheral blood CD34+cells mobilized by granulocyte colony-stimulating factor.** *Br J Haematol*. **92, 788-794, 1996**
14. **Akashi K,** Mizuno S**, Harada M, Kimura N, Kinjyo M, Shibuya T, Shimoda K, Takeshita M, Okamura S, Matsumoto I, Kikuchi M, Niho Y. T lymphoid / myeloid bilineal crisis in chronic myelogenous leukemia.** *Exp Hematol*. **21, 743-748, 1993**
15. Mizuno S**, Akashi K, Hirota Y, Gyoten S, Matsumoto I, Harada M, Niho Y. Filgrastim treatment of leukaemic transformation in myelodysplastic syndrome.** *Lancet*. **341, 1475-1476, 1993.**

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of accumulation of mutations in 3D protein structure

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hidewaki Nakagawa and Tatsuhiko Tsunoda (Riken, IMS, Japan)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Akihiro Fujimoto (Riken, IMS, Japan)	

Background and preliminary data

The accumulation of mutations in a gene is one of the most important signatures of driver genes, and many novel driver genes have been identified by the high number of observed mutations. In most cases, the unit to measure the accumulation of mutations is a gene. But most genes contain multiple domains, and each domain has a different function. In addition, the three dimensional protein structure of the protein is important to the function.

If a location of the mutation in a domain or the 3D structure is critical for carcinogenesis, examination of the mutation accumulation should increase detection power for driver genes. In most previous analyses, probably due to the small sample sizes, the localization of mutations in the 3D protein structure or domain has not yet been considered. Recently, Chapman *et al*. (Nature (2011)), found an accumulation of mutations in the *DIS3* gene within the RNB domain facing the enzyme's catalytic pocket, although statistical examination has not done for the mutations.

For the large number of mutations obtained from 2000 samples, manual review for all mutated genes would be difficult. Therefore, statistical evaluation of mutation accumulation is necessary. We considered 3D structure of a protein gene and tested accumulation of mutations. We divided protein 3D structure into 15 angstrom diameter spheres, and amino acids within these spheres were considered as a unit of statistical test. Statistical significance was obtained under the assumption of Poisson distribution. We adopted this method to *DIS3* gene and found a region of significant accumulation (P-value = 2.9e-06). We also applied this method to liver cancer data, and identified several significant regons.

Timelines & resources dedicated to project

This method requires a list of mutations in genes and protein structures downloaded from PDB. This method does not require a large number of cores nor disks. We think that this method will take a few weeks running on several cores.

Research proposal
Using the method, we will test mutation accumulation in a domain or 3D region in a gene. This method can identify statistically mutated regions in a gene, which are important regions for carcinogenesis, but below significance threshold when whole genes are analyzed, resulting in identification of new significantly mutated genes. **Additionally, this method would provide good implications into the interpretation of mutations; If a mutated gene is an oncogene, mutations should accumulate in a certain region, such as kinase domain, while if a mutated gene is a tumor suppressor gene, mutations are expected to be scattered throughout the gene.** **Using the method, we intend to (1) identify significantly mutated domains or 3D regions in genes, (2) provide suggestions to the rolls of significantly mutated genes and (3) identify gene regions and motifs that have large number of mutations across all cancer samples.** **We are currently doing the DACO and dbGAP approval processes (waiting for the institutional ethic committee's approval - by next February or March).**
Legacy plans
After this project, we will release our program as a tool.

Curriculum Vitae
HIDEWAKI NAKAGAWA, M.D., Ph.D.

Birth: 1966/April/28 at Osaka, Japan
Citizenship & Sex: Japanese, male
Address: Laboratory for Genome Sequencing Analysis, RIKEN (The Institute of Physical and Chemical Research) Center for Inegrative Medical Sciences
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
(within Institute of Medical Science, The University of Tokyo)
Phone: +81-3-5449-5376 FAX: +81-3-5449-5375
E-mail: hidewaki@ims.u-tokyo.ac.jp

Education
1985-1991 Osaka University, School of Medicine (M.D.)
1996-2000 Osaka University, Graduate School of Medicine (Ph.D.)

Training & Occupation
1991-1992 Osaka University Hospital, General Surgery, Resident
1992-1993 Osaka University Hospital, ICU/ Anesthesiology, Resident
1993-1996 National Osaka Hospital, General Surgery, Resident
1996-1999 Osaka University Hospital/ National Osaka Hospital, GI Surgery, Fellow
1999-2003 The Ohio State University, Human Cancer Genetics Program, Postdoctoral Fellow (Supervisor: Prof. A. de la Chapelle)
2003-2007 Institute of Medical Science, The University of Tokyo, Assitant Professor (Supervisor: Prof. Y. Nakamura)
2007-2008 Institute of Medical Science, The University of Tokyo, Associate Professor
2008-2013 Laboratory Head, Laboratory for Biomarker Development, RIKEN Center for Genomic Medicine
2013- Laboratory Head, Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences

Certificates & Licences
1991 Certificate of Medical Doctor, Japan
1995 Certificate of Surgery, Japan
1997 Certificate of GI Surgery, Japan
2008 Certificate of General Clinical Oncologist by Japanese Board of Cancer Therapy

Memberships
- American Association for Cancer Research (AACR), Active member
- The Japanese Cancer Association
- The Japan Surgical Society
- The Japanese Society of Gastroenterological Surgery
- The Japanese Society of Human Genomics
- American Society of Human Genetics (ASHG)

	Name:	Tatsuhiko Tsunoda, Dr. Ph.D. (Medicine) & Ph.D. (Engineering) Group Director
Affiliation:	Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Address:	1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan Phone: +81-45-503-9556 Email: tsunoda@src.riken.jp	
Major Field:	Medical science mathematics, genome medicine, statistical genetics, and cancer genome analysis	
Professional history:	1985-1989: B.S., Department of Physics, Faculty of Science, the University of Tokyo 1989-1991: M.S., Department of Physics (Elementary Particle Physics), the University of Tokyo 1992-1995: Ph.D., Department of Engineering, the University of Tokyo 1995-1997: Assistant Professor, Department of Engineering, Kyoto University 1997-1998: Research Associate, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 1998-2000: Assistant Professor, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 2000-2013: Laboratory Head, Laboratory for Medical Informatics, RIKEN SNP Research Center (2008-present: RIKEN Center for Genomic Medicine) 2011-2013: Director, Research Group for Medical Informatics, RIKEN Center for Genomic Medicine 2012-present: Visiting Professor, the Institute of Statistical Mathematics 2013-present: Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Membership and councilor	Councilor of the Japanese Cancer Association Councilor of the Japanese Society of Human Genetics Associate Editor of Cancer Science Journal Associate Editor of the Journal of Human Genetics Member of American Society of Human Genetics.	

Curriculum Vitae
Akihiro Fujimoto

July, 2013

Current address
Center for Genomic Medicine, RIKEN, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
TEL: +81-45-503-9288
FAX: +81-45-503-9555
E-mail: afujircb@src.riken.jp

Education
March, 2008 Ph. D., Department of Human Genetics, University of Tokyo
March, 2005 M. S., Department of Biological Science, Kyushu University
March, 2003 B. A., Department of Biological Science, Kyushu University

Professional experience
2011-current Center for Genomic Medicine, RIKEN (Senior Researcher)
2010-2011 Center for Genomic Medicine, RIKEN (Special Postdoctoral Researcher)
2008-2010 Data Analysis Fusion Team, Computational Science Research Program, RIKEN (Special Postdoctoral Researcher)

Peer-review papers
Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi SI, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, Miyano S, Nakagama H, Nakamura Y, Tsunoda T, Shibata T and Nakagawa H (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet** 44,760-764.

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. **Nat Genet** 42: 931–936

The International Cancer Genome Consortium (2010) International network of cancer genome projects. **Nature** 464, 993-998.

Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Tsunoda T, Tokunaga K, Ohashi J (2009) FGFR2 is associated with hair thickness in Asian populations. **J Hum Genet** 54: 461-5

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. **J Hum Genet** 53: 886-93

Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K (2008) Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. **Nat Genet** 40: 1324-8

Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K (2008) A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. **Hum Genet** 124: 179-85

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Furusawa T, Nakazawa M, Ohtsuka R, Tokunaga K (2008) A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. **Hum Mol Genet** 17: 835-43

Fujimoto A, Kado T, Yoshimaru H, Tsumura Y, Tachida H (2008) Adaptive and slightly deleterious evolution in a conifer, Cryptomeria japonica. **J Mol Evol** 67: 201-10

Referee for
Journal of Human Genetics, Molecular Biology and Evolution, Genes and Genetic Systems, Biochemical and Biophysical Research Communications, Journal of Molecular Evolution, Bioinformatics and BMC Genomics

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of allele frequencies in normal tissues and their relationship to somatic mutations

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hidewaki Nakagawa and Tatsuhiko Tsunoda (Riken, IMS, Japan)

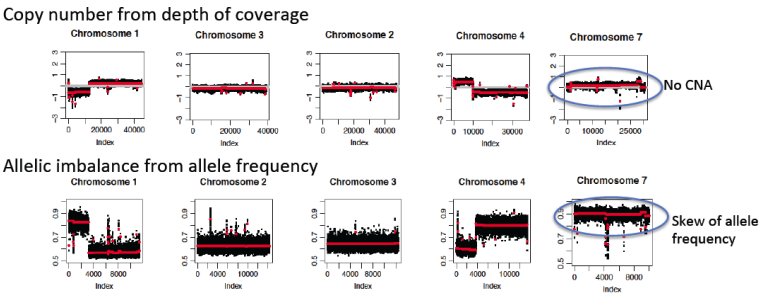
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Akihiro Fujimoto (Riken, IMS, Japan)	

Background and preliminary data

Mosaicism is the presence of different karyotypes in two or more cell lineages within an individual. A recent study analyzed DNA genotyping array data and found mosaicism in non-cancerous tissues in 2–3% of the elderly subjects (Jacobs et al., Nature Genetics, 2012). The mosaicism in non-cancerous tissues is related to cancer risk. This result suggests that proportion of cancer patients with mosaicism may be larger than that of normal populations.

In the NGS data, mosaicism can be identified by variant read frequency. We developed an analysis method to identify uniparental disomy (UPD) from cancer genome sequence.

In our method, we used heterozygote variant call frequency. Major call frequency for 10 neighboring SNVs with depth > 10 were averaged, and each average was clustered with a circular binary segmentation algorithm. After segmentation, the proportion of mosaicism is estimated using a maximum likelihood method. Since we have not adopted yet this method to non-cancerous samples, a result from cancer is shown in the left panel (CNA on chr1, 4 and UPD on chr7). This method is consistent with CNA from depth of coverage, therefore we consider our method can correctly identify UPD in the cancer sample. We would like to adapt this method to the identification of mosaicism in non-cancerous tissues.



Timelines & resources dedicated to project

This method requires VCF files for germline SNV with the number of variant and non-variant reads. If VCF files do not contain the number of variant calls, we can not perform this project, since variant calling from all the gremline BAM files would take a very long time.

Research proposal

In this project, we would like to identify mosaicism in non-cancerous tissues. Since the skew in allele frequency will likely not be large, we will focus on large regions of mosaicism (deletions or amplifications of several Mbp) with small differences in variant call frequency.
Mosaicism in non-cancerous tissues is not a common phenomena and therefore this project is a challenging one. If we can identify mosaicism in non-cancerous tissues, it would be the first example using NGS. We would like to compare mutation patterns in the corresponding cancers, such as the numbers of mutations, substitution patterns between samples with and without mosaicism. Since patients with germline mosaicism may have a predisposition towards genomic instability, we will search for germline variations in known cancer related genes. Germline mosaicism may be a good marker for cancer-risk, and may contribute to the future personalized medicine.

We are currently doing the DACO and dbGAP approval processes (waiting for the institutional ethic committee's approval - by next February or March).

Legacy plans

After this project, we will release our program as a tool.

Curriculum Vitae
HIDEWAKI NAKAGAWA, M.D., Ph.D.

Birth: 1966/April/28 at Osaka, Japan
Citizenship & Sex: Japanese, male
Address: Laboratory for Genome Sequencing Analysis, RIKEN (The Institute of Physical and Chemical Research) Center for Inegrative Medical Sciences
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
(within Institute of Medical Science, The University of Tokyo)
Phone: +81-3-5449-5376 FAX: +81-3-5449-5375
E-mail: hidewaki@ims.u-tokyo.ac.jp

Education
1985-1991 Osaka University, School of Medicine (M.D.)
1996-2000 Osaka University, Graduate School of Medicine (Ph.D.)

Training & Occupation
1991-1992 Osaka University Hospital, General Surgery, Resident
1992-1993 Osaka University Hospital, ICU/ Anesthesiology, Resident
1993-1996 National Osaka Hospital, General Surgery, Resident
1996-1999 Osaka University Hospital/ National Osaka Hospital, GI Surgery, Fellow
1999-2003 The Ohio State University, Human Cancer Genetics Program, Postdoctoral Fellow (Supervisor: Prof. A. de la Chapelle)
2003-2007 Institute of Medical Science, The University of Tokyo, Assitant Professor (Supervisor: Prof. Y. Nakamura)
2007-2008 Institute of Medical Science, The University of Tokyo, Associate Professor
2008-2013 Laboratory Head, Laboratory for Biomarker Development, RIKEN Center for Genomic Medicine
2013- Laboratory Head, Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences

Certificates & Licences
1991 Certificate of Medical Doctor, Japan
1995 Certificate of Surgery, Japan
1997 Certificate of GI Surgery, Japan
2008 Certificate of General Clinical Oncologist by Japanese Board of Cancer Therapy

Memberships
- American Association for Cancer Research (AACR), Active member
- The Japanese Cancer Association
- The Japan Surgical Society
- The Japanese Society of Gastroenterological Surgery
- The Japanese Society of Human Genomics
- American Society of Human Genetics (ASHG)

	Name:	Tatsuhiko Tsunoda, Dr. Ph.D. (Medicine) & Ph.D. (Engineering) Group Director
Affiliation:	Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Address:	1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan Phone: +81-45-503-9556 Email: tsunoda@src.riken.jp	
Major Field:	Medical science mathematics, genome medicine, statistical genetics, and cancer genome analysis	
Professional history:	1985-1989: B.S., Department of Physics, Faculty of Science, the University of Tokyo 1989-1991: M.S., Department of Physics (Elementary Particle Physics), the University of Tokyo 1992-1995: Ph.D., Department of Engineering, the University of Tokyo 1995-1997: Assistant Professor, Department of Engineering, Kyoto University 1997-1998: Research Associate, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 1998-2000: Assistant Professor, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 2000-2013: Laboratory Head, Laboratory for Medical Informatics, RIKEN SNP Research Center (2008-present: RIKEN Center for Genomic Medicine) 2011-2013: Director, Research Group for Medical Informatics, RIKEN Center for Genomic Medicine 2012-present: Visiting Professor, the Institute of Statistical Mathematics 2013-present: Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Membership and councilor	Councilor of the Japanese Cancer Association Councilor of the Japanese Society of Human Genetics Associate Editor of Cancer Science Journal Associate Editor of the Journal of Human Genetics Member of American Society of Human Genetics.	

Curriculum Vitae
Akihiro Fujimoto

July, 2013

Current address
Center for Genomic Medicine, RIKEN, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
TEL: +81-45-503-9288
FAX: +81-45-503-9555
E-mail: afujircb@src.riken.jp

Education
March, 2008 Ph. D., Department of Human Genetics, University of Tokyo
March, 2005 M. S., Department of Biological Science, Kyushu University
March, 2003 B. A., Department of Biological Science, Kyushu University

Professional experience
2011-current Center for Genomic Medicine, RIKEN (Senior Researcher)
2010-2011 Center for Genomic Medicine, RIKEN (Special Postdoctoral Researcher)
2008-2010 Data Analysis Fusion Team, Computational Science Research Program, RIKEN (Special Postdoctoral Researcher)

Peer-review papers
Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi SI, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, Miyano S, Nakagama H, Nakamura Y, Tsunoda T, Shibata T and Nakagawa H (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet** 44,760-764.

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. **Nat Genet** 42: 931–936

The International Cancer Genome Consortium (2010) International network of cancer genome projects. **Nature** 464, 993-998.

Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Tsunoda T, Tokunaga K, Ohashi J (2009) FGFR2 is associated with hair thickness in Asian populations. **J Hum Genet** 54: 461-5

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. **J Hum Genet** 53: 886-93

Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K (2008) Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. **Nat Genet** 40: 1324-8

Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K (2008) A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. **Hum Genet** 124: 179-85

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Furusawa T, Nakazawa M, Ohtsuka R, Tokunaga K (2008) A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. **Hum Mol Genet** 17: 835-43

Fujimoto A, Kado T, Yoshimaru H, Tsumura Y, Tachida H (2008) Adaptive and slightly deleterious evolution in a conifer, Cryptomeria japonica. **J Mol Evol** 67: 201-10

Referee for
Journal of Human Genetics, Molecular Biology and Evolution, Genes and Genetic Systems, Biochemical and Biophysical Research Communications, Journal of Molecular Evolution, Bioinformatics and BMC Genomics

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of cancer heterogeneity and identification of mutated genes and pathways with high clonal proportion

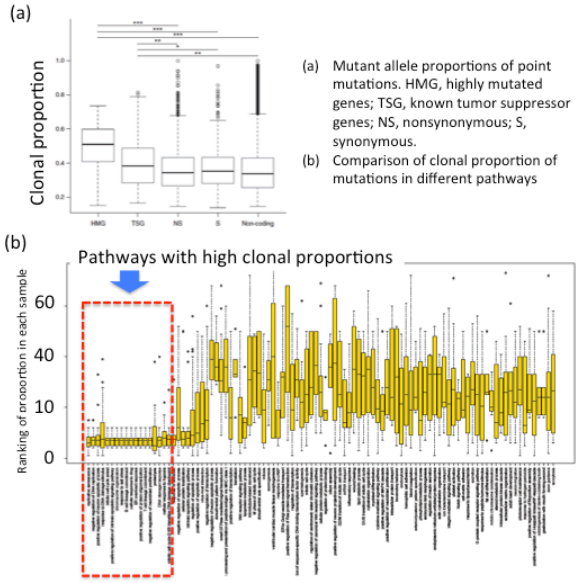
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hidewaki Nakagawa and Tatsuhiko Tsunoda (Riken, IMS, Japan)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Akihiro Fujimoto (Riken, IMS, Japan)	

Background and preliminary data

Tumors are populations of heterogeneous cancer cells. We previously analyzed the clonal proportion of mutations, and found that the more commonly mutated genes had a higher clonal proportion than other genes (Fujimoto et al., Nature Genetics, 2012). To analyze mutations and pathways with high clonal proportions, we performed deep sequencing of PCR amplicons of nonsynonymous mutations in liver cancer samples. We found that genes in sixteen categories, including "replicative senescence", "negative regulation of DNA replication" and "proteolysis", had a higher proportion of mutated alleles after adjustment for multiple testing. This result is consistent with a breast cancer study (Shah et al. Nature (2012)).



Timelines & resources dedicated to project

The analysis requires read counts of mutations, copy number data and allelic imbalance ratio in each region. Identification of pathways with higher clonal proration is not difficult and should finish within a few weeks. The estimation of population structure of cancer population (see below) will take longer.

Research proposal

This project had three steps.

(1) Adjustment of the clonal proportions for copy number and allelic imbalance.

Mutant allele proportion is influenced by copy number status and normal cell contamination. To adjust for this, the copy number of mutations and allelic imbalance ratio are used. First, we estimate variant allele proportions and confidence intervals within a region with high read depth (> 20). We then adjust the clonal proportion with allelic imbalance ratio. For the adjustment, we assume that the mutant allele is harbored on a chromosome in high clonal proportion, and adjust clonal proportion by the allelic imbalance. If phasing information (germline haplotype and read pair with SNP) can be obtained as in Nik-Zainal et al. Cell (2012), the adjustment becomes more accurate. But phasing of all germline genomes is too computationally intensive. Even if possible, the number of read pairs with the SNP allele and mutation would not be so large and it would not strongly influence the result. Contamination rate of normal cells is also necessary for the analysis. If the contamination rate is already available, we will use that rate. However, if the rate is not available, we will adjust the clonal proportions with the highest frequency in each sample.

(2) Comparison of clonal proportion among genes and pathways

We compare the clonal proportion of mutations in different genes and pathways. The clonal proportions are compared among genes and pathways. We also compare the genes and pathways with higher clonal proportion among different cancer types.

(3) Inferring population structure of tumors with mutant allele frequencies

We will also try to estimate population structure in a tumor using the ABC (approximate Bayesian computation) method. We assume several population structures, and perform MCMC estimation with forward simulation of population genetics. We will then select the best-fit model and estimate the number of populations and growth rate. Population structure should be different among cancer types. For example, leukemia cells can freely flow in human body, but solid cancer cells cannot, and therefore the leukemia populations may be more homogeneous than solid tumor ones. Our method would identify population structure of each cancer type and will give explanations for the difference in genetic heterogeneity and the number of mutations in the different cancer types. In addition, this method will provide us with useful suggestions about cancer development process.

We are currently doing the DACO and dbGAP approval processes (waiting for the institutional ethic committee's approval - by next February or March).

Legacy plans

After this project, we will release our program as a tool.

Curriculum Vitae
HIDEWAKI NAKAGAWA, M.D., Ph.D.

Birth: 1966/April/28 at Osaka, Japan
Citizenship & Sex: Japanese, male
Address: Laboratory for Genome Sequencing Analysis, RIKEN (The Institute of Physical and Chemical Research) Center for Inegrative Medical Sciences
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
(within Institute of Medical Science, The University of Tokyo)
Phone: +81-3-5449-5376 FAX: +81-3-5449-5375
E-mail: hidewaki@ims.u-tokyo.ac.jp

Education
1985-1991 Osaka University, School of Medicine (M.D.)
1996-2000 Osaka University, Graduate School of Medicine (Ph.D.)

Training & Occupation
1991-1992 Osaka University Hospital, General Surgery, Resident
1992-1993 Osaka University Hospital, ICU/ Anesthesiology, Resident
1993-1996 National Osaka Hospital, General Surgery, Resident
1996-1999 Osaka University Hospital/ National Osaka Hospital, GI Surgery, Fellow
1999-2003 The Ohio State University, Human Cancer Genetics Program, Postdoctoral Fellow (Supervisor: Prof. A. de la Chapelle)
2003-2007 Institute of Medical Science, The University of Tokyo, Assitant Professor (Supervisor: Prof. Y. Nakamura)
2007-2008 Institute of Medical Science, The University of Tokyo, Associate Professor
2008-2013 Laboratory Head, Laboratory for Biomarker Development, RIKEN Center for Genomic Medicine
2013- Laboratory Head, Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences

Certificates & Licences
1991 Certificate of Medical Doctor, Japan
1995 Certificate of Surgery, Japan
1997 Certificate of GI Surgery, Japan
2008 Certificate of General Clinical Oncologist by Japanese Board of Cancer Therapy

Memberships
- American Association for Cancer Research (AACR), Active member
- The Japanese Cancer Association
- The Japan Surgical Society
- The Japanese Society of Gastroenterological Surgery
- The Japanese Society of Human Genomics
- American Society of Human Genetics (ASHG)

	Name:	Tatsuhiko Tsunoda, Dr. Ph.D. (Medicine) & Ph.D. (Engineering) Group Director
Affiliation:	Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Address:	1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan Phone: +81-45-503-9556 Email: tsunoda@src.riken.jp	
Major Field:	Medical science mathematics, genome medicine, statistical genetics, and cancer genome analysis	
Professional history:	1985-1989: B.S., Department of Physics, Faculty of Science, the University of Tokyo 1989-1991: M.S., Department of Physics (Elementary Particle Physics), the University of Tokyo 1992-1995: Ph.D., Department of Engineering, the University of Tokyo 1995-1997: Assistant Professor, Department of Engineering, Kyoto University 1997-1998: Research Associate, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 1998-2000: Assistant Professor, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 2000-2013: Laboratory Head, Laboratory for Medical Informatics, RIKEN SNP Research Center (2008-present: RIKEN Center for Genomic Medicine) 2011-2013: Director, Research Group for Medical Informatics, RIKEN Center for Genomic Medicine 2012-present: Visiting Professor, the Institute of Statistical Mathematics 2013-present: Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Membership and councilor	Councilor of the Japanese Cancer Association Councilor of the Japanese Society of Human Genetics Associate Editor of Cancer Science Journal Associate Editor of the Journal of Human Genetics Member of American Society of Human Genetics.	

Curriculum Vitae
Akihiro Fujimoto

July, 2013

Current address
Center for Genomic Medicine, RIKEN, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
TEL: +81-45-503-9288
FAX: +81-45-503-9555
E-mail: afujircb@src.riken.jp

Education
March, 2008 Ph. D., Department of Human Genetics, University of Tokyo
March, 2005 M. S., Department of Biological Science, Kyushu University
March, 2003 B. A., Department of Biological Science, Kyushu University

Professional experience
2011-current Center for Genomic Medicine, RIKEN (Senior Researcher)
2010-2011 Center for Genomic Medicine, RIKEN (Special Postdoctoral Researcher)
2008-2010 Data Analysis Fusion Team, Computational Science Research Program, RIKEN (Special Postdoctoral Researcher)

Peer-review papers
Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi SI, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, Miyano S, Nakagama H, Nakamura Y, Tsunoda T, Shibata T and Nakagawa H (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet** 44,760-764.

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. **Nat Genet** 42: 931–936

The International Cancer Genome Consortium (2010) International network of cancer genome projects. **Nature** 464, 993-998.

Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Tsunoda T, Tokunaga K, Ohashi J (2009) FGFR2 is associated with hair thickness in Asian populations. **J Hum Genet** 54: 461-5

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. **J Hum Genet** 53: 886-93

Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K (2008) Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. **Nat Genet** 40: 1324-8

Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K (2008) A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. **Hum Genet** 124: 179-85

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Furusawa T, Nakazawa M, Ohtsuka R, Tokunaga K (2008) A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. **Hum Mol Genet** 17: 835-43

Fujimoto A, Kado T, Yoshimaru H, Tsumura Y, Tachida H (2008) Adaptive and slightly deleterious evolution in a conifer, Cryptomeria japonica. **J Mol Evol** 67: 201-10

Referee for
Journal of Human Genetics, Molecular Biology and Evolution, Genes and Genetic Systems, Biochemical and Biophysical Research Communications, Journal of Molecular Evolution, Bioinformatics and BMC Genomics

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of microsatellite instability (MSI)

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

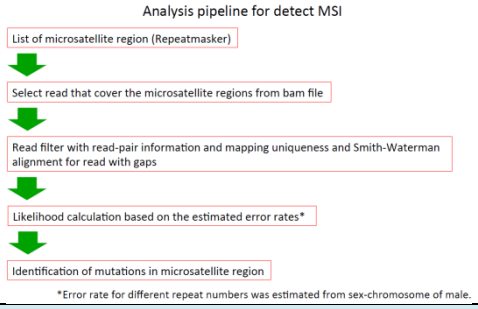
Hidewaki Nakagawa and Tatsuhiko Tsunoda (Riken, IMS, Japan)

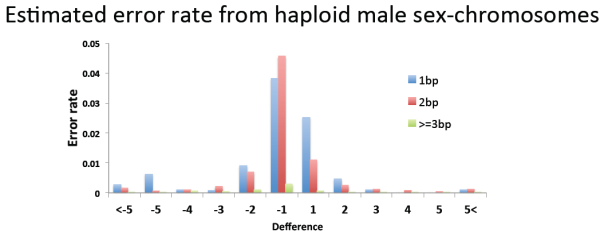
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Akihiro Fujimoto (Riken, IMS, Japan)	

Background and preliminary data

Microsatellite instability (MSI) is a genetic hypermutability that results from impaired DNA Mismatch Repair. MSI is one of the most important features of the mutation signature in cancer. However, MSI identification is difficult by NGS, and only a few studies have been carried out to identify MSI in cancer. We have developed a method to identify mutations in microsatellite regions and applied the method to liver cancer samples.

We hypothesized that sequencing artifacts in microsatellite regions are caused by mapping errors, PCR errors and alignment errors. Mapping errors were excluded by filtering inconsistent read pairs and read pairs with low mapping quality. To exclude alignment errors, we implemented Smith-Waterman alignments for reads that were mapped to microsatellite regions with gaps. To exclude PCR errors, we first estimated the error rate for different patterns of microsatellite, such as mononucleotide, dinucleotide, and trinucleotide repeats. Based on the estimated PCR error rates, we calculated the likelihood for each microsatellite region under the assumption that variation in read length is caused by PCR errors. Based on the likelihood, we identified mutations in microsatellite regions.





Error rate was different among the different repeat numbers, and the mono- and di-nucleotide repeats showed high error rate than tori- or longer repeats.

Timelines & resources dedicated to project

(1) We have already performed verification for 40 candidates, and the false positive rate was less than 10%, but we need more experimental verification to validate the results. If our abstract is adopted, we will first perform additional verification. This verification and parameter adjustment based on the result would take one or two month(s).

(2) Our method require bam file and take a few days to analyze one sample on one CPU core. Therefore, if we can use three hundred cores, we will be able to finish analysis for the 2,000 samples within a month.

If SW alignment is not used, our program is faster and takes about 8 hours for one sample on one CPU core.

Research proposal

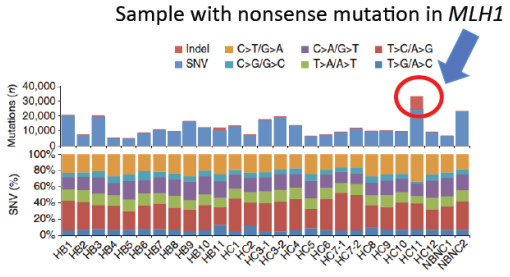
MIS is a key feature of cancer. The pattern of MSI would be influenced by the cellular origins, cancer types, mutated genes and mutagens. The purpose of this research is (1) to compare mutation patterns of MSI among cancer types, (2) to compare MSI pattern and substitution pattern, (3) to identify mutated genes and pathways in the samples with MSI.

(1) Comparison of mutation patterns among cancer types
The number of mutated microsatellites, length, and mutated nucleotides are compared, and a cancer type specific MSI pattern is identified. For the comparison, PCA (principal component analysis) and NMF (non-negative matrix factorization) would be adopted.

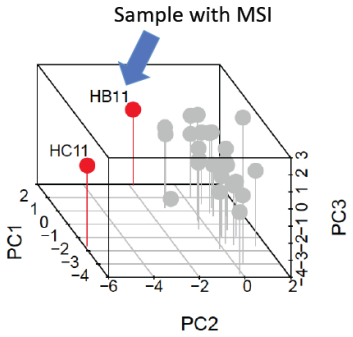
(2) Compare MSI pattern and mutation pattern
In our previous analysis, we found that a sample with nonsense mutations in *MLH1* gene had an exceptionally large number of indels, suggesting MIS. Also, the sample had different substitution pattern compared to other samples, and the sample showed transcription coupled repair deficiency (see below). We thought that *MLH1* inactivation may cause MSI and influence the substitution pattern.
By comparing relationship between the MSI and the substitution patterns, we may be able to identify the genes that influence both types of mutations. We will also compare the pattern of rearrangements and CNAs between samples with and without MSI.

(3) Identification of mutated genes and pathways in the samples with MSI
Based on the detection of MSI and the substitution pattern, we will analyze genes with point mutations, rearrangements, CNAs and indels in the MSI samples. Recurrently mutated genes and pathways in the samples with MSI would be identified.

Comparison of number and pattern of mutations



PCA of substitution pattern



We are currently doing the DACO and dbGAP approval processes (waiting for the institutional ethic committee's approval - by next February or March).

Legacy plans

After this project, we will release our program as a tool.

Curriculum Vitae
HIDEWAKI NAKAGAWA, M.D., Ph.D.

Birth: 1966/April/28 at Osaka, Japan
Citizenship & Sex: Japanese, male
Address: Laboratory for Genome Sequencing Analysis, RIKEN (The Institute of Physical and Chemical Research) Center for Inegrative Medical Sciences
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
(within Institute of Medical Science, The University of Tokyo)
Phone: +81-3-5449-5376 FAX: +81-3-5449-5375
E-mail: hidewaki@ims.u-tokyo.ac.jp

Education
1985-1991 Osaka University, School of Medicine (M.D.)
1996-2000 Osaka University, Graduate School of Medicine (Ph.D.)

Training & Occupation
1991-1992 Osaka University Hospital, General Surgery, Resident
1992-1993 Osaka University Hospital, ICU/ Anesthesiology, Resident
1993-1996 National Osaka Hospital, General Surgery, Resident
1996-1999 Osaka University Hospital/ National Osaka Hospital, GI Surgery, Fellow
1999-2003 The Ohio State University, Human Cancer Genetics Program, Postdoctoral Fellow (Supervisor: Prof. A. de la Chapelle)
2003-2007 Institute of Medical Science, The University of Tokyo, Assitant Professor (Supervisor: Prof. Y. Nakamura)
2007-2008 Institute of Medical Science, The University of Tokyo, Associate Professor
2008-2013 Laboratory Head, Laboratory for Biomarker Development, RIKEN Center for Genomic Medicine
2013- Laboratory Head, Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences

Certificates & Licences
1991 Certificate of Medical Doctor, Japan
1995 Certificate of Surgery, Japan
1997 Certificate of GI Surgery, Japan
2008 Certificate of General Clinical Oncologist by Japanese Board of Cancer Therapy

Memberships
- American Association for Cancer Research (AACR), Active member
- The Japanese Cancer Association
- The Japan Surgical Society
- The Japanese Society of Gastroenterological Surgery
- The Japanese Society of Human Genomics
- American Society of Human Genetics (ASHG)

	Name:	Tatsuhiko Tsunoda, Dr. Ph.D. (Medicine) & Ph.D. (Engineering) Group Director
Affiliation:	Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Address:	1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan Phone: +81-45-503-9556 Email: tsunoda@src.riken.jp	
Major Field:	Medical science mathematics, genome medicine, statistical genetics, and cancer genome analysis	
Professional history:	1985-1989: B.S., Department of Physics, Faculty of Science, the University of Tokyo 1989-1991: M.S., Department of Physics (Elementary Particle Physics), the University of Tokyo 1992-1995: Ph.D., Department of Engineering, the University of Tokyo 1995-1997: Assistant Professor, Department of Engineering, Kyoto University 1997-1998: Research Associate, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 1998-2000: Assistant Professor, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 2000-2013: Laboratory Head, Laboratory for Medical Informatics, RIKEN SNP Research Center (2008-present: RIKEN Center for Genomic Medicine) 2011-2013: Director, Research Group for Medical Informatics, RIKEN Center for Genomic Medicine 2012-present: Visiting Professor, the Institute of Statistical Mathematics 2013-present: Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Membership and councilor	Councilor of the Japanese Cancer Association Councilor of the Japanese Society of Human Genetics Associate Editor of Cancer Science Journal Associate Editor of the Journal of Human Genetics Member of American Society of Human Genetics.	

Curriculum Vitae
Akihiro Fujimoto

July, 2013

Current address
Center for Genomic Medicine, RIKEN, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
TEL: +81-45-503-9288
FAX: +81-45-503-9555
E-mail: afujircb@src.riken.jp

Education
March, 2008 Ph. D., Department of Human Genetics, University of Tokyo
March, 2005 M. S., Department of Biological Science, Kyushu University
March, 2003 B. A., Department of Biological Science, Kyushu University

Professional experience
2011-current Center for Genomic Medicine, RIKEN (Senior Researcher)
2010-2011 Center for Genomic Medicine, RIKEN (Special Postdoctoral Researcher)
2008-2010 Data Analysis Fusion Team, Computational Science Research Program, RIKEN (Special Postdoctoral Researcher)

Peer-review papers
Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi SI, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, Miyano S, Nakagama H, Nakamura Y, Tsunoda T, Shibata T and Nakagawa H (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet** 44,760-764.

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. **Nat Genet** 42: 931–936

The International Cancer Genome Consortium (2010) International network of cancer genome projects. **Nature** 464, 993-998.

Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Tsunoda T, Tokunaga K, Ohashi J (2009) FGFR2 is associated with hair thickness in Asian populations. **J Hum Genet** 54: 461-5

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. **J Hum Genet** 53: 886-93

Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K (2008) Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. **Nat Genet** 40: 1324-8

Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K (2008) A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. **Hum Genet** 124: 179-85

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Furusawa T, Nakazawa M, Ohtsuka R, Tokunaga K (2008) A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. **Hum Mol Genet** 17: 835-43

Fujimoto A, Kado T, Yoshimaru H, Tsumura Y, Tachida H (2008) Adaptive and slightly deleterious evolution in a conifer, Cryptomeria japonica. **J Mol Evol** 67: 201-10

Referee for
Journal of Human Genetics, Molecular Biology and Evolution, Genes and Genetic Systems, Biochemical and Biophysical Research Communications, Journal of Molecular Evolution, Bioinformatics and BMC Genomics

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of mitochondrial heteroplasmy and copy number in cancer tissue

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hidewaki Nakagawa and Tatsuhiko Tsunoda (Riken, IMS, Japan)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Akihiro Fujimoto (Riken, IMS, Japan)	

Background and preliminary data

Mitochondria are organelles that generate cellular energy, produce reactive oxygen species (ROS) and start apoptosis. As each cell has 10-300 mitochondria, the mitochondria can be heterogeneous in both a cancer cell and in a tissue (heteroplasmy). Despite its biological importance, mutations in mitochondria have not been well studied in recent cancer genome analysis.

Since the depth of coverage of the mitochondrial genome is much higher than nuclear genomes, a specific mutation caller for mitochondria should be prepared. We have developed a method to identify mutations based on likelihood (Fujimoto et al, Nature Genetics, 2010), and the method can be used for high coverage regions. Using this method, we will identify mutations in mitochondria, and analyze copy number of mitochondria and heteroplasmy among various cancer types.

Timelines & resources dedicated to project

Cancer and normal BAM files for mitochondria.

Research proposal

The purposes of this study are the following:
(1) Identify mutations in mitochondria, (2) detect heteroplasmy and copy number alternations, and (3) compare copy number and heteroplasmy of mitochondria among various cancer types.

(1) Identify mutations in mitochondria
Mitochondria genome sequence data has a very high depth of coverage. In our data, the depth of coverage of mitochondria is ~2,000 in cancer and ~800 in normal tissue. To identify mutations in mitochondrial genome, we may have to prepare a customized calling method. In a previous study, we developed a method based on likelihood. We applied this method to mitochondria and identified ~30 mutations in cancer mitochondria. Most of them are of low frequency, suggesting heteroplasmy in the cancer tissue. (If the mutation caller for the Pan-cancer project can accurately detect mutations in mitochondria, we will not stick to use our method. But comparison of the methods may be useful.)

(2) Detection of heteroplasmy
Using the proportion of mutant alleles, we can identify heteroplasmy in the mitochondrial genome. We can also identify copy number changes based on the ratio of the depth of coverage between the cancer and the matched normal tissues. The copy number and allele frequency is also influenced by normal cell contamination of the cancer tissue. If tumor content has already been estimated, we will use to adjust the allele frequency and copy number. If not, we will use allele frequency of mutations in autosomes for the adjustment.

(3) Compare copy number and heteroplasmy of mitochondria among various cancer types
We will compare mitochondrial mutations, heteroplasmy and copy number among cancer tissues, and identify highly mutated genes in mitochondria, cancer types with high heteroplasmy, and samples with a larger number of mitochondria. If we can identify samples with larger amount of mitochondrial mutations, we can analyze clinical information, such as survival rate, existence of metastasis, etc. to elucidate the functional role of these mutations. This analysis may help to understand the mechanism of the "Warburg effect", which is when cancer cells produce energy by glycolysis followed by lactic acid fermentation and hypoxia-induced mitochondria respiration.

We are currently doing the DACO and dbGAP approval processes (waiting for the institutional ethic committee's approval - by next February or March).

Legacy plans

After this project, we will release our program as a tool.

Curriculum Vitae
HIDEWAKI NAKAGAWA, M.D., Ph.D.

Birth: 1966/April/28 at Osaka, Japan
Citizenship & Sex: Japanese, male
Address: Laboratory for Genome Sequencing Analysis, RIKEN (The Institute of Physical and Chemical Research) Center for Inegrative Medical Sciences
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
(within Institute of Medical Science, The University of Tokyo)
Phone: +81-3-5449-5376 FAX: +81-3-5449-5375
E-mail: hidewaki@ims.u-tokyo.ac.jp

Education
1985-1991 Osaka University, School of Medicine (M.D.)
1996-2000 Osaka University, Graduate School of Medicine (Ph.D.)

Training & Occupation
1991-1992 Osaka University Hospital, General Surgery, Resident
1992-1993 Osaka University Hospital, ICU/ Anesthesiology, Resident
1993-1996 National Osaka Hospital, General Surgery, Resident
1996-1999 Osaka University Hospital/ National Osaka Hospital, GI Surgery, Fellow
1999-2003 The Ohio State University, Human Cancer Genetics Program, Postdoctoral Fellow (Supervisor: Prof. A. de la Chapelle)
2003-2007 Institute of Medical Science, The University of Tokyo, Assitant Professor (Supervisor: Prof. Y. Nakamura)
2007-2008 Institute of Medical Science, The University of Tokyo, Associate Professor
2008-2013 Laboratory Head, Laboratory for Biomarker Development, RIKEN Center for Genomic Medicine
2013- Laboratory Head, Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences

Certificates & Licences
1991 Certificate of Medical Doctor, Japan
1995 Certificate of Surgery, Japan
1997 Certificate of GI Surgery, Japan
2008 Certificate of General Clinical Oncologist by Japanese Board of Cancer Therapy

Memberships
- American Association for Cancer Research (AACR), Active member
- The Japanese Cancer Association
- The Japan Surgical Society
- The Japanese Society of Gastroenterological Surgery
- The Japanese Society of Human Genomics
- American Society of Human Genetics (ASHG)

	Name:	Tatsuhiko Tsunoda, Dr. Ph.D. (Medicine) & Ph.D. (Engineering) Group Director
Affiliation:	Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Address:	1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan Phone: +81-45-503-9556 Email: tsunoda@src.riken.jp	
Major Field:	Medical science mathematics, genome medicine, statistical genetics, and cancer genome analysis	
Professional history:	1985-1989: B.S., Department of Physics, Faculty of Science, the University of Tokyo 1989-1991: M.S., Department of Physics (Elementary Particle Physics), the University of Tokyo 1992-1995: Ph.D., Department of Engineering, the University of Tokyo 1995-1997: Assistant Professor, Department of Engineering, Kyoto University 1997-1998: Research Associate, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 1998-2000: Assistant Professor, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 2000-2013: Laboratory Head, Laboratory for Medical Informatics, RIKEN SNP Research Center (2008-present: RIKEN Center for Genomic Medicine) 2011-2013: Director, Research Group for Medical Informatics, RIKEN Center for Genomic Medicine 2012-present: Visiting Professor, the Institute of Statistical Mathematics 2013-present: Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Membership and councilor	Councilor of the Japanese Cancer Association Councilor of the Japanese Society of Human Genetics Associate Editor of Cancer Science Journal Associate Editor of the Journal of Human Genetics Member of American Society of Human Genetics.	

Curriculum Vitae
Akihiro Fujimoto

July, 2013

Current address
Center for Genomic Medicine, RIKEN, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
TEL: +81-45-503-9288
FAX: +81-45-503-9555
E-mail: afujircb@src.riken.jp

Education
March, 2008 Ph. D., Department of Human Genetics, University of Tokyo
March, 2005 M. S., Department of Biological Science, Kyushu University
March, 2003 B. A., Department of Biological Science, Kyushu University

Professional experience
2011-current Center for Genomic Medicine, RIKEN (Senior Researcher)
2010-2011 Center for Genomic Medicine, RIKEN (Special Postdoctoral Researcher)
2008-2010 Data Analysis Fusion Team, Computational Science Research Program, RIKEN (Special Postdoctoral Researcher)

Peer-review papers
Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi SI, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, Miyano S, Nakagama H, Nakamura Y, Tsunoda T, Shibata T and Nakagawa H (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet** 44,760-764.

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. **Nat Genet** 42: 931–936

The International Cancer Genome Consortium (2010) International network of cancer genome projects. **Nature** 464, 993-998.

Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Tsunoda T, Tokunaga K, Ohashi J (2009) FGFR2 is associated with hair thickness in Asian populations. **J Hum Genet** 54: 461-5

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. **J Hum Genet** 53: 886-93

Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K (2008) Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. **Nat Genet** 40: 1324-8

Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K (2008) A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. **Hum Genet** 124: 179-85

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Furusawa T, Nakazawa M, Ohtsuka R, Tokunaga K (2008) A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. **Hum Mol Genet** 17: 835-43

Fujimoto A, Kado T, Yoshimaru H, Tsumura Y, Tachida H (2008) Adaptive and slightly deleterious evolution in a conifer, Cryptomeria japonica. **J Mol Evol** 67: 201-10

Referee for
Journal of Human Genetics, Molecular Biology and Evolution, Genes and Genetic Systems, Biochemical and Biophysical Research Communications, Journal of Molecular Evolution, Bioinformatics and BMC Genomics

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Genome-wide search for genetic markers associated with survival time in pan-cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Tatsuhiro Shibata, National Cancer Center Japan, ICGC liver cancer

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Mamoru Kato, National Cancer Center Japan	

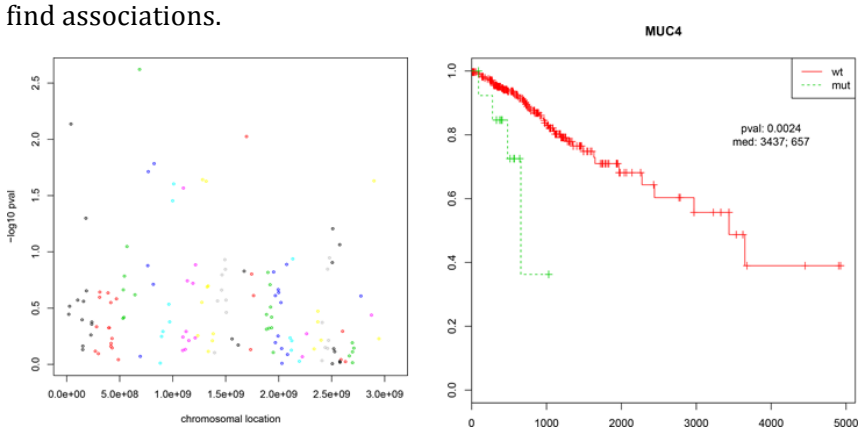
Background and preliminary data

Genetic markers associated with survival time are important because they can predict prognosis of cancer and some of them may be closely related to mutations that drive cancers to malignant states. So far, such genetic markers have been usually searched by, for example, clustering analysis of expression levels in cancer tissues to find subgroups as well as "signatures", or genes that are specifically expressed in subgroups. Then, these signatures are tested to be linked to survival time. However, this approach often fails to find genes linked to survival time for several reasons: for example, it fails to define clear subgroups or it simply does not find an association between subgroups and survival time.

A complementary approach to this is the approach in the opposite way, such as taken in genome-wide association studies (GWAS). GWAS typically examines whether alleles at each genetic locus in the genome are associated with cases and controls. Another type of GWAS examines whether alleles at each locus are associated with quantitative traits such as body weights. Here we propose an analogous approach that regards survival time as a quantitative trait and cancer-mutation status as allelic states in order to extensively search for genetic markers associated with survival time in a hypothesis-free form.

Preliminary data

The following figure illustrates this approach using several hundreds of patients in the current liver-cancer project led by T. Shibata. This is very preliminary data; so, we did not perform multiple testing corrections, did not consider combinations of genes, or did not check covariates or stratifications possibly caused by other clinical information such as age, gender, etc. Nevertheless, we found some genes correlated with survival time: *e.g.*, patients with mutations in the MUC4 gene tend to have significantly shorter survival time. We used the log-rank test for each gene to find associations.



Timelines & resources dedicated to project

Dec 2013 to Jan 2014: Preparation of structured data tables

Jan 2014 to Aug 2014: Uni-variate analysis with stratification in all cancer types
Aug 2014 to Dec 2014: Multi-variate analysis with stratification
Sep 2014 to Feb 2015: Manuscript preparation
Mar 2015: Manuscript submission

Mandatory data
Clinical information: survival time and censoring data
Genomic data: SNV/indel calls
Transcriptome data: RNA expressions of genes and (additionally of non-coding RNAs) in cancers and controls

Additionally necessary data
Clinical information: anything such as age, gender, grade, tumor size, ethnicity, etc.
Genomic data: CNA calls, rearrangement calls
Epi-genomic data: beta values in cancers and controls

Research proposal

We will take the following approach given that we get the mandatory data as well as the additional data indicated above.

We will first make structured data tables in which the rows represent patients and the columns represent survival time, censoring, other clinical attributes, and genes. Values for the gene columns represent the presence or absence of mutations for SNV/indel data. We will assign "presence" if a gene has at least one mutation (assuming dysfunction by this). For transcriptome data, we will just assign log2 expression ratios to genes. For rearrangements such as fusions, we will add such entities to the columns and their values represent their present or absence. For the CNA case, we will first find significant focal CNAs by GISTIC or another method and then add CNA columns whose values represent presence or absence. For the methylation case, we will assign to genes the log2 ratios of beta values between cancers and some controls. This formalization may be extended to pathway columns (values represent the presence or absence of mutations in pathways).

After formatting tables, we will go to the 1st screening, where we will search for factors associated with survival time in a uni-variate approach. For categorical variables such as the presence or absence of mutations, we will use the log-rank test for the Kaplan–Meier estimator. For continuous variables such as expression levels, we will try a semi-parametric or parametric model: Cox proportional hazards model or a parametric model assuming the Weibull distribution. We will consider stratifications caused by other clinical information. We will check validity of assumptions when we use a semi-parametric or parametric model. We will correct for multiple-testing corrections by FDR or by checking abnormalities in a Q-Q plot of –log p-values. We will make a Manhattan plot such as in the figure above.

After this 1st screening, we will get several candidates. We will first look into those genes and non-coding RNAs and CNAs/fusions from a biological aspect, consulting with literature and online databases. Then, we will look into their correlations and cluster them using clustering analysis such as a hierarchal clustering method to see if the significance is improved for a group of factors or for a representative factor.

Finally, we will list genes/CNAs/fusions found to be associated with survival time in every cancer type of the pan-cancer set and discuss about their biological meanings and about specificities/commonalities of markers between cancer types.

Legacy plans

We use a standard library for survival analysis in R. If the commands are so expanded that we think it worth packaging, we will pack them with documentations and make it publicly available.

Curriculum Vitae

CONTACT INFORMATION

Tatsuhiro Shibata, M.D., Ph.D.
Chief, Division of Cancer Genomics, National Cancer Center Research Institute
TEL: 03-3542-2511 (Ext. 3123), FAX: 03-3547-5137, E-mail: tashibat@ncc.go.jp

EDUCATION

1984-1990 University of Tokyo, School of Medicine
1990-1994 University of Tokyo, School of Medicine, Graduating School (Pathology)

EMPLOYMENT HISTORY

1992-1995 Research Fellow, Pathology Division, National Cancer Center Research Institute
1995-1998 Postdoctoral fellow, University of California, Irvine, Developmental Biology Center
1998-2003 Staff Scientist, Pathology Division, National Cancer Center Research Institute
2003-2005 Section Head, Pathology Division, National Cancer Center Research Institute
2005-2010 Project Leader, Cancer Genomics Project, National Cancer Center Research Institute
2010- Chief, Division of Cancer Genomics, National Cancer Center Research Institute

SELECTED PUBLICATIONS

1. Alexandrov LB, **Shibata T**, Stratton MR et al. Signatures of mutational processes in human cancer. **Nature**, 2013 500, 415-421.
2. International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, **Shibata T**, Lopez-Bigas N, et al. Computational approaches to identify functional genetic variants in cancer genomes. **Nature Methods**, 2013 10, 723-9
3. Fujimoto A, **Shibata T**, Nakagawa H, et al. Whole genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet**, 2012, 44:760-764.
4. Kohno T, **Shibata T**, et al. KIF5B-RET fusions in lung adenocarcinoma. **Nat Med**, 2012, 18:375-377.
5. Mitsuishi Y, **Shibata T**, et al. Nrf2 redirects glucose and glutamine into anabolic pathways in metabolic reprogramming. **Cancer Cell**, 2012, 22:66-79.
6. Totoki Y, **Shibata T**, et al. High-resolution characterization of a hepatocellular carcinoma genome. **Nat Genet**, 2011, 43:464-469.
7. Watanabe T, **Shibata T**, et al. Role for piRNAs and a novel non-coding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. **Science**, 2011, 332:848-52.

Curriculum Vitae

Mamoru Kato, Ph.D.
National Cancer Center, 5-1-1, Tsukiji, Chuuoo-ku, Tokyo 104-0045, Japan
E-mail: mamkato@ncc.go.jp, Tel: +81-3-3542-2511 (ext. 3123)

Education and Employment

• 2013/07 – Present	Head, Dept. of Bioinformatics, National Cancer Center, Japan
• 2012/06 – Present	Laboratory Head, Div. of Cancer Genomics, National Cancer Center, Japan
• 2009/04 – 2012/05	Computational Postdoctoral Fellow, Cold Spring Harbor Laboratory, USA
• 2004/04 – 2009/03	Research Scientist, RIKEN, Japan
• 2004/03	Ph.D. in Physics, University of Tokyo, Japan
• 2002/02 – 2003/08	Research Assistant, Cold Spring Harbor Laboratory, USA
• 2000/03	M.S. in Physics, University of Tokyo, Japan
• 1998/03	B.S. in Physics, Kyoto University, Japan

Specialties: bioinformatics, computational biology, population genetics

Selected Papers

- Y. Suenaga, S. M. R. Islam, J. Alagu, Y. Kaneko, M. Kato, ..., and A. Nakagawara. **NCYM, a de novo evolved protein, stabilizes MYCN and characterizes human neuroblastoma.** *PLoS Genetics* (accecpted).
- Mamoru Kato, Takahisa Kawaguchi, ..., Yusuke Nakamura, Hiroyuki Aburatani, and Tatsuhiko Tsunoda. **Population-genetic nature of copy number variations in the human genome.** *Human Molecular Genetics*, 2010, 19, 761-773.
- Mamoru Kato, Yusuke Nakamura, and Tatsuhiko Tsunoda. **An algorithm for inferring complex haplotypes in a region of copy number variation.** *The American Journal of Human Genetics*, 2008, 83, 157-169.
- Mamoru Kato, Yusuke Nakamura, and Tatsuhiko Tsunoda. **MOCSphaser: a haplotype inference tool from a mixture of copy number variation and single nucleotide polymorphism data.** *Bioinformatics*, 2008, 24, 1645-1646.
- Mamoru Kato, Fuyuki Miya, Yonehiro Kanemura, Toshihiro Tanaka, Yusuke Nakamura, and Tatsuhiko Tsunoda. **Recombination rates of genes expressed in human tissues.** *Human Molecular Genetics*, 2008, 17, 577-586.
- Mamoru Kato, Naoya Hata, Nilanjana Banerjee, Bruce Futcher, and Michael Q Zhang. **Identifying combinatorial regulation of transcription factors and binding motifs.** *Genome Biology*, 2004, 5, (R56) 1-13.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Association between mutational signatures and cancer progression

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Yasushi Totoki, National Cancer Center, Japan, ICGC
Tatsuhiro Shibata, National Cancer Center, Japan, ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

Cancer progression can be defined using tumor variant allele frequency. Mutations with high allele frequency are major clonal mutations and were acquired in the early time of cancer progression. In contrast, mutations with low allele frequency are subclonal mutations and were acquired in the late time of cancer progression. We already have analyzed whole exome sequence of liver cancer which we sequenced and found that there are some differences of mutational signatures among genes with distinct variant allele frequencies. Since the number of mutations in whole exome sequence is small, we will expand this analysis to whole genome sequence of liver cancer which Japanese group sequenced and enhance the result. However, the question if there are some differences among cancer types will be remained. We would like to analyze mutational signatures of various cancer types using pan-cancer data and make sure if there are some associations between mutational signatures and cancer progression, and if there are some differences among cancer types.

Timelines & resources dedicated to project

When somatic mutation data of pan-cancer is prepared (until September 2014), we start the analysis. The time which we need to the analysis is about three month (until December 2014).

Research proposal

Using somatic mutation data of pan-cancer, we classify mutations according to tumor variant allele frequency (TVAF) in each cancer type and compare substitution pattern. TVAF is adjusted by tumor purity which is estimated from distribution of TVAF in each sample. We analyze substitution pattern with bases immediately 5' and 3' to each substitution (96-substitution pattern) and also strand bias of 96-substitution pattern. Using principal component analysis (PCA) and non-negative matrix factorization (NMF), we try to extract mutational signatures specific to cancer progression. Although the sequence depth of whole genome sequence is low, we think that we can extract some correlation, if any, since there are an enormous number of mutations. We would like to use whole exome sequence data whose sequence depth is high if necessary.

Legacy plans
We do not have to make new software since we will use the published software and the program already developed in house for mutational signature analysis (ex. NMF and PCA).

Curriculum Vitae

CONTACT INFORMATION

Yasushi Totoki
Section Head, Division of Cancer Genomics
National Cancer Center Research Institute
TEL: 03-3542-2511 (Ext. 3121), FAX: 03-3547-5137
E-mail: ytotoki@ncc.go.jp

EDUCATION

1983-1987 Nagoya University, School of Science, Department of Physics

EMPLOYMENT HISTORY

1990-1999 Information and Mathematical Science Laboratory, Inc
1999-2008 Senior Scientist, RIKEN Genomic Sciences Center
2008-2008 Deputy Team Leader, MetaSystems Research Team, RIKEN Advanced Science Institute
2008-2010 Senior Staff Scientist, Cancer Genomics Project, National Cancer Center Research Institute
2010~ Section Head, Division of Cancer Genomics, National Cancer Center Research Institute

SELECTED PUBLICATIONS

1. Alexandrov LB, **Totoki Y**, Stratton MR et al. Signatures of mutational processes in human cancer. **Nature**, 2013 500, 415-421.
2. Fujimoto A, **Totoki Y**, Nakagawa H et al. Whole genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet**, 2012, 44:760-764.
3. Kohno T, **Totoki Y**, Shibata T et al. KIF5B-RET fusions in lung adenocarcinoma. **Nat Med**, 2012, 18:375-377.
4. **Totoki Y**, Shibata T et al. High-resolution characterization of a hepatocellular carcinoma genome. **Nat Genet**, 2011, 43:464-469.
5. Watanabe T, **Totoki Y**, Sasaki H et al. Role for piRNAs and a novel non-coding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. **Science**, 2011, 332:848-52.
6. International Cancer Genome Consortium. International network of cancer genome projects. **Nature**, 2010, 464:993-8.
7. Watanabe T, **Totoki Y**, Sasaki H et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. **Nature**, 2008, 453:539-43.
8. Taylor TD, **Totoki Y**, Sakaki Y et al. Human chromosome 11 DNA sequence and analysis including novel gene identification. **Nature**, 2006, 440:497-500.
9. Borowsky ML, **Totoki Y**, Lander ES et al. DNA sequence and analysis of human chromosome 18. **Nature**, 2005, 437:551-5.
10. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. **Nature**, 2004, 431:931-45.
11. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. **Nature**, 2001, 409:860-921.
12. Hattori M, **Totoki Y**, Sakaki Y et al. The DNA sequence of human chromosome 21. **Nature**, 2000, 405:311-9.

Curriculum Vitae

CONTACT INFORMATION

Tatsuhiro Shibata, M.D., Ph.D.
Chief, Division of Cancer Genomics
National Cancer Center Research Institute
TEL: 03-3542-2511 (Ext. 3123), FAX: 03-3547-5137
E-mail: tashibat@ncc.go.jp

EDUCATION

1984-1990 University of Tokyo, School of Medicine
1990-1994 University of Tokyo, School of Medicine, Graduating School (Pathology)

EMPLOYMENT HISTORY

1992-1995 Research Fellow, Pathology Division, National Cancer Center Research Institute
1995-1998 Postdoctoral fellow, University of California, Irvine, Developmental Biology Center
1998-2003 Staff Scientist, Pathology Division, National Cancer Center Research Institute
2003-2005 Section Head, Pathology Division, National Cancer Center Research Institute
2005-2010 Project Leader, Cancer Genomics Project, National Cancer Center Research Institute
2010~ Chief, Division of Cancer Genomics, National Cancer Center Research Institute

AWARDS and FELLOWSHIP

Uehara Memorial Research Fellowship (1996)
Incitement Award of the Japanese Cancer Association (2005, Japan Cancer Association)
Tamiya Award (2008, National Cancer Center)
Research Award (2011, The Japanese Society of Pathology)

LICENSES / CERTIFICATION

Medical License (1990)
Certified Board of Pathology (2001)

SELECTED PUBLICATIONS

1. Alexandrov LB, **Shibata T**, Stratton MR et al. Signatures of mutational processes in human cancer. **Nature**, 2013 500, 415-421.
2. International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, **Shibata T**, Lopez-Bigas N, et al. Computational approaches to identify functional genetic variants in cancer genomes. **Nature Methods**, 2013 10, 723-9
3. Fujimoto A, **Shibata T**, Nakagawa H, et al. Whole genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. **Nat Genet**, 2012, 44:760-764.
4. Kohno T, **Shibata T**, et al. KIF5B-RET fusions in lung adenocarcinoma. **Nat Med**, 2012, 18:375-377.
5. Mitsuishi Y, **Shibata T**, et al. Nrf2 redirects glucose and glutamine into anabolic pathways in metabolic reprogramming. **Cancer Cell**, 2012, 22:66-79.
6. Totoki Y, **Shibata T**, et al. High-resolution characterization of a hepatocellular carcinoma genome. **Nat Genet**, 2011, 43:464-469.
7. Watanabe T, **Shibata T**, et al. Role for piRNAs and a novel non-coding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. **Science**, 2011, 332:848-52.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

A comprehensive evaluation of mutations in micro-RNA genes, promoter elements, and target sites across multiple cancers and cancer sub-types

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Tatsuhiko Tsunoda (RIKEN, IMS, Japan)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Todd A. Johnson (RIKEN, IMS, Japan)	

Background and preliminary data

Micro-RNAs (miRNA) have been identified as potential dysregulators of cancer suppressor genes or oncogenes. By relaxing suppression of cell growth or increasing cell proliferation, dysregulation of normal miRNA functions may lead to neoplasia or tumor differentiation into sub-types. One typical method of identifying cancer associated miRNA genes has been differential gene expression analysis of neoplastic versus normal tissues or between different cancer/patient classifications. Follow-up analyses may then search for the underlying genetic abnormalities that led to a differentially expressed miRNA gene. Some reports have identified somatic copy-number changes as responsible for particular associated miRNA genes. For example, a deletion at chr13q14 in a large proportion of CLL patients removes miR-15/16, thus abrogating its normal function to suppress BCL2 expression. Besides larger structural genomic abnormalities, somatic single-base substitutions within an miRNA gene potentially could disrupt its binding site and/or the hairpin structure that is necessary for proper processing into a mature miRNA. However, studies that have searched for such mutations in limited tumor types have been inconclusive as to whether such changes play a main role in miRNA dysregulation. In addition to somatic changes involving miRNA genes, reports have also found that miRNA target-sites in the 3'-UTR of protein-coding genes may be disrupted/modified through mutation. Similarly, somatic changes may also target miRNA gene regulatory elements such as promoters, but besides small, single disease studies, there have been no comprehensive examination of these potentially important regulatory regions. Recent datasets from the ENCODE and FANTOM4 projects such as deepCAGE and epigenetic data can be used to predict miRNA promoter elements across the genome using programs such as the newly released PROmiRNA (Genome Biol, 2013, Vol.14, R84). Using the current release of the miRNA database (miRbase Ver. 20; 1,871 genes) with ENSEMBL Rel. 73 genomic features, we predicted 9,827 promoter elements for 1,472 miRNA genes. This dataset should help quantify the contribution of somatic SV and point mutations to dysregulation of miRNA function across multiple cancer types.

Timelines & resources dedicated to project

The primary IGCG resource used for this analysis will be the variant calls from cancer tissues, although in follow-up analyses, we will examine germline variant calls in any miRNA genes/promoters/target-sites that we find to be enriched across multiple cancer types. If datasets are available and techniques allow, we will also examine the RNA-Seq data for whether expression of miRNAs are affected by mutations in miRNA promoter elements. That analysis could alternatively utilize any other limited datasets such as in TCGA that might have interrogated miRNA expression levels. As the primary analyses make use of existing variant calls, computational resource use should be limited or restricted to our own local computers, and the first stage analysis of cancer mutations would likely take between two to four months to finish.

Research proposal

For this research, we propose to identify miRNA related genomic elements harboring excessive somatic mutations (both structural and point mutations) across multiple cancer types/sub-types, and then place any positive findings in the context of miRNA/transcription-factor(TF)/gene interaction networks. Our research will be directed at quantifying mutations in three general categories of miRNA related genomic elements using the Pan-cancer dataset: 1) miRNA genes (mutations that impact miRNA hairpin structure or stability of binding site), 2) miRNA promoter elements and associated TFBS, 3) miRNA target sites in protein-coding genes. We will extract the dataset of hairpin/mature sequences from the current version of miRBase (Ver. 20 = 1,871 miRNA genes). To predict miRNA promoters, we use PROmiRNA (http://promirna.molgen.mpg.de/) with the set of miRNA genes and the latest genomic feature annotations from ENSEMBL and deepCAGE data, while more likely causal mutations will be extracted that intersect TF binding sites or predicted TF motifs. TF predictions will likely come from ENCODE and/or the TRANSFAC database, but as annotation of genomic features (epigenomic, TF binding sites, etc.) is continually advancing, we will make every effort to have our pipeline flexible enough to include new foundation datasets as the project advances. Our current miRNA promoter annotation pipeline utilizes deepCAGE data originally produced by FANTOM4, but it is likely that new FANTOM5 data (which should significantly increase the number of tissues and celltypes that can be used for prediction) will be available during the course of this project. We will obtain predicted miRNA target sites from one or more of the common prediction databases such as TargetScan, EIMMO, although new methods have proposed predictions based on biophysical modeling, which may prove more accurate compared to other methods. In addition, we will utilize other online tools such as staRBase (http://starbase.sysu.edu.cn/) that identify higher confidence target sites from multiple prediction methods that overlap experimentally derived Chip-Seq RNA binding analyses for argonaute proteins. For any positive miRNA genes, we will determine putative target genes and evaluate their validity based on gene expression levels from RNA-Seq data if that is feasible based on available data.

Legacy plans

We plan on releasing our analysis pipeline methods and the database of mutated miRNA related elements using our internal web-server (http://emu.src.riken.jp/) after publication of the results and/or algorithms.

	Name:	Tatsuhiko Tsunoda, Dr. Ph.D. (Medicine) & Ph.D. (Engineering) Group Director
Affiliation:	Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Address:	1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan Phone: +81-45-503-9556 Email: tsunoda@src.riken.jp	
Major Field:	Medical science mathematics, genome medicine, statistical genetics, and cancer genome analysis	
Professional history:	1985-1989: B.S., Department of Physics, Faculty of Science, the University of Tokyo 1989-1991: M.S., Department of Physics (Elementary Particle Physics), the University of Tokyo 1992-1995: Ph.D., Department of Engineering, the University of Tokyo 1995-1997: Assistant Professor, Department of Engineering, Kyoto University 1997-1998: Research Associate, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 1998-2000: Assistant Professor, the Human Genome Center, the Institute of Medical Science, the University of Tokyo 2000-2013: Laboratory Head, Laboratory for Medical Informatics, RIKEN SNP Research Center (2008-present: RIKEN Center for Genomic Medicine) 2011-2013: Director, Research Group for Medical Informatics, RIKEN Center for Genomic Medicine 2012-present: Visiting Professor, the Institute of Statistical Mathematics 2013-present: Director, Chief Scientist, Research Group (Laboratory) for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences	
Membership and councilor	Councilor of the Japanese Cancer Association Councilor of the Japanese Society of Human Genetics Associate Editor of Cancer Science Journal Associate Editor of the Journal of Human Genetics Member of American Society of Human Genetics.	

Curriculum Vitae
Todd A. Johnson, Ph.D.

Contact Information:

Residence: Ookawa 7-10-908, Kanazawa-ku
Yokohama, Kanagawa-ken 236-0043 JAPAN
Home Phone: +81-45-783-2604
Cell Phone: +81-90-5309-5867

Laboratory: RIKEN, Yokohama Institute
Center for Integrative Medical Sciences
Laboratory for Medical Science Mathematics
Suehiro-cho 1-7-22, Tsurumi-ku
Yokohama, Kanagawa-ken 230-0045 JAPAN
e-mail: tjohnson@src.riken.jp/bbnmore@pair.com

Education:

2005-2010: Ph.D. in Medical Science (Bioinformatics)
Graduate school of Tokyo Medical and Dental University
Department of Bioinformatics, Medical Research Institute

1985-1990: B.A. Biochemistry/Cell-Biology
(1988-89, UC Education Abroad Program, Georg-August-Universität Göttingen, Germany)
University of California, San Diego

Work experience:

2004-Present: Researcher
Laboratory for Medical Science Mathematics
RIKEN, Center for Integrative Medical Sciences (prev. Center for Genomic Medicine)
Research in the fields of genome-wide association studies for analysis of diseases and quantitative traits, population genetics, statistical genetics, software development, and immunology using self-developed software implemented on high-performance compute clusters.

2001-2004: Programmer Analyst
Laboratory of Dr. Joel E. Dimsdale
UCSD, Department of Psychiatry
Development of systems for integration and management of laboratory databases, hardware, software, and network systems for psychiatric clinical research.

1990-2001: Research Associate
Laboratory of Dr. Thomas J. Kipps
Chronic Lymphocytic Leukemia Research Consortium
UCSD, Department of Medicine, Hematology/Oncology Division
Research in the fields of immunology, molecular biology, and chronic lymphocytic leukemia. Conducted immunological assays using CLL patients from investigator-sponsored clinical trials and development of data management system.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of long-non coding RNA expression patterns and its correlation with structural aberrations in cancer genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Alfredo Hidalgo Miranda, Instituto Nacional de Medicina Genomica, Mexico. Mexico-US breast cancer project PI, member of the ICGC Scientific Steering Commite

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Sandra Lorena Romero-Cordoba, Rosa Rebollar-Vega, Instituto Nacional de Medicina Genomica, Mexico	**Claudia Rangel-Escareño, Enrique Hernandez-Lemus Instituto Nacional de Medicina Genomica, Mexico**

Background and preliminary data

Modern high-throughput genomic experiments have paved the way to a large-scale characterization of living organisms. This usually involves the generation and, more importantly, the interpretation of data at an unprecedented scale. Computational tools and mathematical algorithms have been created in order to integrate, organize and also mine the gargantuan wealth of information generated. Technologies for the detection of several kinds of genomic alterations have been developed and applied to analyses of almost any living organism, even to cancer genomes. Cancer research in particular, has proven that studies based on a single technology platform result extremely limited in scope when compared with the extent of knowledge that can be acquired when using different platforms all together. For this reason, there is a need for systematic methodologies to facilitate data management, visualization and integration. The goal of these methodologies should be to permit a proper analysis of the biological implications for the findings, without sacrificing mathematical and statistical rigour and computational efficiency. Development of such integrated analytical methods will help to address a wide range of biological questions. In this proposal, we will focus on the analysis of the expression of Long non-coding RNAs and their correlation with other somatic genomic alterations in cancer. LncRNAs play important roles in cell physiology and alteration in their expression has been associated to several pathological processes, including cancer. Aberrant expression of lncRNAs in human cancer has been related with tumor development, recurrence of disease, metastasis and prognosis. Similar to messenger RNA, lncRNA expression patterns have also been used as tools to identify tumor specific expression signatures. LncRNA expression has been carried out through the analysis of RNAseq data from several tumors, including some of the tumors included in the TCGA (ovarian and glioblastoma). These analyses have identified correlations between changes in lncRNA expression and somatic DNA copy number aberrations, as well as specific signatures associated with relevant clinical features. Recent analysis of transcriptome data across diverse cancer types has also revealed a set of microRNAs that regulate the expression of genes in cancer related pathways. However, there is no information in the literature about the integrated analysis of somatic DNA copy number changes, point mutations and their impact on lncRNA expression across diverse cancer types.

Timelines & resources dedicated to project

First trimester 2014. Data acquisition, conversion of BAM files to FASTQ and re-alignment to the Human Body Map lncRNA catalog. Second trimester: Normalization of expression data and identification of tumor specific and pan-cancer lncRNA expression signatures. Third trimester: Correlation analysis between somatic DNA copy number changes, point mutations, DNA methylation and changes in lncRNA expression. Fourth semester: validation of lncRNA expression profiles on an independent sample set from a selected tumor (breast) through microarray analysis.

Resources devoted to the project: three post-graduate students and four staff scientists from the computational genomics department of the National Institute of Genomic Medicine in Mexico City will participate in data analysis and discussion. The resources needed for sample collection and processing, as well as microarray analysis of validation of lncRNA expression patterns in breast tumors will be covered by the proponent's lab.

Research proposal
With view to the construction of an integral view of genomic alterations, a data driven combinatorial approach has been proposed. It is based on the enumeration of all possible genomic alterations scenarios that may be present in a N-platform integrative analysis relying on a so- called three-state model applied to the set of statistically significant genes. Each scenario is represented as a sequence of states, where Sk denotes the state of a gene for platform k. Each state is defined to take values coding for (Down, NoChange, Up). This list represents the universe of hypotheses that describe structural variations in the genome as well as transcription activity in coding and non-coding regions. Hypotheses can be chosen for their clear biological relevance but also for their quantitative importance. We may find that a large set of genes follow a particular scenario or that genes commonly share a set of more specific scenarios leading to other important questions to be answered. As an example of this kind of integrative analysis, we aim to define tumor specific and pan-cancer lncRNA expression patterns in diverse human tumor types, and their potential relation with somatic DNA copy number aberrations and mutation profiles, we propose to analyze the RNAseq data from the ICGC through re-alignment of the FASTQ converted BAM files to the current version of the Human Body Map lncRNA catalog. Read counts will be normalized and compared between diverse tumor types in order to define tumor specific signatures, and to identify lncRNAs whose expression is common between diverse cancer types. Correlation between somatic DNA copy number aberrations, point mutation, DNA methylation and the expression of lncRNAs will be explored in the cases where data from the relevant platforms is available. LncRNA expression profiles will also be tested as a classification tool to identify specific tumors and tumor subtypes within selected cancers (PE breast and ovary), compared to messenger RNA expression profiles. To further explore the findings derived from the RNAseq analysis, and in order to compare the data generated through NGS with a microarray based platform, we will analyze the messenger RNA and lncRNA expression data of an independent set of breast tumors with the Affymetrix Human Transcriptome 2 array. As a result of this proposal, we will generate: **A catalog of lncRNA expression across diverse human cancer types** **lncRNA expression signatures associated to specific tumor types** **A pan-cancer lncRNA expression signature** **Insights into the correlation between changes in lncRNA expression profiles and several types of somatic aberrations in human tumors, including DNA copy number changes, point mutation and DNA methylation.**
Legacy plans
All the tools and results generated from the proposed analyses will be available to the research community

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Pan-cancer IMAGE.

(**I**dentification and **M**apping of **A**ctionable **GE**notypes)

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Andrew V. Biankin (Australian Pancreatic Cancer Genome Initiative); TCGA Pancreas
Regius Professor of Surgery
Director, Wolfson Wohl Cancer Research Centre,
University of Glasgow
Garscube Estate, Switchback Road, Bearsden,
Glasgow Scotland G61 1BD, United Kingdom
Ph: +44 141 330 5670 (direct)
Email: **Andrew.Biankin@glasgow.ac.uk**

Sean M. Grimmond (ICGC Australia); TCGA Pancreas
Professor of Medical Genomics
Wolfson Wohl Cancer Research Centre,
University of Glasgow
Garscube Estate, Switchback Road, Bearsden,
Glasgow Scotland G61 1BD, United Kingdom
Ph: +44 141 330
Email: **Sean.Grimmond@glasgow.ac.uk**

Name(s) & institute(s) of junior investigators

(Name no more than 2; append 1 page CV for each)

David K. Chang
Clinical Senior Lecturer
Wolfson Wohl Cancer Research Centre,
University of Glasgow
Garscube Estate, Switchback Road, Bearsden,
Glasgow Scotland G61 1BD
United Kingdom
Ph: +44 141 330 5834
Email: **David.Chang@glasgow.ac.uk**

Peter Bailey
Queensland Centre for Medical Genomics and
Wolfson Wohl Cancer Research Centre,
University of Glasgow
Garscube Estate, Switchback Road, Bearsden,
Glasgow Scotland G61 1BD
United Kingdom
Email: p.bailey@imb.uq.edu.au

Name(s) & institute(s) of non-ICGC collaborators

(Name no more than 2; append 1 page CV for each)

Oliver Hofmann
Associate Director, Bioinformatics Core,
Harvard School of Public Health.
Senior Research Scientist,
Harvard School of Public Health.
Affiliated Faculty, Harvard Stem Cell Institute
655 Huntington Ave, Boston, MA 02115, USA
Phone: +1 617 365 0984
Email: ohofmann72@gmail.com

Background and preliminary data

Over the last two decades, a deeper understanding of the genetic and molecular basis of cancer has led to therapies that selectively target molecular mechanisms of specific importance to cancer cells. Clinical benefit from these therapies is dependent on the presence of specific cellular targets and is optimally directed by the presence of a biomarker. Large-scale genomics and other "-omic" technologies are providing opportunities to improve current approaches to cancer therapy, however, successful translation into improvements in patient care requires a fundamental shift in clinical oncology to include a molecular taxonomy, where individual cancers are grouped and selected for optimal therapy depending on their molecular phenotype (genotype). As a consequence, the **overall aim** of this proposal is to map therapeutically actionable phenotypes (genotypes) across all cancers types within the ICGC pan-cancer analysis. Defining the prevalence of actionable molecular phenotypes (genotypes) across many cancer types will identify opportunities for rationalising, rescuing and repurposing existing therapies, and enhancing the development of emerging therapeutics. In addition, defining actionable genotypes across cancer types will define optimal clinical trial strategies to test these approaches in the clinic.

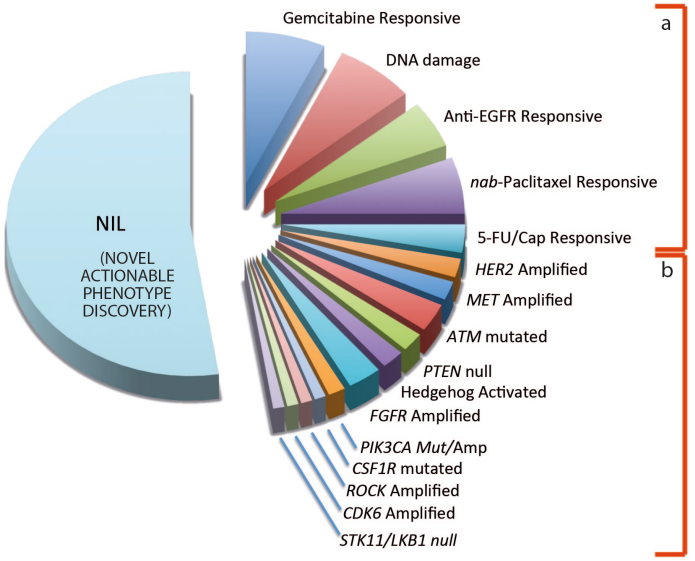


Figure 1: Actionable Molecular Phenotypes (Genotypes) of pancreatic cancer. Analysis of current data from the Australian Pancreatic Cancer Genome Initiative has allowed us to map actionable molecular phenotypes (genotypes) in detail for a) current therapies that do not have a predictive biomarker, and b) opportunities for rescuing and repurposing therapeutics where the molecular target is present in PC.

We have mapped actionable phenotypes in pancreatic cancer that we have sequenced and uploaded to the ICGC portal (>100 WGS + >100 WES, all with CNV, mRNA expression and methylation data) (Figure 1). Whilst the higher prevalence of some makes clinical testing feasible, lower prevalence phenotypes are unlikely to recruit sufficient patients for clinical trials using current approaches, nor interest industry in pursuing indication extension studies. Importantly, although these genotypes are often individually small, cumulatively they account for a significant proportion of patients. Some may be tested using emerging clinical trial strategies that use an efficient "umbrella" approach, others may be optimally tested by recruiting patients based on genotype irrespective of organ of origin (basket studies). As a consequence, defining actionable genotypes across all cancers within the ICGC pan-cancer project will identify opportunities for therapeutic development.

Timelines & resources dedicated to project

The planned analyses build on the essential analyses of the Pan-cancer initiative. Actionable genotypes can be defined based on a variety of genomic events, and combinations of these. As a consequence, we will liaise with groups performing the essential analyses so that we can mine these data for actionable genotypes that confer sensitivity or resistance to specific therapies.

It is envisaged that as essential analyses are completed (December 2013 to September 2014) we will define actionable phenotypes as these datasets are made available, which will provide sufficient time before manuscript preparation and submission in march 2015.

The resources available are the informatics team of Australia's ICGC contribution, primarily the Australian Pancreatic Cancer Genome Initiative based in Australia, and the team developing in Glasgow, which will be comprised of 8 informaticians at the commencement of this work. Biankin and Grimmond have established the Howat Cancer Genomics Facility at the Wolfson Wohl Cancer Research Centre in Glasgow which will be fully equipped for cancer genomic sequencing and analysis. In addition, further analysis will be performed in collaboration with Oliver Hofmann, Associate Director of the Bioinformatics Core at the Harvard School of Public Health.

Research proposal

The proposal builds on the essential analyses performed by the Pan-cancer initiative, with data summaries generated by those teams used to map actionable genotypes.

Actionable molecular phenotypes for cancer therapy can be broadly classified into 5 groups:

1. *Clinical trial evidence that a therapeutic is effective when the actionable molecular phenotype is present in that cancer type (e.g. HER2 amplification and trastuzumab therapy in breast and gastric cancer, BRAF V600E mutation and vemurafenib therapy in melanoma).*
2. *Those where clinical trials have shown incremental overall benefit, but often with significant responses in subgroups that are not well defined (e.g. platinum based therapy).*
3. *Those that have failed in clinical trials, but with subgroups of responders and supportive preclinical data (e.g. trastuzumab and hedgehog inhibitors in pancreatic cancer).*
4. *Opportunities for repurposing therapies used in other cancers or other diseases, and for early indication extension for emerging therapies (e.g. mTOR inhibitors)..*
5. *Preclinical evidence and computational modelling to define therapeutic susceptibility based on dominant pathways, mechanisms, or surrogate measures of these processes.*

Actionable genotypes can be defined based on a number of genomic events including SNV (eg: *EGFR* mutations), CNV (eg: *HER2* amplification), SV (eg: *BCR-ABL* gene fusions) and surrogates of processes that putatively confer sensitivity to specific drugs (eg: *BRCA* mutational signature). In addition, some drug susceptibility phenotypes require examination of the germ line (eg: biallelic *BRCA* inactivation).

An important initial step is to verify the diagnosis of cancers uploaded to the ICGC portal by identifying inconsistencies and outliers on variant analyses to prompt histopathological and clinical reassessment. Histopathological and clinical diagnoses are imperfect, and may significantly impact on the clinical application of genomics in the future. For example, a UV light mutational signature in a liver lesion, suggests a lesion of cutaneous origin and may alter the diagnosis in light of the clinical scenario.

To map actionable phenotypes (genotypes) across cancer types, we will:

1. Define the overall and organ specific prevalence of ***well-characterized genomic events that have been proven as effective targets and/or biomarkers of therapeutic responsiveness in clinical trials*** *of any individual cancer type. These will be grouped both here, and subsequently into 3 categories: a) Specific events that are associated with therapeutic response (eg: BRAF pV300E mutations), b) those where experimental data, but not clinical data support functional and probable therapeutic relevance, and c) remaining genomic events with predicted functional consequences.*

2. Define the overall and organ specific prevalence of ***genomic events that are hypothesized as biomarkers of therapeutic responsiveness for existing and emerging therapies*** *that are as yet not clinically proven, but are in clinical trials.*

3. Define the overall and organ specific prevalence of ***genomic events with supportive preclinical evidence as candidate therapeutic targets****.*

4. Computationally modeling to define susceptibility to specific therapies using surrogates of genomic events defined above. Eg: Genomic or expression array signatures that are associated with the events above, but the specific genomic event is not present.

Legacy plans
Searchable database and criteria for actionable phenotypes will be made publicly available, which may be built on in the future to expand actionable phenotypes as therapies emerge. Executable code that is sufficiently well documented to enable replication by third parties will be made available for software developed through this project.

Andrew V. Biankin

CURRENT POSITIONS

Regius Professor of Surgery, University of Glasgow.
Director, Wolfson Wohl Cancer Research Centre, University of Glasgow.
Head, Pancreatic Cancer research, Garvan Institute of Medical Research,
Professor, Conjoint Appointee University of New South Wales

QUALIFICATIONS

1988	B. Med. Sc.	University of New South Wales
1992	M.B.,B.S. (HONS)	University of New South Wales
1999	F.R.A.C.S.	Royal Australasian College of Surgeons
2003	Ph. D.	University of New South Wales
2011	F.F.S (RCPA)	Royal College of Pathologists of Australasia
2012	F.R.C.S. (Glasg.)	Royal College of Physicians and Surgeons of Glasgow
2013	F.R.C.S. (Edin.)	Royal College of Surgeons of Edinburgh

HONOURS and AWARDS (selected)

2012	Cancer Institute NSW Wildfire Award
2010	Landon Foundation-AACR INNOVATOR Award
2008	Hirshberg Award for Pancreatic Cancer, American Pancreatic Association
2007	Cancer Institute NSW Premier's Award for Outstanding Cancer Research Fellow
2005	Cure Cancer Australia Young Researcher of the Year (Open Division)
2004	Excellence in Translational Research Award, Johns Hopkins University
2003	Garvan Institute of Medical Research, Thesis Prize

RESEARCH INTERESTS

Biankin's primary scientific focus is on the molecular pathology of pancreatic cancer, the development of early detection and novel therapeutic strategies based on molecular phenotyping and the delineation and implementation of biomarkers that facilitate clinical decision-making. He contributes to the International Cancer Genome Consortium through extensively characterising the genomic, transcriptomic and epigenomic aberrations in pancreatic cancer, and is extending this knowledge to a personalized model of cancer care, where molecular characteristics guide treatment decisions.

PREVIOUS APPOINTMENTS

2005 – 2014	Head, Pancreatic Cancer Research, *The* Kinghorn Cancer Centre, Cancer Research Program, Garvan Institute of Medical Research
2005 – 2013	Consultant HPB and Upper GI Surgeon, Sydney South West Area Health.
2005 – 2014	Chairman, Australian Pancreatic Cancer Network
2009 – 2014	Clinical Lead, Australian Pancreatic Cancer Genome Initiative (ICGC).
2006 – 2012	Chairman Bankstown Hospital Multidisciplinary Team (GI Oncology).

SELECTED RECENT PUBLICATIONS *(from a total of 100)*

1. Chang DK, Johns A, … Kench JG and **Biankin AV**. (2009) Margin clearance and outcome in resected pancreatic cancer. ***J Clin Oncol*** 27: 2855-2862
2. **Biankin AV,** Waddell N, … Pearson JV, McPherson JD, Gibbs RA, Grimmond SM. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. ***Nature***. 2012;491:399-405
3. Chang DK, Colvin EK, Johns A, … Kench JG and **Biankin AV**. Histomolecular Phenotypes and Outcome in Adenocarcinoma of the Ampulla of Vater**.** ***J Clin Oncol.*** 2013 31:1348-56
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S, Behjati S, **Biankin AV**, … Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. ***Nature*** 2013 500:415-21.
5. Chou A, Waddell N, Cowley MJ, Gill AJ, Chang DK, Patch AM, Nones K, Wu J, Pinese M, Johns AL, Miller DK, Kassahn KS, … Grimmond SM, **Biankin AV**. Clinical and molecular characterization of HER2 amplified pancreatic cancer. ***Genome Med***. 2013 Aug 31;5:78.

International Cancer Genome Consortium

Sean Michael GRIMMOND

Current Poitions: Chair of Medical Genomics,
Director, Howat Canceer Genomics Facility
Wolfson Wohl Cancer Research Centre,
University of Glasgow, Garscube Estate,
Switchback Road, Bearsden,
Glasgow Scotland G61 1BD
United Kingdom.

Education and Degrees:

2011 Founding Fellow (Faculty of Science), Royal College of Pathologists of Australasia.
1994 PhD (University of Queensland
1987 B.Sc. Hons University of New England

Awards & Scholarships

2013 Royal Society Wolfson Merit Award
2012 NH&MRC Principal Research Fellowship
2011 Julian Wells Medal for Transcriptomics
2007 NH&MRC Senior Research Fellowship (SRFA)
2004 Eppendorf Australian Genomics Research Medal
2002 NH&MRC Career Development Fellowship
1997 NH&MRC CJ Martin Travelling Fellow

Selected publications

Andrew V. Biankin, Nicola Waddell, Karin S. Kassahn, Marie-Claude Gingras, Amber L. Johns, David K. Miller, Peter J. Wilson, Ann-Marie Patch, Jianmin Wu, David K. Chang[3], Mark J. Cowley, Brooke B. Gardiner, Sarah Song, Ivon Harliwong, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Shivangi Wani, Milena Gongora, Marina Pajic, Christopher J. Scarlett, Anthony J. Gill, Elizabeth A. Musgrove, Robert L. Sutherland, Andreia V. Pinho, Ilse Rooman, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor Scott Wood, Christina Xu, Katia Nones, J. Lynn Fink, Angelika Christ, Tim Bruxner, Nicole Cloonan, Gabriel Kolle, Felicity Newell, Mark Pinese, Scott Mead, Jeremy L. Humphris, Warren Kaplan, Marc D. Jones, Emily K. Colvin, Adnan M. Nagrial, Emily S. Humphrey Angela Chou, Venessa T. Chin, Lorraine A. Chantrill. Jaswinder S. Samra, James G. Kench, Jessica A. Lovell, Roger J. Daly, Neil D. Merrett, Christopher Toon, Krishna Epari, Nam Q. Nguyen, Andrew Barbour, Nikolajs Zeps, Australian Pancreatic Cancer Genome Initiative, Nipun Kakkar, Fengmei Zhao, Yuan Qing Wu, Min Wang, Donna M. Muzny, William E. Fisher, F. Charles Brunicardi, Sally E. Hodges, Jennifer Drummond, Kyle Chang, Yi Han, Lora L. Lewis, Huyen Dinh, Christian J. Buhay, Lakshmi Muthuswamy, Timothy Beck, Lee Timms, Michelle Sam, Kimberly Begley, Andrew Brown, Deepa Pai, Ami Panchal, Nicholas Buchner, Richard De Borja, Robert E. Denroche, Christina K. Yung, Stefano Serra, Nicole Onetto, Debabrata Mukhopadhyay, Ming-Sound Tsao, Patricia A Shaw, Gloria Petersen, Steven Gallinger, Lincoln D. Stein, Ralph H. Hruban, Anirban Maitra, Christine A. Iacobuzio-Donahue Richard D. Schulick, Christopher L. Wolfgang, Richard A. Morgan, Rita T. Lawlor, Stefania Beghell, Vincenzo Corbo, Maria Scardoni, Claudio Bassi, Margaret A. Tempero, Karen M. Mann, Nancy A. Jenkins, Pedro A. Perez-Mancera, David J. Adams, David A. Largaespada, Lodewyk F. Wessels, Alistair G. Rust, David A. Tuveson, Neal G. Copeland, Thomas J. Hudson, Aldo Scarpa, James R. Eshleman, David A. Wheeler, John V. Pearson, John D. McPherson, Richard A. Gibbs and Sean M. Grimmond (2012) Genomic Analysis Reveals Roles for Chromatin Modification and Axon Guidance in Pancreatic Cancer. ***Nature, epub 24th Oct, 2012.***

Pedro A. Pérez-Mancera, Alistair G. Rust, Louise van der Weyden, Glen Kristiansen, Allen Li, Aaron L. Sarver, Kevin A. T. Silverstein, Robert Grützmann, Daniela Aust, Petra Rümmele, Thomas Knösel, Colin Herd, Derek L. Stemple, Ross Kettleborough, Jacqueline A. Brosnan, Ang Li, Richard Morgan, Spencer Knight, Jun Yu, Shane Stegeman, Lara S. Collier, Jelle J. ten Hoeve, Jeroen de Ridder, Alison P. Klein, Michael Goggins, Ralph H. Hruban, David K. Chang, Andrew V. Biankin, **Sean M. Grimmond**, Lodewyk F. A. Wessels, Stephen A. Wood, Christine A. Iacobuzio-Donahue, Christian Pilarsky, David A. Largaespada, David J. Adams, David A. Tuveson (2012) The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. ***Nature****, 486, 266–270*.

Karen M. Mann, Jerrold M. Ward, Christopher Chin Kuan Yew, Anne Kovochich, David W. Dawson, Michael A. Black, Benjamin T. Brett, Todd E. Sheets, Adam J Dupuy, David K. Chang, Andrew V. Biankin, Nic Waddell, Karin S. Kassahn, **Sean M Grimmond**, Alistair G. Rust, David J. Adams, Nancy A. Jenkins, and Neal G. Copeland (2012) Sleeping Beauty Mutagenesis Reveals Cooperating Mutations and Pathways in Pancreatic Adenocarcinoma ***Proc Natl Acad Science USA*** 2012 Mar 15. 109(16):5934-41.

David K. Chang

CURRENT POSITIONS
Clinical Senior Lecturer
Wolfson Wohl Cancer Research Centre,
Institute of Cancer Sciences, University of Glasgow.

QUALIFICATION

1999	**M.B., B.S**	University of Sydney
2007	**F.R.A.C.S.**	Royal Australasian College of Surgeons
2008	**M.S.**	University of Sydney
2012	**Ph.D.**	University of New South Wales

HONOURS and AWARDS
Cancer Institute NSW Clinical Fellowship 2007 (AU $140,000)
ACORD Fellowship 2008 (Australia and Asia Pacific Clinical Oncology Research Development Workshop)
Australasian Pancreatic Club Travel Award 2008
American Pancreatic Association Young Investigator Award 2009 & 2010
ASCO Cancer Foundation Merit Award 2011
Cancer Institute NSW Premier's Outstanding Cancer Research Scholar 2011
Pfizer Oncology International Studentship Award 2011
Cancer Institute NSW The Wildfire Award 2012
Garvan Institute Thesis Prize of the Year 2012
WIN Symposium Best Abstract Award 2013

RESEARCH INTERESTS
Active translational and basic science for thediscovery and development of biomarker guided clinical decision-making for cancer.

FUNDING
Cancer Institute NSW Research Scholar Award 2009 – 2010 (A $50,000)
RACS Sir Roy McCaughey Surgical Research Fellowship 2009 – 2010 (AU $39,240)
NHMRC Postgraduate Research Scholarship 2009 – 2011 (AU $99,713)
NHMRC Early Career Fellowship (Peter Doherty) 2012 – 2015 (AU $354,892; CIA)
Cancer Institute NSW Early Career Fellowship 2012-2014 (AU $365,491; CIA)
NHMRC Project Grant (Personalising care in operable pancreas cancer. GAP-T: a study of imaging and molecular biomarkers to guide treatment of patients receiving preoperative chemotherapy followed by surgery) 2012 – 2014 (AU $391,175; CIE)

PUBLICATIONS (from a total of 37)

1. Biankin AV, … **Chang DK,** … Sutherland RL. Expression of S100A2 calcium-binding protein predicts response to pancreatectomy for pancreatic cancer. ***Gastroenterology***. 2009: 558
2. **Chang DK,** Johns AL, …. Kench JG, Biankin AV. Margin Clearance and Outcome in Resected Pancreatic Cancer. ***J Clin Oncol***. 2009: 2855-62
3. Humphris JL, **Chang DK,** … Sutherland RL, Biankin AV. The Prognostic and Predictive Value to Serum CA19.9 in Pancreatic Cancer. ***Ann Oncol***. 2012: 1713-22.
4. Mann KM, Ward JM, … **Chang DK**, … Jenkins NA, Copeland NG. *Sleeping Beauty* Mutagenesis Reveals Cooperating Mutations and Pathways in Pancreatic Adenocarcinoma. ***Proc Natl Acad Sci USA***. 2012: 5934-41
5. Pérez-Mancera PA, Rust AG, … **Chang DK**, …Adams DJ, Tuveson DA. The deubiquitinase *USP9X* suppresses pancreatic ductal adenocarcinoma. ***Nature***: 266-70
6. Biankin AV, Waddell N, … **Chang DK,** … McPherson JD, Gibbs RA, Grimmond SM. Pancreatic Cancer Genomes Reveal Aberrations in Axon Guidance Pathway Genes. ***Nature***. 2012: 399-405
7. **Chang DK,** Jamieson NB, … McKay CJ, Biankin AV. Histomolecular Phenotypes and Outcome in Adenocarcinoma of the Ampulla of Vater. ***J Clin Oncol***. 2013:1348-56
8. Chou A, Waddell N, …**Chang DK,** … Grimmond SM, Biankin AV. Clinical and Molecular Characterization of *HER2* amplified-pancreatic cancer. ***Genome Med***. 2013: 78

Peter Bailey

CURRENT POSITIONS
Associate Lecturer (School of Chemistry and Molecular Biosciences, UQ)
Senior Research Officer (Queensland Centre for Medical Genomics, Institute of Molecular Bioscience)

QUALIFICATION
1995 B.Sc. (Hons) The University of Queensland
1999 Ph.D. (Biochemistry) The University of Queensland
2010 MIP (Masters of Intellectual Property Law) The University of Melbourne

HONOURS and AWARDS
2012 – 2013 Research & Teaching Appointment (UQ ResTeach), The University of Queensland
2000 – 2004 NHMRC, C.J. Martin Postgraduate Fellowship

RESEARCH INTERESTS

Dr Bailey's primary research focus is directed towards the identification of molecular pathways associated with the pathogenesis of pancreatic cancer by the integration of genome, epigenome and transcriptome data. Of particular interest, is the integration of mutation signatures with "-omics" data to identify actionable molecular phenotypes as targets for adjuvant therapy.

PREVIOUS APPOINTMENTS
2010 – 2012 Research Officer, School of Chemistry and Molecular Biosciences (UQ)
2006 – 2010 Cullen & Co., Patent and Trade Mark Attorneys
2002 – 2006 Research Fellow, Centre for Molecular Biology, The Karolinska Institute
1999 – 2002 Research Fellow, Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School

SELECTED PUBLICATIONS

Juliet D. French, Maya Ghoussaini, Kerstin B. Meyer, Stacey Edwards, Kyriaki Michailidou, Shahana Ahmed, Sofia Khan, Mel J. Maranian, Martin O'Reilly, Kristine M Hillman, Joshua A Betts, Thomas Carroll, **Peter J. Bailey**, …, Melissa A. Brown, Georgia Chenevix-Trench, Douglas F. Easton, Alison M. Dunning (2013) Fine scale mapping and functional analysis of the breast cancer 11q13 (CCND1) locus. ***Am J Hum Genet.: 92(4):489-503***.

Dowhan DH*, **Bailey PJ***, Harrison MJ, Eriksson NA, Pearen MA, Fuller PJ, Funder JW, Simpson ER, Leedman PJ, Tilley WD, Brown MA, Clarke CL, Muscat GE (2012) Protein arginine methyltransferase 6-dependent gene expression and splicing: association with breast cancer outcomes**. *Endocr Relat Cancer.: 19(4):509-26.*** **The authors contributed equally***.**

Wee EJ, Peters K, Nair SS, Hulf T, Stein S, Wagner S, **Bailey PJ**, Lee SY, Qu WJ, Brewster B, French JD, Dobrovic A, Francis GD, Clark SJ, Brown MA (2012) Mapping the regulatory sequences controlling 93 breast cancer-associated miRNA genes leads to the identification of two functional promoters of the Hsa-mir-200b cluster, methylation of which is associated with metastasis or hormone receptor status in advanced breast cancer. ***Oncogene.: 31(38):4182-95.***

Albin Sandelin*, **Peter J. Bailey***, Boris Lenhard, Johan Ericson (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. ***BMC Genomics* 5: 99-105.** **The authors contributed equally.*

Jean-Bernard Lazaro*, **Peter J. Bailey*** and Andrew Lassar (2002) Cyclin D/cdk4 activity modulates the subnuclear localisation and interaction of MEF2 with SRC-family co-activators during skeletal muscle differentiation. ***Genes and Development* 16(14):1792-805.** **The authors contributed equally.*

Oliver Hofmann

CURRENT POSITIONS

Senior Research Scientist, Harvard School of Public Health
Affiliated Faculty, Harvard Stem Cell Institute
Associate Director, Harvard School of Public Health Bioinformatics Core

QUALIFICATIONS

1999	Diploma Biology	University of Cologne
2004	PhD Biochemistry	University of Cologne

RESEARCH INTERESTS

The data management, analysis and integration of biomedical high-throughput data requires a thorough understanding of the involved technologies and knowledge of the right analytical processes. As part of the HSPH Bioinformatics Core team Dr Hofmann has overseen the development and deployment of multiple analytical systems, including workflows for array quality control, expression array analysis and methylation analysis, as well as sequencing pipelines for RNA-Seq, RRBS and re-sequencing projects. His training in large-scale data management, network-base data integration standards development for biological sample description (as part of the OBO Foundry group) have provided him with the required skills to further support this project, and his research focus on biological data integration will enable him to guide the development of methods and applications for this grant proposal.

PREVIOUS APPOINTMENTS

2004 – 2007	Research Fellow, South African National Bioinformatics Institute
2007 – 2008	Research Fellow, Harvard School of Public Health
2008 – 2009	Research Associate, Harvard School of Public Health
2009 – 2012	Research Scientist, Harvard School of Public Health

RECENT PUBLICATIONS

1. P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, **O. Hofmann**, S. Neumann, P. Sterk, W. Tong, S.A. Sansone, ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, **Bioinformatics**, (2010). PMID 20679334
2. **The Fantom4 consortium**, An atlas of combinatorial transcriptional regulation in mouse and man, **Cell**, (2010). PMID 20211142
3. C. Huttenhower and **O. Hofmann**, A quick guide to large-scale genomic data mining, **PLoS Comput Biol,** (2010). PMID 20523745
4. **DIAGRAM consortium**, Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis, **Nature Genetics**, (2010). PMID 20581827
5. H.Zhao, W. Lin, **O.Hofmann**, A.W.Tai, K.Goto, L.Zhang, K.Kumthip, L.F. Peng, F. Dahlene, N. Jilg, W. Hide, R.X. Shao, J.Y. Jang, M-N. Jorge and R.T. Chung}, A functional genomic screen reveals novel host genes that mediate interferon-alpha's antiviral effects against hepatitis C virus, **J Hepatol** (2011). PMID 21888876
6. S. Fu, L. Yang, P. Li, **O. Hofmann**, L. Dicker, W. Hide, X. Lin, S.M. Watkins, A.R. Ivanov and G.S. Hotamisligil, Aberrant lipid metabolism disrupts calcium homeostasis causing liver endoplasmic reticulum stress in obesity. **Nature** (2011). PMID 21532591

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Alternative splicing in cancer transcriptomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)

Steven G. Rozen, Duke-NUS Graduate Medical School Singapore; ICGC Singapore, Biliary Duct Cancer

Patrick Tan, Duke-NUS-Graduate Medical School, Genome Institute of Singapore, National Cancer Centre Singapore; ICGC Singapore; Singapore, Biliary Duct Cancer

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Ioana Cutcutache, Duke-NUS Graduate Medical School Yuka Suzuki, Duke-NUS Graduate Medical School	

Background and preliminary data

There is considerable evidence that changes in alternative splicing and related changes in the transcriptome such as alternative 5' exon usage and alternative polyadenylation contribute to carcinogenesis. *CD44* is a well-studied example of a gene with multiple alternative splice events (ASEs) that have been implicated in cancer. Studies have found that the *CD44*v8-v10 spliceform promotes tumorigenesis by conferring resistance to oxidative stress and that another spliceform, known as *CD44*v6-v14, promotes cell growth in cancer cell lines. The *CD44*v5 spliceform was found to be expressed preferentially in poorly differentiated gastric tumors and in metastatic lymph nodes, correlating with shorter patient survival times. Other genes that are functionally altered through ASEs include the *FGFR2* gene, which switches isoforms during cancer progression, and *MST1R* (macrophage stimulating 1 receptor [c-met-related tyrosine kinase], also known as *RON*). The gene *ENAH* [(Enabled Homolog (Drosophila), also known as *MENA*] has a spliceform that has been extensively studied in the context of epithelial- mesenchymal transition in cancer.

We recently completed a study of alternative splicing in gastric cancers (Liu et al., submitted, 2014), and found spliceforms of several genes, including *KITLG* (*KIT* ligand), that have spliceforms that are highly enriched in gastric cancer compared to non-malignant gastric epithelium. For this study we used an analytical pipeline based to TopHat (Trapnell et al., *Bioinformatics*, 2009) and MISO (Katz et al., *Nature Methods*, 2010).

Nevertheless, the complexity of transcriptome has made study of the breadth of alternative splicing across many genes in many cancers difficult. **We propose, as a high-risk non-essential analysis, to examine the available ICGC RNA-seq data for alternative splicing, alternative 5'-exon usage and alternative polyadenylation across multiple cancer types.** We propose to examine whether there are significant differences in relative spliceform abundance across different cancer types, and whether clustering by relative spliceform abundance reveals similarities or differences within and across cancer types. (By "spliceform" we include variation in 5'-exon usage in alterative polyadenylation.) We also propose to examine whether particular spliceforms are enriched in multiple cancers relative to non-malignant tissues. This proposal is high-risk because differences in RNA preparation and differences in the details of sequencing protocols (number of reads, read length, paired or single-end sequencing, fragment size) across centers may confound the analysis. We will address this high risk by doing an early analysis with available data and deciding at that point whether this analysis should proceed further.

Timelines & resources dedicated to project

Timeline: Preliminary analyses and methodology development will begin immediately with available data.

We will decide whether a full analysis is possible in June, 2014. If we decide to proceed, data will be re-analyzed as additional data become available. In this case, we anticipate a data "freeze" in September 2014 to allow analysis for publication.

Resources: Rozen will spend 20% of his time on the study of alternative splicing in cancer. Cutcutache and Suzuki will carry out most of the in silico analyses under Rozen's supervision. We have the expertise to make use of ICGC-provided cloud computing. In addition, we have access to a Linux cluster and additional compute resources should these be necessary for running the TopHat / MISO pipeline or other analysis.

Research proposal

We will rely on the pipeline based on TopHat and MISO that we used in our previous study of alternative splicing in gastric cancer. MISO considers possible alternative splicing events, "ASEs"; an ASE is a set of alternative splicing configurations. The "Percent Spliced Isoform" (PSI, sometimes also referred to as "Percent Spliced In") is the proprtion of the transcripts with one of the possible configuration in an ASE. For the situation in which the ASE has only two configurations, one skipping an exon, PSI can be literally the percent of transcripts in which the exon is spliced in. Thus, PSIs can be thought of as the relative abundance of particular spliceforms in the ASE. The accuracy with which PSIs can be estimated depends on read depth, read length, and whether reads are paired or unpaired. As a result, comparisons between RNA-seq data set that vary in these characteristics is likely to be problematic. Therefore we will decide in June 2014 whether the RNA-seq data associated with the ICGC genomes is likely to be suitable for a global study of alternative splicing in cancer, for example a study that would attempt clustering of tumors based on relative abundance of spliceforms or that would look for associations between particular isoforms. If we conclude that the data (or, more likely, a subset of the data) are suitable for such global analyses we will refine our methodology and then carry out an analysis for publication based on a data freeze in September, 2014. We may also decide that the data are unsuitable for a global analysis, but that they may be analyzed to look for spliceforms that have higher relative abundance in tumors than in non-malignant tissues). In this case we will proceed with this analysis on data freeze in Sepember.

Legacy plans

The catalog of alternative splicing events and their PSIs across the analyzed tumors will be released as a database. The pipelines for used for discovering alternative spliceforms will be made available as "literate programming" (e.g. knitr) scripts. Statistical analyses of alternative splicing across tumors and across tumor types will be made available as knitr documents.

Name Steven G. ROZEN
Mailing address Duke-NUS Graduate Medical School, 8 College Road, 169857 Singapore
Email steve.rozen@duke-nus.edu.sg

Current positions

Associate Professor with tenure, Duke-NUS Graduate Medical School, primary faculty appointment
Director, Duke-NUS Centre for Computational Biology
Associate Professor Track V, Duke University Medical Center, U.S.A

Academic qualifications B.A. University of California at Riverside, 1972; M.S., New York University, 1986; Ph.D., New York University, 1993

Research interests Translational research based on cancer genetics, genomics, and transcriptomics; next-generation sequencing for variation and mutation discovery, bioinformatics for cancer genomics.

Selected publications (from a total of 83)

1. Y. Liu, ..., P. Tan, **S. G. Rozen**. "Alternative splicing in gastric cancer". *submitted* (2013)
2. S. L. Poon, J. R. McPherson, P. Tan, B. T. Teh, **S. G. Rozen**. "Genome wide carcinogen signatures promise new opportunities for cancer prevention." (review) *submitted* (2013)
3. I. Cutcutache, ..., P. Tan, **S. G. Rozen,** "Abundant hemizygous deletions of CYCLOPS and STOP genes in gastric adenocarcinoma. " *submitted* (2013)
4. Z. Lei,..., P. Tan, **S. G. Rozen**. "Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase Inhibitors and 5-fluorouracil." Gastroenterology 145:554-565 (2013) (received journal editorial) IF 12.8
5. W. Chan-on, ..., **S. G. Rozen**,* P. Tan P*, B. T. Teh*. "Distinct mutational patterns of infection and non-infection-related bile duct cancers revealed by exome sequencing." Nat. Genet. 45:1474:1478 IF 35.2
6. S. L. Poon, ..., **S. G. Rozen***, P. Tan*, B. T. Teh* "Genome wide mutational signatures of aristolochic acid and its application as a screening tool." Sci. Transl. Med. 5:197ra101 (2013) (Received journal cover image and focus article; also highlighted by Science Magazine and The Scientist; * = corresponding author) IF 10.8
7. C. K. Ong, …, **S. Rozen***, P. Tan*, B. T. Teh*. "Exome sequencing of liver fluke-associated cholangiocarcinoma." Nat. Genet. 44:690-3 (2012) (* = corresponding author) IF 35.2
8. Z. J. Zang, …, **S. Rozen***, B. T. Teh*, P. Tan*. "Exome sequencing of gastric adenocarcinoma reveals recurrent somatic mutations in cell adhesion and chromatin remodeling genes." Nat. Genet. 44:570-74 (2012) (* = corresponding author) IF 35.2
9. A. Untergasser, …, **S. G. Rozen**. "Primer3-new capabilities and interfaces." Nucleic Acids Res. 40:e115 (2012) IF 8.3
10. **S. G. Rozen**, …, D. C. Page. "*AZFc* Deletions and Spermatogenic Failure: A Population-Based Survey of 20,000 Y Chromosomes." Am. J. Hum. Genet. 91(15):890-896 (2012) IF 11.2
11. J. F. Hughes, **S. Rozen**. "Genomics and genetics of human and primate Y chromosomes" (review), Annu. Rev. Genomics Hum. Genet. 13:83-108 (2012) IF 9.5
12. N.Deng, …, **S.G.Rozen**, P.Tan. "A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets." Gut. 61:673-684 (2012) IF 10.7
13. J. Ye, …, **S. Rozen**, and T. Madden. "Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction". BMC Bioinformatics 13:134 (2012)
14. Z. J. Zang, …, **S. Rozen***, B. T. Teh*, and P. Tan*. "Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing." Cancer Res. 71:29-39 (2011) (* = corresponding author) IF 7.9
15. **S. Rozen**, …, D. C. Page. "Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection." Am. J. Hum. Genet. 85:923-928 (2009) IF 10.2

Selected patents

Grouping for Classifying Gastric Cancer and Methods of Using the Same; Singapore filing 201206943-1 (2012)
Markers of Alterations in the Y Chromosome and Uses Therefor; EP Patent 1,794,321 (2007); US Patent App. 11/195,344 (2005); US Patent App. 12/251,270 (2008)

Selected scientific awards

2003 Breakthrough of the Year #9 by the News and Editorial Staffs of Science, for sequencing the human Y chromosome, reported in **Rozen** et al., *Nature*, (2003), and in H. Skaletsky et al.,…, **S. Rozen**, D. C. Page, *Nature*, (2003) IF 31.0
2003 Faculty of 1000, Exceptional Paper: S. Repping, …, **S. Rozen**. *Nat. Genetics*, (2003) IF 26.5
2006 Faculty of 1000 Medicine, Must Read for Repping, ..., **S. Rozen.**, *Nat. Genetics,* (2006) IF 24.2

Current support as PI

1. "Developing a clinically testable biomarker-based predictor for early stage colorectal cancer likely to metastasize" 1/4/13 to 31/3/16, BMRC 13/1/96/191684; SG$ 506,143 direct (approx US$ 400,000)
2. "Identification of drugs for targeted treatment of PTEN-deficient tumors", NMRC/GMS/CIRG/1324/2012; 6/19/12-6/18/15; SG$ 1,389,000 direct (approx. US$ 1,132,768)
3. "A striatal synaptic dysfunction hypothesis for repetitive behaviors in autism evaluated by re-sequencing of candidate genes", NMRC/GMS/1248/2010; 9/1/10-7/31/14; SG$ 1,500,000 direct (approx. US$ 1,223,292)

Name Patrick Tan MD, PhD
Mailing address Duke-NUS Graduate Medical School Singapore, 8 College Road, Singapore 169857
Email address gmstanp@duke-nus.edu.sg

Current positions

2012 - Present Professor, Cancer and Stem Cell Biology, Duke-NUS Graduate Medical School Singapore
2013 - Present Senior Group Leader, Genome Institute of Singapore
2012 - Present Senior Principal Investigator, Cancer Science Institute of Singapore (NUS)
2009 - Present Research Associate Professor, Institute of Genome Sciences and Policy, Duke Uni.
2009 - Present Director, Duke-NUS Genome Biology Facility
2006 - Present Principal Investigator (Adjunct), National Cancer Centre, Singapore
2012 - Present Professor (Adjunct), Dept of Physiology, National University of Singapore

Academic qualifications 1992, Harvard University, B.A.(Biochemistry); 2000, Stanford University School of Medicine, M.D.,Ph.D (Developmental Biology)

Research interests Gastric cancer & cancer genomics

Publications (from a total of 104)

1) Chan-on W, ..., Rozen* SG, **Tan* P***,Teh* BT. (2013) Distinct mutational patterns of infection and non-infection-related bile duct cancers revealed by exome sequencing. Nat. Genet. 45:1474:1478 * = corresponding author
2) Poon SL, ..., **Tan* P**, Teh* BT. (2013) Genome wide mutational signatures of aristolochic acid and its application as a screening tool. Sci. Transl. Med. 5:197ra101 (Received journal cover image and focus article; also highlighted by Science Magazine and The Scientist) * = corresponding author IF 10.8
3) Z. Lei,,....., **Tan* P**, Rozen* SG. (2013) Identification of molecular subtypes of gastric cancer with different responses to PI3-Kinase inhibitors and 5-fluorouracil. Gastroenterology 145(3):554-65 (2013) (Featured in Editorial) * = corresponding author
4) Zouridis H, Deng N, Ivanova T, ..., Teh BT, Rozen S, **Tan P**. (2012) Methylation subtypes and large scale epigenetic alterations in gastric cancer. Sci Transl Med. 4(156):156ra140
5) Zang ZJ, ..., Rozen* SG, BT Teh* BT, **Tan* P**. (2012) Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet. 44(5):570-4 * = corresponding autho
6) N Deng, LK, ..., Rozen S, **Tan P**. (2012) A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets.Gut. 61(5):673-84.
7) Tan IB, ..., **Tan P.** (2011) Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. Gastroenterology 141(2):476-85

Patents

2006 Singapore Patent 200404696-7 " Materials and Methods for Cancer Diagnosis"
2008 European Patent EP1668151B1 "Materials and Methods Related to Breast Cancer Diagnosis"
2008 Singapore Patent 160726 "Materials and Methods Related to Breast Cancer Diagnosis"

Scientific awards

2001 Young Scientist Award (Singapore National Academy of Sciences)
2004 SingHealth Investigator Excellence Award
2011 Singapore General Hospital Scientist Award
2013 Chen New Investigator Award, Human Genome Organization
2013 American Society for Clinical Investigation

Current support

Jul 2011 - Jun 2014 Genomic Interrogation of FGFR2-amplified Gastric Cancer Leading to a Phase 2 Adaptive Clinical Trial with Molecular Targeted Therapies and Predictive Biomarker Development, BMRC / NMRC, A*STAR, SG$606,000
Aug 2011 - Jul 2014 Lineage-Specific Survival Oncogenes in Gastric Cancer: Functional Characterization, Genomic Dissection and Integration with Classical Oncogenic Circuits , Biomedical Research Council, A*STAR , SG$1,131,500
Oct 2012 - Nov 2015 Developing clinically implementable multiplex gene assays for specific indications across the continuum of care in stomach and kidney cancer, BMRC-NMRC, A*STAR, SG$1,421,600
Feb 2013 - Jan 2018 TCR Flagship Program (Tier 2): Singapore Gastric Cancer Consortium – Re-defining Gastric Cancer Management, NMRC, Subaward from NUHS as theme PI, SG$4,780,000
Mar 2012 - Apr 2014 CMOS nanoplasmonic array for the rapid determination of pathogen growth and antibiotic susceptibility, A*STAR Joint Council Office, SG$514,000
Dec 2012 – Dec 2015 The POLARIS Programme- Transforming Disease Management Through Personalised OMIC Profiling, BMRC SPF, SG$19,999,998.50

Name Ioana CUTCUTACHE
Mailing address Duke-NUS Graduate Medical School, 8 College Road, 169857 Singapore
Email ioana.cutcutache@duke-nus.edu.sg

Current position

Research Associate, Centre for Computational Biology, Duke-NUS Graduate Medical School

Academic qualifications

B.Sc. Politehnica University of Bucharest, Faculty of Automatic Control and Computers, 2005;
M.Sc., National University of Singapore, School of Computing, 2009.

Research interests

Cancer genetics and genomics; next-generation sequencing for variation and mutation discovery, bioinformatics for genomics, transcriptomics, and cancer genetics.

Publications

1. **I. Cutcutache** et al**,** "Abundant hemizygous deletions of CYCLOPS and STOP genes in gastric adenocarcinoma. " *submitted* (2013)
2. Z. J. Zang*, **I. Cutcutache***, … S. Rozen, B. T. Teh, P. Tan. "Exome sequencing of gastric adenocarcinoma reveals recurrent somatic mutations in cell adhesion and chromatin remodeling genes." Nat. Genet. 44:570-74 (2012) (* = equal contribution) IF 35.2
3. W. Chan-on, M.-L. Nairismägi, C. K. Ong, S. Dima, C. Pairojkul, K. H. Lim, J. R. McPherson, W. K. Lim, **I. Cutcutache,** … S. G. Rozen, P. Tan P, B. T. Teh. "Distinct mutational patterns of infection and non-infection-related bile duct cancers revealed by exome sequencing." Nat. Genet. 45:1474:1478 (2013) IF 35.2
4. S. L. Poon, S.-T. Pang, J. R. McPherson, W. Yu, K. K. Huang, P. Guan, W.-H. Weng, E,.Y. Siew, Y. Liu, H. L. Heng, S. C. Chong, A. Gan, S. T. Tay, W. K. Lim, **I. Cutcutache**, … S. G. Rozen, P. Tan, B. T. Teh. "Genome wide mutational signatures of aristolochic acid and its application as a screening tool." Sci. Transl. Med. 5:197ra101 (2013) (Received journal cover image and focus article; also highlighted by Science Magazine and The Scientist) IF 10.8
5. C. K. Ong, C. Subimerb, C. Pairojkul, S. Wongkham, **I. Cutcutache,** … S. Rozen, P. Tan, B. T. Teh. "Exome sequencing of liver fluke-associated cholangiocarcinoma." Nat. Genet. 44:690-3 (2012) IF 35.2
6. A. Untergasser*, **I. Cutcutache***, … S. G. Rozen. "Primer3-new capabilities and interfaces." Nucleic Acids Res. 40:e115 (2012) (* = equal contribution) IF 8.3
7. J. Ye, G. Coulouris, I. Zaretskaya, **I. Cutcutache**, …, S. Rozen, and T. Madden. "Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction". BMC Bioinformatics 13:134 (2012)
8. N. Nagarajan, D. Bertrand, A. M. Hillmer, Z. J. Zang, F. Yao, P. E. Jacques, A. S. Teo, **I. Cutcutache**, … P. B. Tan, Y. Ruan. “Whole-genome reconstruction and mutational signatures in gastric cancer.” Genome Biology 13:R115 (2012)
9. G. C. Koo, S. Y. Tan, T. Tang, S. L. Poon, G. E. Allen, L. Tan, S. C. Chong, W. S. Ong, K. Tay, M. Tao, R. Quek, S. L., Kheng-Wei Yeoh, S. P. Yap, K. A. Lee, L. C. Lim, D. Tan, C. Goh, **I. Cutcutache**, … S. Rozen, P. Tan, B. T. Teh, S. T. Lim. “Janus Kinase 3-Activating Mutations Identified in Natural Killer/T-cell Lymphoma.” Cancer Discovery 2(7):591-7 (2012)
10. W. Yu, W. Chan-On, M. Teo, C. K. Ong, **I. Cutcutache**, … S. Rozen, K. C. Soo, P. Tan, B. T. The. “First somatic mutation of E2F1 in a critical DNA binding residue discovered in well differentiated papillary mesothelioma of the peritoneum.” Genome Biology 12:R96 (2011)
11. L. Goh, G. B. Chen, **I. Cutcutache**, B. Low, B. T. Teh, S. Rozen, P. Tan. “Assessing Matched Normal and Tumor Pairs in Next-Generation Sequencing Studies.” PLoS ONE 6(3): e17810 (2011)
12. I. B. Tan, **I. Cutcutache**, … S. Rozen, E. H. Tan, P. Tan. “Fanconi's anemia in adulthood: Chemoradiation-induced bone marrow failure and a novel FANCA mutation identified by targeted deep sequencing.” Journal of Clinical Oncology 29(20):e591-4 (2011)
13. Z. J. Zang, C.K. Ong, **I. Cutcutache**, … S. Rozen, B. T. Teh, P. Tan. "Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing." Cancer Res. 71:29-39 (2011)
14. Q. Zhao, **I. Cutcutache**, W. F. Wong. “PiPA: Pipelined Profiling and Analysis on Multi-core Systems.” ACM Transactions on Architecture and Code Optimization, Volume 7, No. 3, Article 13 (2010)
15. **I. Cutcutache**, W. F. Wong. “Fast, frequency-based, integrated register allocation and instruction scheduling.” Software: Practice and Experience, Volume 38, Issue 11, Pages 1105-1126 (2008)
16. **I. Cutcutache**, … F. E. H. Tay, W. F. Wong. “BSN Simulator: Optimizing Application Using System Level Simulation.” 6th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2009)
17. K. D. Nguyen, **I. Cutcutache**, … T. Mitra, W. F. Wong. “Fast and Accurate Simulation of Biomonitoring Applications on a Wireless Body Area Network”, 5th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2008)
18. K. D. Nguyen, **I. Cutcutache**, … F. T. E. Hock, T. Mitra. “A SystemC-based Fast Simulator for Biomonitoring Applications on Wireless BAN.” Workshop on Software and Systems for Medical Devices and Services (SMDS 2007)

Name Yuka Suzuki

Mailing address Duke-NUS Graduate Medical School, 8 College Road, 169857 Singapore

Email yuka.suzuki@nus.edu.sg

Current position Ph.D Candidate at Duke-NUS Graduate Medical School, Cancer and Stem Cell Biology Program

Academic qualifications B.Sc. (Hons), National University of Singapore, 2010; M.Sc., University College London, 2011

Research interests Genomic characterization of tumors using targeted next-generation sequencing; identification of cancer-specific transcript variants using high-throughput RNA sequencing

Publications

1. I. Cutcutache, A. Y. Wu**, Y. Suzuki**, …, S. G. Rozen. "Abundant hemizygous deletions of CYCLOPS and STOP genes in gastric adenocarcinoma." *submitted* (2013)

2. I. B. Tan, K. Ramnarayanan, S. Malik, J. McPherson, **Y.Suzuki**, … P. Tan. "High-depth targeted next-generation sequencing of over 750 cancer-associated genes reveals high degree of similarity in paired colorectal primaries and liver metastases." *submitted* (2013)

G.A. Wilson, P. Dhami, A. Feber, D. Cortázar, **Y. Suzuki**, … S. Beck. "Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers." *GigaScience*, *1*:3 (2012)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Examination of signatures of physical mutational processes to infer genotoxic exposures

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Prof. Steven G. Rozen, Duke-NUS Graduate Medical School Singapore; ICGC Singapore, Biliary Duct Cancer
Prof. Bin Tean Teh, National Cancer Centre Singapore; Duke-NUS Graduate Medical School Singapore; ICGC Singapore; Singapore, Biliary Duct Cancer

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

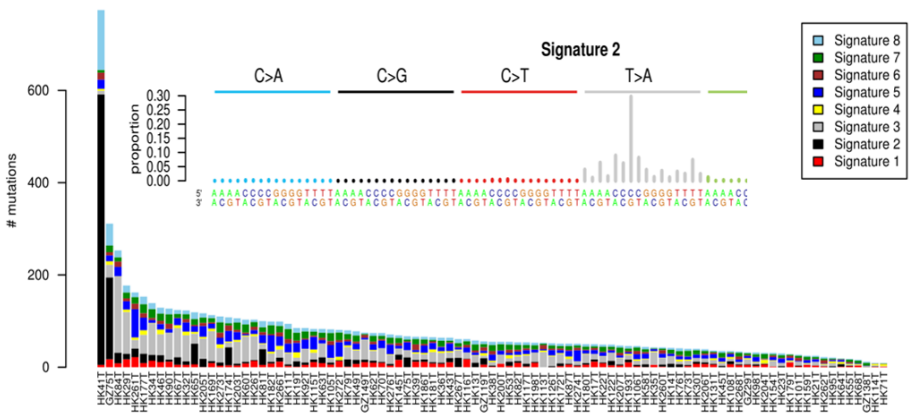
Dr. Song Ling Poon, National Cancer Centre Singapore
Dr. John. R. McPherson, Duke-NUS Graduate Medical School Singapore

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

The characteristic signatures of physical mutational processes observed in tumors provide valuable new information on cancer epidemiology and etiology because they can indicate specific environmental exposures or endogenous mutational processes that contributed to carcinogenesis. Alexandrov and colleagues (*Nature* and *Cell Reports*, 2013) developed an effective method for extracting signatures from the catalogs of mutations across thousands of tumors. This approach is based on non-negative matrix factorization (NMF). Using NMF they detected 11 signatures associated with known environmental exposures (e.g. UV light or tobacco smoke) or endogenous mutational processes (e.g. activated APOBECs or defective proofreading in some DNA polymerases). Nevertheless, the exposures or processes responsible for 10 signatures were unknown. Conversely, and equally important, detailed sequence-level signatures of many known mutagenic carcinogens are unknown.

Furthermore, given that environmental exposures likely vary by geography, we hypothesize that many signatures remain undetected because no or very few tumors with those signatures have been resequenced. The signature of exposure to aristolochic acid (AA) is an example of one not detected by Alexandrov and colleagues, presumably because tumors with this signature were rare or absent among the tumors studied. We and another group recently delineated the sequence-level mutation signature of AA exposure in upper urinary tract urothelial tumors and in AA-exposed cell lines (Poon et al. and Hoang et al. *Science Translational Medicine,* 2013). We also detected, by inspection, the presumed signature of AA in hepatocellular carcinomas (HCCs), which had not previously been associated with AA exposure. Thus, AA exposure through herbal remedies likely contributes to HCC as well as urothelial tumors. We subsequently applied the NMF procedure of Alexandrov and colleagues to a series of tumors, including HCCs, and detected the signature of AA exposure in additional HCCs, as shown in this figure, where Signature 2 is the AA-associated signature.



Under the hypothesis that many signatures have yet to be sampled, the ~2,000 ICGC genomes and ~8,000 exomes from diverse geographical locations should harbor additional, novel signatures.

At present, we are experimentally determining the physical mutation signatures of

several mutagenic carcinogens in cell culture and in mice.

We propose to answer the following questions: (1) What signatures are present in the ~2,000 ICGC genomes? **(2)** Do differences in variant calling procedures affect the set of detected signatures? **(3)** Can we reliably detect additional signatures in the ~2,000 whole genomes combined with the ~8,000 ICGC exomes and additional, publicly available, data? **(4)** With regard to the signatures detected in (1), (2), and (3), which ones correlate with clinical characteristics or can be attributed to known environmental exposures or endogenous mutational processes based on other evidence? **(5)** What are the changes in signatures before and after treatment and between primary tumors and distant metastases? **(6)** What computational methods can best be used for simultaneously recognizing known mutation signatures and discovering unknown signatures?

Timelines & resources dedicated to project

Timeline: Preliminary analyses and methodology development will begin immediately. Data will be re-analyzed as additional data become available. We anticipate a data freeze in September 2014 to allow analysis for publication.

Resources: Rozen will spend 20% of his time on study of physical mutation signatures. Poon with support from Teh will focus on in vitro and in vivo studies of detailed sequence-level signatures of known mutagenic carcinogens. McPherson (co-first author on our AA signature paper) will carry out most of the in silico analysis. We have the expertise to make use of ICGC-provided cloud computing. In addition, we have access to a Linux cluster and to additional compute resources for the computationally intensive tasks of signature discovery using NMF.

Research proposal

Here we describe our approach to each of the 6 questions posed above:

(1) What signatures are present in the ~2000 ICGC genomes? For this we will initially use the NMF approach developed by Alexandrov and colleagues. We will compare the signatures detected in this data set with those previously reported. We will also use these data to develop alternative analyses for question (6).

(2) Do differences in variant calling procedures affect the set of detected signatures? We will investigate this question by (a) analyzing the 2000 genomes separately for each variant caller and comparing the resulting signatures and (b) analyzing the 2000 genomes times 3 variant callers in one analysis to determine if some signatures are "split" by the variant caller used.

(3) Can we reliably detect additional signatures in the ~2000 whole genomes combined with the ~8000 ICGC exomes and additional, publicly available, data? We will first determine if similar signatures are extracted from exome and from genome data when the genome data are "projected" onto the exome. We will then proceed to analyze combined genome and exome data, and assess the stability of the signatures.

(4) With regard to the signatures detected in (1), (2), and (3), which ones correlate with clinical characteristics or can be attributed to known environmental exposures or endogenous mutational processes based on other evidence? We will assess statistical associations between strengths of signatures and clinical and epidemiological covariates, and will also compare signatures to signatures that we ascertained by in vitro or in vivo studies.

(5) What are the changes in signatures before and after treatment and between primary tumors and distant metastases? Using procedures for detecting signatures, we can assess whether mutations that occur after treatment or only metastases reflect different mutational processes than occurred during initial development of the tumor.

(6) What computational methods can best be used for simultaneously recognizing known mutation signatures and discovery of unknown signatures? This question is exploratory, but as one approach, one might "spike in" pure signatures into the NMF procedure's input and assess whether this affects the results and whether NMF returns a close approximation of the spiked-in signatures.

Legacy plans

The mutation signatures detected will be published, including the associated exposures when known. The mixtures of signatures identified in the analyzed genomes and exomes will be made available in a database. Algorithms for combined discovery of unknown mutation signatures and detection of signatures due to known exposures or mutational processes will be published: the "feature sets" (including e.g. single-nucleotide substitutions in trinucleotide context) analyzed will be made available in a database; documented software

packages and “literate programming” (e.g. knitr) top-level scripts for extracting signatures will be released.

Name Steven G. ROZEN
Mailing address Duke-NUS Graduate Medical School, 8 College Road, 169857 Singapore
Email steve.rozen@duke-nus.edu.sg

Current positions

Associate Professor with tenure, Duke-NUS Graduate Medical School, primary faculty appointment
Director, Duke-NUS Centre for Computational Biology
Associate Professor Track V, Duke University Medical Center, U.S.A

Academic qualifications B.A. University of California at Riverside, 1972; M.S., New York University, 1986; Ph.D., New York University, 1993

Research interests Translational research based on cancer genetics, genomics, and transcriptomics; next-generation sequencing for variation and mutation discovery, bioinformatics for cancer genomics.

Selected publications (from a total of 83)

1. Y. Liu, ..., P. Tan, **S. G. Rozen**. "Alternative splicing in gastric cancer". *submitted* (2013)
2. S. L. Poon, J. R. McPherson, P. Tan, B. T. Teh, **S. G. Rozen**. "Genome wide carcinogen signatures promise new opportunities for cancer prevention." (review) *submitted* (2013)
3. I. Cutcutache, ..., P. Tan, **S. G. Rozen,** "Abundant hemizygous deletions of CYCLOPS and STOP genes in gastric adenocarcinoma. " *submitted* (2013)
4. Z. Lei,..., P. Tan, **S. G. Rozen**. "Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase Inhibitors and 5-fluorouracil." Gastroenterology 145:554-565 (2013) (received journal editorial) IF 12.8
5. W. Chan-on, ..., **S. G. Rozen**,* P. Tan P*, B. T. Teh*. "Distinct mutational patterns of infection and non-infection-related bile duct cancers revealed by exome sequencing." Nat. Genet. 45:1474:1478 IF 35.2
6. S. L. Poon, ..., **S. G. Rozen***, P. Tan*, B. T. Teh* "Genome wide mutational signatures of aristolochic acid and its application as a screening tool." Sci. Transl. Med. 5:197ra101 (2013) (Received journal cover image and focus article; also highlighted by Science Magazine and The Scientist; * = corresponding author) IF 10.8
7. C. K. Ong, …, **S. Rozen***, P. Tan*, B. T. Teh*. "Exome sequencing of liver fluke-associated cholangiocarcinoma." Nat. Genet. 44:690-3 (2012) (* = corresponding author) IF 35.2
8. Z. J. Zang, …, **S. Rozen***, B. T. Teh*, P. Tan*. "Exome sequencing of gastric adenocarcinoma reveals recurrent somatic mutations in cell adhesion and chromatin remodeling genes." Nat. Genet. 44:570-74 (2012) (* = corresponding author) IF 35.2
9. A. Untergasser, …, **S. G. Rozen**. "Primer3-new capabilities and interfaces." Nucleic Acids Res. 40:e115 (2012) IF 8.3
10. **S. G. Rozen**, …, D. C. Page. "*AZFc* Deletions and Spermatogenic Failure: A Population-Based Survey of 20,000 Y Chromosomes." Am. J. Hum. Genet. 91(15):890-896 (2012) IF 11.2
11. J. F. Hughes, **S. Rozen**. "Genomics and genetics of human and primate Y chromosomes" (review), Annu. Rev. Genomics Hum. Genet. 13:83-108 (2012) IF 9.5
12. N.Deng, …, **S.G.Rozen**, P.Tan. "A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets." Gut. 61:673-684 (2012) IF 10.7
13. J. Ye, …, **S. Rozen**, and T. Madden. "Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction". BMC Bioinformatics 13:134 (2012)
14. Z. J. Zang, …, **S. Rozen***, B. T. Teh*, and P. Tan*. "Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing." Cancer Res. 71:29-39 (2011) (* = corresponding author) IF 7.9
15. **S. Rozen**, …, D. C. Page. "Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection." Am. J. Hum. Genet. 85:923-928 (2009) IF 10.2

Selected patents

Grouping for Classifying Gastric Cancer and Methods of Using the Same; Singapore filing 201206943-1 (2012)
Markers of Alterations in the Y Chromosome and Uses Therefor; EP Patent 1,794,321 (2007); US Patent App. 11/195,344 (2005); US Patent App. 12/251,270 (2008)

Selected scientific awards

2003 Breakthrough of the Year #9 by the News and Editorial Staffs of Science, for sequencing the human Y chromosome, reported in **Rozen** et al., *Nature*, (2003), and in H. Skaletsky et al.,…, **S. Rozen**, D. C. Page, *Nature*, (2003) IF 31.0
2003 Faculty of 1000, Exceptional Paper: S. Repping, …, **S. Rozen**. *Nat. Genetics*, (2003) IF 26.5
2006 Faculty of 1000 Medicine, Must Read for Repping, ..., **S. Rozen.**, *Nat. Genetics,* (2006) IF 24.2

Current support as PI

1. "Developing a clinically testable biomarker-based predictor for early stage colorectal cancer likely to metastasize" 1/4/13 to 31/3/16, BMRC 13/1/96/191684; SG$ 506,143 direct (approx US$ 400,000)
2. "Identification of drugs for targeted treatment of PTEN-deficient tumors", NMRC/GMS/CIRG/1324/2012; 6/19/12-6/18/15; SG$ 1,389,000 direct (approx. US$ 1,132,768)
3. "A striatal synaptic dysfunction hypothesis for repetitive behaviors in autism evaluated by re-sequencing of candidate genes", NMRC/GMS/1248/2010; 9/1/10-7/31/14; SG$ 1,500,000 direct (approx. US$ 1,223,292)

Name Bin Tean TEH
Mailing address 11 Hospital Drive, Singapore 169610
Email teh.bin.tean@singhealth.com.sg
Current positions
Director & Professor, NCCS-VARI Translational Research Laboratory, National Cancer Centre
Professor, Duke-NUS Graduate Medical School, Singapore
Senior Principal Investigator, Cancer Science Institute of Singapore
Adjunct Professor, Baylor College of Medicine, USA
Adjunct Professor, Karolinska Institute, Sweden
Adjunct Professor, Sun Yat-sen University Cancer Center, China
Academic qualifications Medical Degree, University of Queensland, Australia, 1992; PhD, Karolinska Institute, Sweden, 1997
Research interests Cancer Genomics and Biology
Selected publications (from a total of 314)

1. G.L. Dalgliesh, …, **B. T. Teh,** …, P.A. Futreal. " Systemic sequencing of renal cell carcinoma reveals inactivation of histone modifying genes." Nature 463(7279) 360-363 (2010) (* = corresponding author) IF 38.6
2. D. Huang, ..., **B.T. Teh,** "Interleukin-8 mediates resistance to anti-angiogenic sunitinib in renal cell carcinoma. " Cancer Res 70(3):1063-1071 (2010) IF 8.7
3. I.Varela, ..., **B. T. Teh*,** … " Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma." Nature 469(7331):539-542 (2011) (* = corresponding author) IF 38.6
4. A. Ooi, ..., **B. T. Teh** *, K.A. Furge "An antioxidant response phenotype shared between hereditary and sporadic type 2 papillary renal cell carcinoma." Cancer Cell 20(4):511-523 (2011) *Highlighted in Cancer Cell Previews 20(4): 418-420 (2011), Nature Reviews Cancer 11 (2011) and Nature Chemical Biology (2012) (* = co-corresponding author) IF 27.1
5. C. K. Ong, …, S. Rozen*, P. Tan, **B. T. Teh**. "Exome sequencing of liver fluke-associated cholangiocarcinoma." Nat. Genet. 44:690-3 (2012) IF 35.2
6. Z. J. Zang, …, S. Rozen*, **B. T. Teh***, P. Tan. "Exome sequencing of gastric adenocarcinoma reveals recurrent somatic mutations in cell adhesion and chromatin remodeling genes." Nat. Genet. 44:570-74 (2012) (* = co-corresponding author) IF 35.2
7. G.C. Koo, …, S. Rozen, P. Tan, **B. T. Teh***, S.T. Lim "Janus kinase 3-activating mutations identified in natural killer/T-cell lymphoma." Cancer Discov. 2(7):591-597 (2012) (* = co-corresponding author) IF 10.1
8. Ooi A ..... B.T.Teh,Furge KA. "CUL3 and NRF2 mutations confer a NRF2 activation phenotype in a sporadic form of paillary renal cell carcinoma." Cancer Res 73(7): 2044-2051,(2013)
9. S.L. Poon, …, S. G. Rozen, P. Tan, **B. T. Teh**. "Genome wide mutational signatures of aristolochic acid and its application as a screening tool." Sci. Transl. Med. 5:197ra101 (2013) (Received journal cover image and focus article; also highlighted by Science Magazine) IF 10.8
10. W. Chan-on, ..., S. G. Rozen, P. Tan, **B. T. Teh**. "Distinct mutational patterns of infection and non-infection-related bile duct cancers revealed by exome sequencing." Nat. Genet. 45:1474:1478 IF 35.2

Selected patents
Exome sequencing identifies frequent mutation of the SWI/SNF complex gene, PBRM1, in renal carcinoma. - US Provisional Patent Application no. 61/385, 426.
JAK3 Mutations Identified in Natural-Killer/T-Cell Lymphoma - Singapore Patent Application no. SG 201108800-2 (Filed 25 Nov 2011)

Selected scientific awards
2013 Honorable Member, Romanian Academy of Medical Sciences
2009 – 2014 Singapore Translational Research Investigator Award (STaR Award)
2008 University of Queensland International Alumnus of the Year Highly Commended Nominee, Australia
2007 Biotechnology Healthcare Distinguished Scientist Award, Malaysian Academy of Sciences
2004 – present Kidney Cancer Association Medical Advisory Board
Current support as PI

"Overcoming Drug Resistance in Cancer Targeted Therapy by Combining Genomics and Molecular Studies" 28/8/09 to 31/08/14 NMRC, SGD$5,500,000
"Whole-Genome Sequencing of Bile Duct Cancer or Cholangiocarcinoma (CCA) as a Singapore-led project of the International Cancer Genome Consortium", 2013-2016, NCCRF, donated by Mrs Irene Bronsveld & Ms Babara Cusick, 1,000,000 SGD

Name Song Ling POON
Mailing address Division of Medical Sciences, National Cancer Centre Singapore, 11 Hospital Drive, 169610 Singapore
Email songling.poon@gmail.com

Current position
Research Fellow, Division of Medical Sciences, National Cancer Centre Singapore
Academic qualifications B.Sc., National Cheng Kung University, Taiwan 2001; M.Sc., National Cheng Kung University, Taiwan; Ph.D., University of British Columbia, Canada, 2011
Research interests Translational research based on cancer cell biology and genomics; using high throughput screening to elucidate the roles of aberrant signalling pathways and interrogate potential novel therapeutic drugs in cancer.
Publications

1. **Poon SL**, McPherson JR, Tan P, Teh BT, Rozen SG. "Genome wide carcinogen signatures promise new opportunities for cancer prevention." (review) (submitted, 2013)
2. **Poon SL**, ..., Rozen SG, Tan P, Teh BT. "Genome wide mutational signatures of aristolochic acid and its application as a screening tool." Sci. Transl. Med. 5:197ra101 (2013) (Received journal cover image and focus article; also highlighted by Science Magazine and The Scientist) IF 10.8
3. Zang ZJ, Cutcutache I, **Poon SL**, ..., Rozen SG, Teh BT, Tan P. **S. G.** "Exome sequencing of gastric adenocarcinoma reveals recurrent somatic mutations in cell adhesion and chromatin remodeling genes." Nat. Genet. 44:570-74 (2012) IF 35.2
4. Koo GC, Tan SY, Tang T, **Poon SL**,.., Teh BT, Lim ST. "Janus inase 3-activating mutations identifies in natural killer/T-cell lymphoma. Cancer Discov. 2:591-7 (2012) IF 10.14
5. **Poon SL**, Hammond GL, Klaussen C, Leung PC. "37-kDa laminin receptor precursor mediates GnRH-II-induced MMP-2 expression and invasiveness in ovarian cancer cells." Mol. Endocrinol 25:327-28 (2011) (Received journal cover) IF 4.76
6. **Poon SL**, Lau MT, Hammond GL, Leung PC. "Gonadotropin-releasing hormone-II increases membrane type-I metalloproteinases production via beta-catenin signalling in ovarian cancer cells." Endocrinology 152:764-72 (2011) IF 4.71
7. **Poon SL**, Hammond GT, Leung PC. "Epidermal growth factor induced GnRH-II synthesis contributes to ovarian cancer cell invasion." Mol. Endocrinol. 10:1646-56 (2009) IF 4.76
8. **Poon SL**, An BS, So WK, Hammond GL, Leung PC. "Temporal recruitment of transcription factors at the 3',5'-cyclic adenosine 5'-monophosphate-response element of the human GnRH-II promoter." Endocrinology 149:5162-71 (2008) IF 4.71
9. Leu SF*, **Poon SL***, Pao HY, Huang BM. "The in vivo and in vitro stimulatory effects of Cordycepin on mouse Leydig cell steroidogenesis." Biosci. Biotechnol. Biochem. (2011) (*= co-first author) IF 1.3
10. Lin Q*, **Poon SL***,..., Leung PC. "Leptin interferes with 3',5'-cyclic adenosine monophosphate (cAMP) signalling to inhibit steroidogenesis in human granulosa cells." Reprod. Biol. Endocrinol. (2009) (*= co-first author) IF 2.14
11. An BS*, **Poon SL***, So WK, Hammond GL, Leung PC. "Rapid effect of GnRH-I on follicle-stimulating hormone β gene expression in LβT2 mouse pituitary cells requires the progesterone receptor." Biol. Reprod. 81:243-9 (2009) (*= co-first author) IF 4.02
12. Choi JH, Chen CL, **Poon SL**, Wang HS, Leung PC. "Gonadotropin-stimulated EGFR expression in human ovarian surface epithelial cells: involvement of cyclic AMP-dependent Epac pathway." Endocr Relat Cancer. 16:179-88 (2009) IF 5.26
13. So WK, Cheng JC, **Poon SL**, Leung PC. "Gonadotropin-releasing hormone and ovarian cancer: a functional and mechanistic overview". (review article) FEBS J. 275:5496-511 (2008) IF 4.25
14. Lin YM, **Poon SL**, Choi JH, Lin JS, Leung PC, Huang BM. "Transcript of testicular gonadotropin releasing hormone, steroidogenic enzymes, and intratesticular testosterone levels in infertile men." Fertil Steril. 90:1761-8 (2008) IF 4.17
15. Lin YM, Liu MY, **Poon SL**, Huang BM. "Gonadotropin releasing hormone-I and II stimulate steroidogenesis in prepubertal murine Leydig cells in vitro" Asian. J. Androl. 10:929-36 (2008) IF 2.14
16. **Poon SL**, Leu SF, Hsu HK, Liu MY, Huang BM. "Regulatory mechanism of Toona Sinensis on mouse Leydig cells in vitro." Life Science 13:1473-87 (2005) IF 2.55
17. Lo HC, Yang JG, Liu BC, Huang YL, Chen YW, **Poon SL**, Liu MY, Huang BM. "The effects of Tremella aurantia on testosterone and corticosterone production in normal and diabetic rats. Archives of Andrology 6:395-404 (2004) IF 0.81

Selected scientific awards
2013 4th DUNES Scientific Symposium, Duke-NUS, Best Oral Presentation, Singapore
2009 Child & Family Research Institute Outstanding Achievement of a Doctoral Student, Canada
2005 Outstanding Research Award in College of Medicine, National Cheng Kung University, Taiwan

Name: John Richard McPHERSON

Email: john.mcpherson@duke-nus.edu.sg

Position:Associate in Research,
Laboratory of Computational Systems Biology and Human Genetics (Rozen lab),
Cancer and Stem Cell Biology Program & Centre for Computational Biology,
Duke-NUS Graduate Medical School, Singapore

Education: Ph.D. University of Waikato, New Zealand (2007)
B. Sc. (Hons) (Computer Science) University of Canterbury, New Zealand (1999)

Fields: Bioinformatics/high-throughput genomic sequencing analysis and genetics (cancer, developmental disorders, germline diseases), mutational signatures in cancer, high-performance Computing.

Recent publications:

Abundant Hemizygous Deletions of CYCLOPS and STOP Genes in Gastric Adenocarcinoma
Cutcutache I., Wu Y.T., Suzuki Y., **McPherson J.R.**, Lei Z.D., Deng N.T., Wong W.K., Soo K.C., Chan W.H., Ooi L., Welsch R., Tan P., Rozen S.G.
(in review, Dec 2013)

Evaluation of a targeted gene panel for massively parallel sequencing of patients with intellectual disability, congenital anomalies and/or autism spectrum disorders
Brett M., **McPherson J.R.**, Zang Z.J., Lai A., Tan E.S., Ong L.C., Cham B., Tan P., Rozen S.G., Tan E.C.
(submitted, Dec 2013)

Alternative Splicing in Gastric Cancer
Liu Y.J., Tay S.T., Lee M.H., Wu J., Ramnaryanan K., **McPherson J.R.**, Wong W.K., Soo K.C., Tan P., Rozen S.G.
(in review, Nov 2013)

Exome sequencing identifies distinct mutational patterns in liver fluke-related and non-infection-related bile duct cancers
Chan-On W., Nairismägi M.L., Ong C.K., Lim W.K., Dima S., Pairojkul C., Lim K.H., **McPherson J.R.**, Cutcutache I., Heng H.L., Ooi L., Chung A., Chow P., Cheow P.C., Lee S.Y., Choo S.P., Tan I.B., Duda D., Nastase A., Myint S.S., Wong B.H., Gan A., Rajasegaran V., Ng C.C., Nagarajan S., Jusakul A., Zhang S., Vohra P., Yu W., Huang D., Sithithaworn P., Yongvanit P., Wongkham S., Khuntikeo N., Bhudhisawasdi V., Popescu I., Rozen S.G., Tan P., Teh B.T.
Nature Genetics vol 45:12, 1474–1478 (2013)

Genome-Wide Mutation Signatures of Aristolochic Acid and Its Application as a Screening Tool
Poon S.L.*, Pang S.-T.*, **McPherson J.R.***, Yu W., Huang K. K., Guan P., Weng W.-H., Siew E.Y., Liu Y.J., Heng H.L., Chong S.C., Gan A., Tay S.T., Lim W.K., Cutcutache I., Huang D.C., Ler L.D., Nairismägi M.-L., Lee M.H., Chang Y.-H., Yu K.-J., Chan-on W., Li B.-K., Yuan Y.-F., Qian C.-N., Ng K.-F., Wu C.-F., Hsu C.-L., Bunte R.M., Stratton M.R., Futreal P.A., Sung W.-K., Chuang C.-K., Ong C.K., Rozen S.G., Tan P., Teh B.T.
Science Translational Medicine vol 5:197ra101 (2013)
* denotes equal contribution

Exome sequencing of liver fluke-associated cholangiocarcinoma
Ong C.K., Subimerb C., Pairojkul C., Wongkham S., Cutcutache I., Yu W., **McPherson J.R.**, Allen G.E., Ng C.C., Wong B.H., Myint S.S., Rajasegaran V., Heng H.L., Gan A., Zang Z.J., Wu Y., Wu J., Lee M.H., Huang D., Ong P., Chan-on W., Cao Y., Qian C.N., Lim K.H., Ooi A., Dykema K., Furge K., Kukongviriyapan V., Sripa B., Wongkham C., Yongvanit P., Futreal P.A., Bhudhisawasdi V., Rozen S., Tan P., Teh B.T.
Nature Genetics vol 44:6, 690–693 (2012)

Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes
Zang Z.J., Cutcutache I., Poon S.L., Zhang S.L., **McPherson J.R.**, Tao J., Rajasegaran V., Heng H.L., Deng N., Gan A., Lim K.H., Ong C.K., Huang D., Chin S.Y., Tan I.B., Ng C.C., Yu W., Wu Y., Lee M., Wu J., Poh D., Wan W.K., Rha S.Y., So J., Salto-Tellez M., Yeoh K.G., Wong W.K., Zhu Y.J., Futreal P.A., Pang B., Ruan Y., Hillmer A.M., Bertrand D., Nagarajan N., Rozen S., Teh B.T., Tan P.
Nature Genetics vol 44:5, 570–574 (2012

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

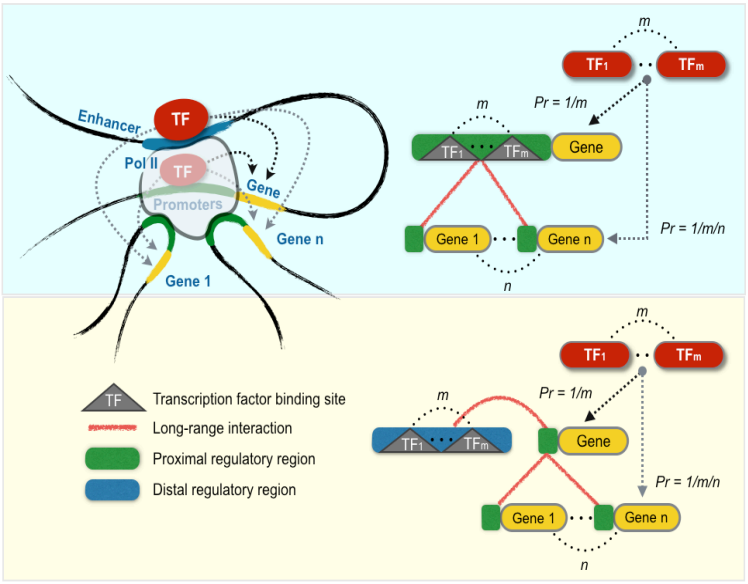
Three-dimensional and functional annotation of noncoding regulatory mutations

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hyung-Lae Kim, Ewha Womans University (Member of ICGC)
Keun-Chil Park, Samsung Medical Center (Member of ICGC)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Woojin Yang, KAIST Youngil Koh, Seoul National University Hospital	Jung Kyoon Choi, KAIST Jong-Sun Jung, Syntekabio, Inc.

Background and preliminary data



Transcription perturbation is thought to be the mechanism by which noncoding variations contribute to disease phenotypes (**Nature** 473:43-49; **Science** 337:1190-1195; **Cell** 152:633-641; **Science** 342:1235587). However, it is difficult to predict genes functionally affected by noncoding variations without systematic modeling of regulatory elements interacting with regulators and target genes. We established a probabilistic model that incorporated three-dimensional chromatin interactions connecting distal and proximal regulatory elements for 485 TFs (figure on the right). Functional regulator-target relationships were inferred from the physical interaction model based on our machine learning algorithm for Bayesian network analysis of ~1,400 TCGA gene expression profiles in breast cancer. Our breast cancer network revealed general properties of transcription regulation (especially via long-distance enhancers) and successfully identified the key regulators underlying subclass-specific gene expression programs and transcriptional response to 155 drug treatments. Our network is now being analyzed to predict the functional target genes of breast cancer risk loci identified by previous GWASs and noncoding mutations identified from the ICGC Breast Cancer Working Group (**Cell** 149:979-993, 2012; **Cell** 149:994-1007, 2012).

Timelines & resources dedicated to project

~March 2014: DNase I footprinting analysis of noncoding mutations

~April 2014: Publication of the current network model and findings in breast cancer

~May 2014: Development of an advanced regulatory network model

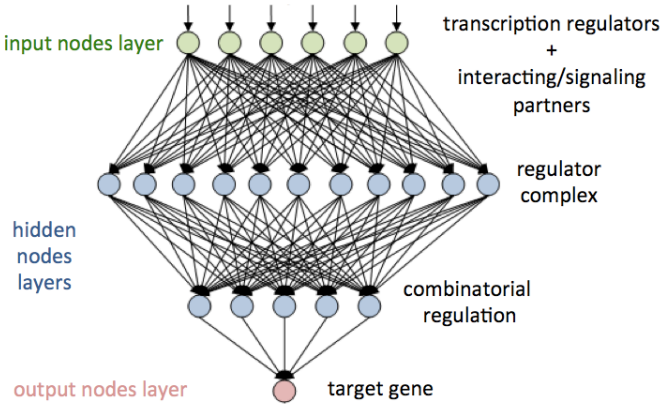
~June 2014: Identification of putative target genes based on the advanced network model and footprinting

~December 2014: Prioritization of epidriver candidates and prediction/validation of their functional effects

- WGS and microarray or RNA-seq data of ICGC breast cancer and leukaemia
- ChIA-PET or Hi-C data (ENCODE or published), DNase I footprints (ENCODE), TFBSs (ENCODE) in relevant cell lines (MCF7, T47D, K562, and NB4)

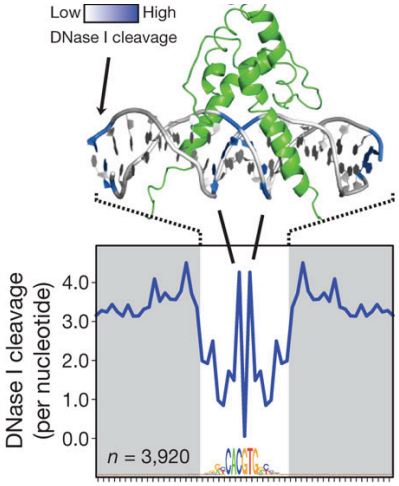
Research proposal

Development of an integrated network model



Our current network described above helps to predict the transcriptional target gene of noncoding variations on the basis of the physical and functional interaction of regulatory regions in the genome. The affected genes will contribute to the disease phenotype by participating in regulatory networks, signaling pathways, and protein-protein interactions. Transcription regulators are also regulated at the protein level via signaling pathway or protein interaction, which is not addressed in our current network model. Moreover, combinatorial regulation by two or more TFs cannot be inferred in the Bayesian networks. These necessitate integrated network modeling. Artificial neural networks are expected to provide a proper modeling framework in which transcription regulators and their interacting or signaling partners are engaged in combinatorial regulation by multiple TFs (figure above). This will be coupled with our previous Bayesian model and computational algorithm to infer causal relationships among the "meta-regulators" and target genes.

Filtering of regulatory and driver mutations



Deep sequencing of DNase I hypersensitive sites enables a fine-scale inspection of chromatin accessibility surrounding TF binding regions, a technique referred to as DNase I footprinting analysis (**Nature** 489:83–90, 2012). DNA sequences in direct contact with TFs binding to a non-canonical or unknown recognition motif can be identified at a nucleotide resolution (figure on the right). Therefore, regulatory mutations can be selected based on these data in addition to consensus motif information. A method to predict driver mutations is by screening the regulatory regions enriched for rare variants (**Science** 342:1235587, 2013). Mutation recurrence will be also used a criterion to filter driver mutations.

Prediction and validation of the functional consequences of high-priority mutations

Functional target genes of the predicted driver mutations can be inferred by exploring our integrated network. Genes that are physically connected to the mutation-containing regulatory regions (via three-dimensional chromatin interaction) and that are functionally associated with binding TFs or their protein partners (in terms of expression causality between the meta-regulator and target gene) will be identified. Downstream genes that are altered by the direct target genes will be predicted based on Bayesian causal network analysis. The putative epidriver genes and their downstream genes may be analyzed in various biological contexts, particularly in association with clinical traits such as tumor subclasses, prognosis, and treatment outcome. Experimental validation by a multiplex genome-editing technique such as CRISPR followed by DNA microarray analysis will be performed for selected epidriver candidates.

Legacy plans

The following can be provided for other ICGC working groups.

- Integrated networks in breast cancer and leukaemia (can be used to understand the mechanism of action of known or newly identified driver or epidriver genes)

- List of predicted epidriver genes in breast cancer and leukaemia

- Algorithm and/or software for the construction of the integrated networks in other cancers (including a version that can be used without DNase I footprinting and chromatin interaction data)

Curriculum Vitae

Hyung-Lae Kim, MD, Ph.D

School of Medicine
Ewha Wonans Uiversity
Mok-5-Dong, Yangcheon-Gu,
Seoul , Korea
hyung@ewha.ac.kr
+82-2-2650-5727

CAREER & EDUCATION

2012 ~, Executive Committee Member, International Scientific Steering Member, ICGC

2011 ~, Director General, The National Project for Personalized Genomic Medicine, Korea

2013 ~, President, Korean Society of Biochemistry and Molecular Biology

2011, President, Korea Genome Organization

2008 ~2010, Director General, Korea National Institute of Heath, Korea

2006 ~ 2008, Director, Center for Genome Science, Korea National Institute of Heath, Korea

1995 ~, Professor, Dept. Biochemistry, School of Medicine, Ewha Womans University, Korea

1989 ~ 2002, Visiting Fellow, Lab Mol Biology, NINDS, NIH, USA

1986, PhD in Biochemistry, Graduate School, Seoul National University, Korea

PUBLICATION (5 years)

[1] Kim et al. A genome-wide association study of a coronary artery disease risk variant
J Hum Genet. 58(3):120-6 (2013).

[2] Kim et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. (2012) Mar 27;14(2):R56

[3] Cho et al. Meta-analysis of genome-wide association studies identifies 8 new loci for type 2 diabetes in East AsiansNat Genet. 44(1): (2011) 67-72

[4] The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret et al.Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367) (2011):103-9.

[5] Kim et al. Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. Nat Genet. 43(10): (2011); 990-5.

[6] Fox et al. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study.Hum Mol Genet. 20(2011): 2273-84.

[7] Jung et al. Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 5(7) (2010): e11855

[8] HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 326 (2009):1541-5.

[9] Cho et al. A large scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics 41(2009):527-34.

Curriculum Vitae

Keunchil Park, M.D., Ph.D.

Keunchil Park is Professor of the Division of Hematology-Oncology, Sungkyunkwan University School of Medicine in Seoul, Korea. Prof. Park is Director of the Medical Nano Element Development Center, and is the Principal Investigator of the 'Identification of Novel Therapeutic Targets in Lung Cancer with Unmet Need' of the National Project for Personalized Genomic Medicine(PGM21), both of which are funded by the Ministry of Health and Welfare, Korea.

Professor Park has served many domestic academic societies, e.g., Chair of the Scientific Committee of the Korean Cancer Association, Chair of the Lung Cancer Committee of the Korean Cancer Study Group(KCSG). Prof. Park also served as Chairman of the Board of Directors, Korean Association for Clinical Oncology (KACO) since June 2010 until May 2012.

Prof. Park has been also very actively involved in and served many international activities, such as the Scientific Secretary of the 12th WCLC (Sept, 2007), and the Chairman of the 4th Asia Pacific Lung Cancer Conference (Dec, 2010). He was elected as the Board of Directors of the IASLC and is serving as Associate Editor for the Journal of Thoracic Oncology (JTO) and on editorial board of the Asia-Pacific Journal of Clinical Oncology.

Prof. Park's main interests include the translational and early clinical researches for the treatments of upper aero-digestive tract cancers, especially lung cancer. Recently Dr. Park is leading several early clinical trials of the targeted agents as well as many pre-clinical development programs internationally. Prof. Park has several book chapters and authored more than 200 peer-reviewed publications in national and international journals.

Woojin Yang
(December 2013)

E-mail: boodmain@gmail.com

Academic Degrees

2012-Present	Ph.D.,	Department of Bio and Brain Engineering, KAIST
2000-2003	M.S.,	Department of Computer Science and Engineering, Seoul National University
1996-2000	B.S.,	Department of Computer Science, Seoul National University

Employment Record

2012~2011 **Senior Researcher, Future IT R&D Lab, LG Electronics**
Cloud support for Home Energy Management System (HEMS)
Connectivity between HeMS and Renewable Energy Resources

2011~2009 **Manager, SW Development Section, SeAH ICT**
Architecture design for multi-processor ASN-GW system (support for high-capacity and high-availability)
Enhancing reliability of packet classification and processing
Optimizing performance and architecture of multi-core ASN-GW

2009~2007 **Manager, SW Development Team, WiMAX R&D Center, POSDATA**
Architecture design for multi-core ASN-GW
Requirement analysis and high level design
Design and development of data structure with shared memory of 16 CPU cores
Design and development of WIMAX packet processor (parallel processing in 12 CPU cores)
Development of Firewall algorithm with modified RFC (Recursive Flow Classification)
Performance optimization of packet processor (modifying algorithm, enhancing concurrency between CPU cores, compiler and cache memory optimization)

2007~2003 **Research Engineer, ETRI (Electronics and Communications Research Institute)**
Development project: IPv6 Router supporting IP Mobility (2006-2007)
- Architecture design of mobile IPv6 router (interfaces and shared data structure)
- Test-bed construction of IPSec protocol (routers and mobile terminals)

Coordinated a joint project titled "Research for Seamless Handover and Service by Optimizing Handover" (with INC Lab, Dept. of EECS, Seoul National University) (2006)
Coordinated international joint project titled "Research for Traffic Engineering in MPLS network" (with Posts and Telecommunications Institute of Technology, Veitnam) (2006)
Development project: IPv6 router supporting Quality of Service (2005-2006)
- Design and development of data structure for managing interfaces
- Optimization in communication and event-driven synchronization of network system
- Development of control module for switch fabric in multi-processor network system

Development project: IPv6 router (IPv6 Router S/W Team, 2004)
- Research in algorithm and data structure for enhancing performance of traffic monitoring with IPFIX (IP Flow Information Export)
- Development of MPLS label switching in Intel IXP2400 network processor
- Optimization in connection between routing protocol and kernel (RSVP-TE and LDP)

Development project: 10Gb Ethernet Switch (10GE S/W Team, 2003)
- Performance optimization of IPC (Inter-Processor Communication) module
- Research in communication and synchronization of data structure between line-card processor and routing-processor
- Development in packet forwarding with LACP (Link Aggregation Control Protocol) in IBM Network Processor

Presentations & Papers

2007 **Woo-jin Yang**, Tae-il Kim, Hae-won Jung, "Method to Provide Host Mobility and Handover Based on Distributed Proxy," *International Conference on Advanced Communication Technology 2007*

Sungro Yoon, Jiwoong Jeong, Chong-kwon Kim, **Woo-jin Yang**, Tae-il Kim, Hae-won Jung, "New Approach for Reducing DAD delay using Link Layer Assistance in Mobile IPv6," *International Conference on Multimedia and Ubiquitous Engineering 2007*

Hayoung Oh, Kibaek Yoo, Chong-kwon Kim, **Woo-jin Yang**, Tae-il Kim, Hae-won Jung, "An *Enhanced* Fast Handover Scheme with Temporal Reuse of CoAs and PBP in IPv6-Based Mobile Networks," *International Conference on Multimedia and Ubiquitous Engineering 2007*

Tran Cong Hung, Nguyen Hoang Thanh, Nguyen Duc Thang, Hae Won Jung, Tae Il Kim, Sung Hei Kim, **Woo Jin Yang**, "Advanced Routing Algorithms and Load Balancing on MPLS," *International Conference on Advanced Communication Technology 2007*

2006 **Woo-jin Yang**, *Tae*-il Kim, Hae-won Jung, "Optimizing Hash Table Structure of Flow Exporting Software," *International Conference on Advanced Communication Technology 2006*

Patents

2008 Method and Apparatus for Transmitting IP Packet in Network Based on Tunneling

Apparatus and method for classifying packet in wideband wireless communication system

2007 APPARATUS AND METHOD FOR MANAGING ADDRESS TO PROVIDE HOST MOBILITY IN NETWORK

Apparatus for managing mobility of mobile terminal in distribution and method using the same

2006 Apparatus for controlling cursor on display and method there of

2005 Apparatus and method for measuring per-flow information of traffic

2003 Interface module for implementing single high speed interface by aggregating plurality of low speed interfaces and communication device including the same

Curriculum Vitae

Youngil Koh, MD

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
Go01@chol.com
+82-1091175012

CAREER & EDUCATION

2013.5 - Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea
2010.3 - 2013.4 Public Service Doctor, Kkotdongnae, Gapyeong, Korea (Military service)
2010.3 - PhD. Candidate, Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Masters in Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
2005.2, MD, Seoul National University College of Medicine, Seoul, Korea

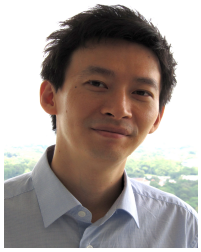
AWARDS

2012 Best Oral Presentation Award, Korean Association for Clinical Oncology Annual Meeting
2011 Best Oral Presentation Award, Korea Cancer Association Annaul Meeting
2008 Travel Award, American Society of Hematology Annual Meeting
2008 Best doctor for patients, Seoul National University Hospital
1998 Bronze medal, 39th International Mathematics Olympiad, Taiwan
1997 Silver medal, 38th International Mathematics Olympiad, Argentina

PUBLICATION (recent 2 years, 1st author only, including co-first author)

[1] Park S, Koh Y, Jung SH, Chung YJ. Application of array comparative genomic hybridization in chronic myeloid leukemia. Methods in molecular biology 2013;973:55-68.
[2] Kim I, Koh Y, Yoon SS, et al. Fludarabine, cytarabine, and attenuated-dose idarubicin (m-FLAI) combination therapy for elderly acute myeloid leukemia patients. American journal of hematology 2013;88:10-5.
[3] Koh Y, Lim HY, Ahn JH, et al. Phase II trial of everolimus for the treatment of nonclear-cell renal cell carcinoma. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2013;24:1026-31.
[4] Koh Y, Kim I, Shin DY, et al. Polymorphisms in genes that regulate cyclosporine metabolism affect cyclosporine blood levels and clinical outcomes in patients who receive allogeneic hematopoietic stem cell transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation 2012;18:37-43.
[5] Koh Y, Lee HE, Oh DY, et al. The lack of CD34 expression in gastrointestinal stromal tumors is related to cystic degeneration following imatinib use. Japanese journal of clinical oncology 2012;42:1020-7.

JUNG KYOON CHOI



Dept. Bio and Brain Engineering
KAIST
335 Gwahak-ro, Yooseong-goo
Daejeon 305-701
Republic of Korea
T +82-42-350-4327
F +82-42-350-8834
jungkyoon@kaist.ac.kr
http://omics.kaist.ac.kr

ASSISTANT PROFESSOR — OCT 2009 ~ PRESENT
Dept. Bio and Brain Engineering, KAIST

PRINCIPAL INVESTIGATOR — JAN 2010 ~ DEC 2012
Genome Institute of Singapore (Joint appointment)

Publications - As corresponding or first author

Genome-wide reorganization of histone h2AX toward particular fragile sites on cell activation. *Nucleic Acids Res.* **in press.**

Regulation of the boundaries of accessible chromatin. *PLoS Genet.* **9**, e1003778 (2013).

Genetic landscape of open chromatin in yeast. *PLoS Genet.* **9**, e1003229 (2013).

Genome-wide profiles of H2AX and gamma-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. *Nucleic Acids Res.* **40**, 5965-5974 (2012).

Controlling transcriptional programs for cellular adaptation by chromatin regulation. *Mol. BioSyst.* **7**, 1713-1719 (2011).

Genetic and metabolic characterization of insomnia. *PLoS ONE* **6**, e18455 (2011).

Systems biology and epigenetic gene regulation. *IET Syst. Biol.* **4**, 289-295 (2010).

Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol.* **11**, R70 (2010).

Nucleosome deposition and DNA methylation at coding region boundaries. *Genome Biol.* **10**, R89 (2009).

Implications of the nucleosome code in regulatory variation, adaptation and evolution. *Epigenetics* **4**, 291-295 (2009).

Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat. Genet.* **41**, 498-503 (2009).

Stochastic and regulatory role of chromatin silencing in genomics response to environmental changes. *PLoS ONE* **3**, e3002 (2008).

Epigenetic regulation and the variability of gene expression. *Nat. Genet.* **40**, 141-147 (2008).

Environmental effects on gene expression phenotype have regional biases in the human genome. *Genetics* **175**, 1607–1613 (2007).

Impact of transcriptional properties on essentiality and evolutionary rate. *Genetics* **175**, 199–206 (2007).

Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21**, 4348-4355 (2005).

Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett.* **565**, 93-100 (2004).

Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19** Suppl. 1, i84-i90 (2003).

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

Abstract of proposed research for WGS pan-cancer analysis
Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Ethnic specific risk prediction for blood and liver cancers by comparing the ICGC data to 1000 Genomes Project data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hyung-Lae Kim, *PGM 21*; Sungsoo Yoon, *Seoul National University Hospital, Seoul, Republic of Korea*

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Seong Gu Heo & Eun Pyo Hong, *Hallym University College of Medicine, Chuncheon, Republic of Korea*	Ji Wan Park, *Hallym University College of Medicine;* Jongsun Jun, *Synteca Bio*

Background and preliminary data

Background:
In previous indirect genome wide association studies, meta-analysis has been widely used to increase statistical power by combining results of independent studies. Currently, the ICGC completed catalogues of somatic mutations and provides both common and ethnic specific mutations in 6 and 4 independent projects for blood and liver cancers, respectively. Integrating individual datasets will increase statistical power; however, also could lead to the reduction in statistical power caused by genetic heterogeneity across study populations. Heritability of a disease is a key population parameter to assess the genetic contributions to a population's phenotypic variance. Heritability can be estimated in the lineage of germ cells and disease causal variants can be identified by comparing patients and healthy controls. We could search for germline mutations by comparing the ICGC data to the reference panels of 1000 Genomes projects for different ethnic groups using linear mixed model (LMM), which has become a standard method to account for the confounding effects of population structure and cryptic relatedness (*Nat Genet*. 44:821, 2012). Liability refers to the individual's innate tendency to develop a disease. When the liability exceeds a specific threshold, it leads to the development of disease. The liability threshold (LT) model can be generalized to assess clinical risk factors (*Am J Hum Genet.* 88:548, 2011). We will develop liability threshold models to predict risk/prognosis of blood and liver cancers.

Preliminary Data:
Our group has studied on disease prediction models using GWAS. In a preliminary study, we analyzed whole exome sequencing data of 65 normal/tumor sample pairs obtained from Korean patients with AML. We found 4,962 somatic point mutations from the samples using MuTect v1.1.4. We evaluated the associations between the mutations and AML with PLINK/SEQ v0.08. We also identified subtype specific mutations of AML in 17, 12, and 11 normal/tumor pairs of M2, M3, and M4, respectively, and developed models to predict the subtype of AML (i.e., 93, 97, 93, and 90% accuracy, respectively).

Timelines & resources dedicated to project

Timelines:
12/2013 – 03/2014: Detect germline and somatic mutations for each tumor type.
03/2014 – 06/2014: Select potent mutations for each tumor type.
07/2014 – 08/2014: Develop genetic risk models for each tumor type and ethnic groups.
09/2014 – 12/2015: Evaluate and validate the risk prediction models.
01/2015 – 03/2015: Manuscript writing and submission.

Resources:

- **Blood cancer: Acute Lymphoblastic Leukemia - US (United States, 229), Chronic Lymphocyclic Leukemia - ES (Spain, 264), Acute Myeloid Leukemia - TCGA, US (200), KR (South Korea, 55), Chronic Myeloid Disorders - UK (United Kingdom, 129), Malignant Lymphoma - DE (Denmark, 53), and their survival data.**
- **Liver cancer: Liver Hepatocellular carcinoma – TCGA (US 73), Liver Cancer - NCC, JP (Japan 213), RIKEN, JP (Japan, 45), FR (France, 30), and their survival data.**
- 1000 genome reference data – European penal (EUR 379), East Asian panel (ASN 286: CHS 100, CHB 97, JPT 89)

Research proposal

We will estimate heritability and search for both common and ethnic specific germline mutations by comparing the ICGC data to the reference panels of 1000 Genomes project for different ethnic groups (i.e., Europeans and East Asians) available for blood and liver cancers. We will also develop genetic risk models for prediction of developing blood and liver cancers and prognostic models (i.e., survival).

Step 1. Detection of germline and somatic mutations

The Genome Analysis Toolkit and MuTect will be used to call germline and somatic mutations, respectively. We will estimate heritability and test association of these mutations with the diseases of interest using LMM. Firstly, we will look at susceptibility germline mutations for blood cancer by comparing between 930 cases with blood cancer (ICGC) and 665 samples without blood cancer obtained from the 1000 Genomes project. Secondly, we will compare 361 cases with liver cancer (ICGC) to 665 reference samples to identify germline mutations for liver cancer. We will examine somatic mutations in 930 normal/tumor pairs of blood cancer and 361 normal/tumor pairs of liver disease, respectively. Multivariate analyses will be performed in all subjects and subgroups stratified by ethnicity and subtype of each cancer.

Step 2. Heterogeneity test

We will examine the consistency in the results across studies based on I^2 and Cochran's Q-statistics for both germline and somatic mutations; and then will identify ethnic- and subtype-specific mutations. The variants passed heterogeneity test will be further validated in the combined data set.

Step 3. Survival analysis

We will perform survival analysis using germline and somatic mutations selected for blood and liver cancers. We will compare the survival probability of the patients according to having mutations by using Kaplan-Meier method with log-rank test. Those mutations that show significant difference between case/control status and/or tumor types will be included in the ethnic specific risk assessment model.

Step 4. Risk assessment model

We will develop ethnic specific risk assessment model based on liability threshold model. The LT model calculates both life-time risk and the risk within a specified period of time (e.g., 1 year- or 5 year-risk) for the occurrence or the prognosis of a certain type of cancer. The model will be evaluated in each of Asian and European data sets with the area under the receiver-operating characteristic curve. We can develop cancer risk models using in-house programs written in R and Python.

Legacy plans

1. **Provide n×n genotype matrix for the ICGC cancer patients and 1000 Genomes controls.**
2. **Provide germline and somatic mutation calling pipeline.**
3. **Provide the script for association analysis using linear mixed model**
4. **Provide the script for testing heterogeneity and survival analysis.**
5. **Provide the script to develop liability threshold model.**

Curriculum Vitae

Hyung-Lae Kim, MD, Ph.D

School of Medicine
Ewha Wonans Uiversity
Mok-5-Dong, Yangcheon-Gu,
Seoul , Korea
hyung@ewha.ac.kr
+82-2-2650-5727

CAREER & EDUCATION

2012 ~, Executive Committee Member, International Scientific Steering Member, ICGC

2011 ~, Director General, The National Project for Personalized Genomic Medicine, Korea

2013 ~, President, Korean Society of Biochemistry and Molecular Biology

2011, President, Korea Genome Organization

2008 ~2010, Director General, Korea National Institute of Heath, Korea

2006 ~ 2008, Director, Center for Genome Science, Korea National Institute of Heath, Korea

1995 ~, Professor, Dept. Biochemistry, School of Medicine, Ewha Womans University, Korea

1989 ~ 2002, Visiting Fellow, Lab Mol Biology, NINDS, NIH, USA

1986, PhD in Biochemistry, Graduate School, Seoul National University, Korea

PUBLICATION (5 years)

[1] Kim et al. A genome-wide association study of a coronary artery disease risk variant
J Hum Genet. 58(3):120-6 (2013).

[2] Kim et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. (2012) Mar 27;14(2):R56

[3] Cho et al. Meta-analysis of genome-wide association studies identifies 8 new loci for type 2 diabetes in East AsiansNat Genet. 44(1): (2011) 67-72

[4] The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret et al.Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367) (2011):103-9.

[5] Kim et al. Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. Nat Genet. 43(10): (2011); 990-5.

[6] Fox et al. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study.Hum Mol Genet. 20(2011): 2273-84.

[7] Jung et al. Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 5(7) (2010): e11855

[8] HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 326 (2009):1541-5.

[9] Cho et al. A large scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics 41(2009):527-34.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012

[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7

[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7

[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea

1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea

1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.

1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.

1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea

1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press

[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.

[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7

[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.

[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.

[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

Curriculum Vitae

Seong Gu Heo, MSc

Department of Medical Genetics
College of Medicine
Hallym University
Chuncheon , Korea
lukeheo@hallym.ac.kr
http://web.hallym.ac.kr/~de1688/
T: +82-33-248-2693
F: +82-33-248-2693

PROFESSIONAL EXPERIENCE

[1]Developed T2DM risk prediction models based on liability threshold model using Python, 2013
[2]Physics data analysis using Python, C++, Linux, and Grid Computing of CERN, 2009-2011
[3]Sun Certified Java Programmer, 2000

CAREER & EDUCATION

2012~present, PhD student in Medical Genetics, Hallym University, Chunchon, Korea
2009~2011, MSc in Physics, Kangwon National University, Chunchon, Korea
2007~2008, Assistant System Manager, eScience Centre, Cambridge University, UK
2004~2005, IT Support, the Oxford Group, London, UK
2003~2004, Computer Network Engineer, Sekyee Computer Ltd, London, UK
1988~1992, BSc in Physics, Kangwon National University, Chunchon, Korea

PUBLICATION

[1]Heo et al, Male-specific genetic effect on hypertension and metabolic disorders. Hum. Genet. 2013 October;doi:10.1007/s00439-013-1382-4
[2]Heo et al, Genetic Risk Prediction for Normal-Karyotype Acute Myeloid Leukemia Using Whole Exome Sequencing. Genomics Inform. 2013 Mar;11(1):46-51.
[3]Chatrchyan et al, Inclusive search for supersymmetry using raxor variables in pp collisions at 7 Tev, Phys Rev Lett. 2011 Nov 4;107(19):191802.

Curriculum Vitae

Eunpyo Hong, MSc

Dept. of Medical Genetics
College of Medicine
Hallym University
Chuncheon 200-702, Korea
ghddmsvy0305@hallym.ac.kr
http://web.hallym.ac.kr/~de1688/
T: +82-33-248-2693
F: +82-33-248-2693

PROFESSIONAL EXPERIENCE

[1]Python language (Middle level).
[2]Sequencing analysis (Middle level): BI Award, Risk prediction for NK-AML using Whole Exome Sequencing, KOGO & PGM21, Feb. 2013.
[3]GWAS and Meta-analysis; Gene-Environment interaction analysis.
[4]Survival study (follow-up data analysis).
[5]Statistical S/W: R, STATA, PLINK, PLINK/SEQ, GCTA, etc.

CAREER & EDUCATION

2012~, Ph.D. Student in Medical Genetics, Hallym University, Korea
2009~2012, MSc in Medical Genetics, Hallym University, Korea
2008~2009, Apprentice in Dpt. of Medical Genetics, Hallym University, Korea
2002~2006, BSc in Information Statistics, Hallym University, Korea

PUBLICATION (5 years)

[1]Hong et al, Sample Size and Statistical Power Calculation in Genetic Association Studies. Genomics Inform. 2012 Jun;10(2):117-22.
[2]Hong et al, Analyses of Longitudinal Effects of Gene-Environment Interactions on Plasma C-reactive Protein Levels: The Hallym Aging Study. Genes & Genomics. 2013 Jan;35(1):131-139.
[3]Heo et al, Genetic Risk Prediction for Normal-Karyotype Acute Myeloid Leukemia Using Whole Exome Sequencing. Genomics Inform. 2013 Mar;11(1):46-51.

Curriculum Vitae

Ji Wan Park, MPH, PhD

Hallym University Colledge of Medicine
1 Hallymdaehak-gil,
200-702, Chuncheon, Korea
jwpark@hallym.ac.kr
+82- 33-248-2691

PROFESSIONAL EXPERIENCE

2010-present Committee member, Korean Genome Organization
2007-present Primary Instructor, Asian Institute in Statistical Genetics and Genomics
2006 Travel Award, the 2nd North American Congress of Epidemiology

CAREER & EDUCATION

2009–present Associate professor, Dept. of Medical Genetics, Hallym Univ. College of Medicine
2006-2009 Senior Scientist, Samsung Biomedical Research Institute, Seoul, Korea
2005-2006 Post-doctoral fellow/Instructor, Dept. of Epidemiology, Johns Hopkins University, Baltimore, MD, USA
2000–2005 Ph.D. in Epidemiology (Human Genetics/Genetic Epidemiology Track), The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
1987–1989 MPH in Epidemiology, Seoul National University, School of Public Health, Seoul, Korea
1983–1987 BS, Korea University, Seoul, Korea
1990-1992 Epidemiologist, US Army Medical Research Center, Seoul, Korea

PUBLICATIONS (Selected)

[1]Heo SG, Hwang JY, Uhmn S, Go MJ, Oh B, Lee JY, Park JW. Male-specific genetic effect on hypertension and metabolic disorders. Hum. Genet. 2013 October;doi:10.1007 (Epub ahead of print).
[2]Heo SG, Hong EP, Park JW. Genetic Risk Prediction for Normal-Karyotype Acute Myeloid Leukemia Using Whole Exome Sequencing. Genomics Inform. 2013;11:46-51.
[3]Yoon D, Park SK, Kang D, Park T, Park JW. Meta-analysis of homogeneous sub-groups reveals association between PDE4D gene variants and ischemic stroke. Neuroepidemiology. 2011;36:213-22.
[4]Jee SH, Sull JW, Lee JE, Shin C, Park J, Kimm H, Cho EY, Shin ES, YUN JE, Park JW, et al. Adiponectin concentrations: a genome-wide association study. Am J Hum Genet. 2010;87:545-52.
[5]Cho YS, Go MJ, Kim YJ,…, Park JW, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet. 2009;41:527-34.
[6]Park JW, et al. Family history of diabetes and risk of atherosclerotic cardiovascular disease in Korean men and women. Atherosclerosis 2008;197:224-31.
[7]Park JW, et al. BMI and stroke risk in Korean women. Obesity 2008;16:396-401.
[8]Park JW, et al. Association between IRF6 and nonsyndromic cleft lip with or without cleft palate in 4 populations. Genet Med. 2007;9:219-27.
[9]Park JW, et al. High throughput SNP and expression analyses of candidate genes for nonsyndromic oral clefts. J Med Genet. 2006;43:598-608.
[10]Park JW, et al. Comparing whole genome amplification methods and sources of biological sample for single-nucleotide polymorphism genotyping. Clin Chem. 2005;51:1520-3.

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science
and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Genetic variation profiling based on network module across pan-cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)

Hyung-Lae Kim, ***PGM 21, Director***
Sungsoo Yoon, ***Seoul National University Hospital, Seoul, Republic of Korea***

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Youngwook Kim, Samsung Medical Center	**Kiejung Park, KRIBB (Korea Research Institute of Bioscience and Biotechnology)** **Jongsun Jung, DCC in PGM21, Korea**

Background and preliminary data

Recently, high-throughput sequencing technologies and computational analysis pipelines can identify multiple mutation genes. Due to technical limitation and poor algorithms, genetic variations could contain false positives. Gene level approach is insufficient to find comprehensive biological meaning of cancer mechanism and characteristics. Hence, we need to develop new approach to understand multi-omics data sets.

We focused on identify and investigate functional relationships among cancer casual genes. Remarkably, the pathway-driven approach can identify the gene set related to the diseases, and the cancer driver genes from protein-protein interactions.

Therefore, we present the pathway-based network analysis, Human integrated Functional Interaction (HiFI). This identifies core network modules, candidate oncogenic processes and biological pathways related to specific cancer. We try to show profile of ethnic specific mutation spectrum using the gene set of pan-cancer data.

Timelines & resources dedicated to project

a. Target variation calling – January 2014
b. Gene Set selection - January to February 2014
c. Ethnic specific level calculation – March 2014
d. Hi-Fi network module update – April to May 2014
e. Driver gene selection by Hi-Fi – May to June 2014
f. Profiling – June to July 2014
g. Manuscript preparation – August to December 2014

Research proposal

We constructed the Human functional interaction network extending curated Pathways (hiPathDB) using multiple human protein-protein interaction (PPI) sources and gene co-expressions. In order to compensate low-coverage interactions with confident pathways and pairwise interactions, we used a supervised learning method that can provide accurate predictions even with violation of strong independence assumptions between the features.

To assess the usability of our Human integrated Functional Interaction network (HiFI), we conducted an integrated network analysis to identify frequently mutated gene network modules and candidate driver mutations from the TCGA GBM (glioblastoma) data set. We obtained mutation genes from 91 GBM cases, for which both CNV and somatic mutation data were available, also included genes mutated in two or more of the cases. Total 582 genes passed the threshold and 259 genes had interactions in the HiFI and then projected these 259 genes into the functional interaction network.

As a result, we identified the sub-network with 64 GBM mutated genes and 6 linker genes. Lastly, we applied network clustering algorithm to the obtained GBM mutation network and produced 10 core modules. From these identified GBM modules, we confirmed identical driver mutations reported in previous TCGA studies. In addition, these identified modules using our HiFI system were enriched GBM-related functional pathways such as PI3K/AKT, RB and p53 tumor suppressor pathway.

To investigate pathway modules with the genes have mutational association with multiple cancers, we select the set of multiple cancer-related genes including oncogenes. Using our HiFI system, we may identify the ethnic specific mutations in pan-cancer genome and can extend to the target module across multiple cancers.

Legacy plans

a. **Raw data (Neo4j Graph database) for HiFI system**
b. Webservice image and documentation of HiFI
c. HiFI output
d. RStudio/R source code

Curriculum Vitae

Hyung-Lae Kim, MD, Ph.D

School of Medicine
Ewha Wonans Uiversity
Mok-5-Dong, Yangcheon-Gu,
Seoul , Korea
hyung@ewha.ac.kr
+82-2-2650-5727

CAREER & EDUCATION

2012 ~, Executive Committee Member, International Scientific Steering Member, ICGC

2011 ~, Director General, The National Project for Personalized Genomic Medicine, Korea

2013 ~, President, Korean Society of Biochemistry and Molecular Biology

2011, President, Korea Genome Organization

2008 ~2010, Director General, Korea National Institute of Heath, Korea

2006 ~ 2008, Director, Center for Genome Science, Korea National Institute of Heath, Korea

1995 ~, Professor, Dept. Biochemistry, School of Medicine, Ewha Womans University, Korea

1989 ~ 2002, Visiting Fellow, Lab Mol Biology, NINDS, NIH, USA

1986, PhD in Biochemistry, Graduate School, Seoul National University, Korea

PUBLICATION (5 years)

[1] Kim et al. A genome-wide association study of a coronary artery disease risk variant
J Hum Genet. 58(3):120-6 (2013).

[2] Kim et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. (2012) Mar 27;14(2):R56

[3] Cho et al. Meta-analysis of genome-wide association studies identifies 8 new loci for type 2 diabetes in East AsiansNat Genet. 44(1): (2011) 67-72

[4] The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret et al.Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367) (2011):103-9.

[5] Kim et al. Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. Nat Genet. 43(10): (2011); 990-5.

[6] Fox et al. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study.Hum Mol Genet. 20(2011): 2273-84.

[7] Jung et al. Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 5(7) (2010): e11855

[8] HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 326 (2009):1541-5.

[9] Cho et al. A large scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics 41(2009):527-34.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012
[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7
[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7
[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea
1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea
1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.
1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.
1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7
[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.
[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.
[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

Curriculum Vitae

Youngwook Kim

Samsung Biomedical Research Institute

Senior Researcher

Rm.188 B4 Cancer Center

50 Irwon-dong Gangnam-gu

Seoul, Korea zip: 135-710

Office: 82-2-2148-7349

Cell: 82-10-2300-9856

Publications

1. **Kim Y**, Hammerman P, Kim JG, Yoon J, Lee Y, Sun J, Wilkerson M, Pedamallu C, Cibulskis K, Yoo Y, Lawrence M, Stojanov P, Carter S, Hayes N, Getz G, Meyerson M, Park K Integrative and comparative genomic analysis of lung squamous cell carcinomas in East-Asians Journal of Clinical Oncology *in press (2014)*

2. **Kim Y**, Kim J, Lee J, Bae K, Min J, Park T, Lee J, Nam Y, Park K *Tumor-Targeted Delivery of Paclitaxel using Solid Lipid Nanoparticles* Nature Communications *in review*

3. **Kim Y**, Ko J, Cui Z, Abolhoda A, Ahn JS, Ou SH, Ahn MJ, Park K *The EGFR T790M mutation in acquired resistance to an irreversible second-generation EGFR inhibitor* Mol Cancer Ther. (2012) Mar;11(3):784-91.

4. Lee S*, **Kim Y***, Sun JM, Choi YL, Kim JG, Shim YM, Park YH, Ahn JS, Park K, Han JH, Ahn MJ *Molecular profiles of EGFR, K-ras, c-met, and FGFR in pulmonary pleomorphic carcinoma, a rare lung malignancy*. J Cancer Res Clin Oncol. (2011) Aug;137(8):1203-11. * equally contributing authors

5. Oh YH*, **Kim Y***, Kim YP, Seo SW, Mitsudomi T, Ahn MJ, Park K, Kim HS *Rapid detection of the epidermal growth factor receptor mutation in non-small-cell lung cancer for analysis of acquired resistance using molecular beacons.* J Mol Diagn. (2010) Sep;12(5):644-52. *equally contributing authors

Curriculum Vitae

Kiejung Park, Ph.D
KRIBB(Korea Research Institute of Bioscience and Biotechnology),
KOBIC(Korean Bioinformation Center),
125 Gwahak-ro, Yuseong-gu, Daejeon, 306-809, Korea
kjpark63@gmail.com
+82-42-879-8500

PROFESSIONAL EXPERIENCE

<Bioinformatics Research>

2009.12 – 2013.1, Div. of Bioinformatics/Korea NIH (Director)
2000. 3 - 2009.12, SmallSoft Co., Ltd. (CEO/CTO)
1998. 6 - 2000. 2, KAIST Medical Science Center (researcher)
1990.11 - 1991. 2, Japan RIKEN (visiting researcher)
1989. 2 - 1998. 6, KRIBB (senior researcher)

CAREER & EDUCATION

2013.1 – current, KRIBB(principal investigator), KOBIC(director)
1982. 3 - 1986. 2, Dep. Of Computer Science & Engineering, Seoul National University (Bs.E.)
1986. 3 - 1987. 2, Dep. Of Computer Science, KAIST (master course)
1987. 3 - 1989. 2, Dep. Of Biological Science, KAIST (M.S./bioinformatics)
1991. 3 - 2002. 2, Dep. Of Biological Science, KAIST (Ph.D./bioinformatics)

PUBLICATION (5 years)

[1] Park et al, An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data, Bioinformatics, 2012, Volume 28, Issue 24Pp. 3306-3315(2012)
[2] Park et al, Genovar: a detection and visualization tool for genomic variants, BMC Bioinformatics, 2012,13(Suppl 7):S12
[3] Park et al, Identification of methylation-dependent regulatory elements for intergenic miRNAs in human H4 cells, Biochemical and Biophysical Research Communications, 2012, Volume 420, Issue 2
[4] Park et al, Exome sequencing identifies KIAA1377 and C5orf42 as susceptibility genes for monomelic amyotrophy, Neuromuscular Disorders, 2012, Volume 22, Issue 5
[5] Park et al, Exome sequencing and subsequent association studies identify five amino acid-altering variants influencing human height, Hum Genet, 2011, 10.1007/s00439-011-1096-4
[6] Park et al, WeGAS: a web-based microbial genome annotation system, Biosci. Biotechnol. Biochem, 2009, 73(1), 213~216
[7] Park et al, Development of an analysis program of type I polyketide synthase gene clusters usin g homology search and profile hidden Markov model, Journal of Microbiology and Biotechnology, 20 09, 19(2):140-146
[8] Park et al, MapsiDB: an integrated web database for type I polyketide synthases, Bioprocess Bios yst Eng, 2009, 32:723~727

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science
and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

Abstract of proposed research for WGS pan-cancer analysis
Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mimicking cancer mechanisms by direct comparison with instable stem cell genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Hyung-Lae Kim, National Project for Personalized genomic medicine 21, Korea.
Sung-Soo Yoon, Seoul National University Hospital (SNUH), member of ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Youngil Koh, SNUH, Seoul, Korea **Kwang-Sung Ahn, PDgene, Inc**	**Jongsun Jung, DCC in PGM21, Korea** **Kinarm Ko, Ph.D in Konkuk University, Korea**

Background and preliminary data

Background

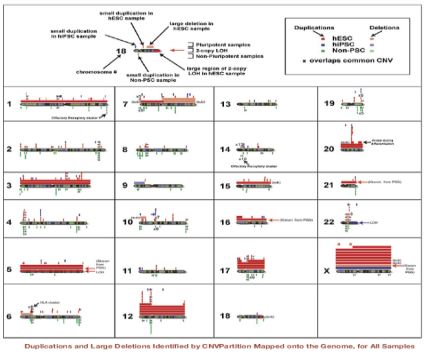
During the period when NGS technology was not well established, the PCR-based measurement of genomic instability by Storer et.al[1] was commonly used, but now, the advanced NGS technology enables for the direct comparison of instable genomes between instable cancer and stem cells by NGS sequencing data alone.

Stoler et. al. has measured genomic instability in invasive breast carcinomas and assessed the relationship of genomic instability to known tumor prognostic factors. DNAs from tumors and adjacent normal tissue of 18 breast cancer patients were subjected to inter-Simple Sequence Repeat (inter-SSR) PCR for quantitation of tumor genomic instability. Associations between genomic instability level and known breast cancer prognostic factors were evaluated using the Pearson Product Moment Correlation, the Kruskal-Wallis test of independent samples and the Mann-Whitney non-parametric test. Genomic instability was detected by inter-SSR PCR in over 90% of the breast tumors. The mean instability index was 3.08% (0-7.59%), approximately the same mean value observed in studies of colorectal and thyroid carcinomas. Stoler hypothesizes that the higher levels of genomic instability detected in necrotic tumors is a consequence of hypoxia-associated DNA damage.

Preliminary Study,

Full manipulation of Variants including SV, Indel, CNV, Variant Calling, viral integration mapping and Hypervariable region mapping using 2,000 WES and 10 WGS sequence data.

Full manipulation of exome, expression and methylation analysis of human mesenchymal stem cells, neural stem cells, ESCs, and iPSCs reprogrammed from somatic cells and their instability parameters.



Cell, 2010, Louise et. al,

Table 1. Summary of recurrent changes in human embryonic stem cells and human induced pluripotent stem cells following prolonged culture

Recurrent changes	hESCs	hiPSCs	References
Trisomy 12 or Amplification of 12p	+	+	[4–6, 33, 35–37]
Trisomy X	+	+	[4, 33]
Trisomy 17	+	–	[4, 5, 33]
Amplification of 17q	+	+	[6, 14]
Amplification of 20q11.21	+	+	[11, 12, 35, 37–39]
Isodicentric X	+	–	[13]
Deletion of 18q12.1	+	–	[11]
Amplification of 1p36.13	+	–	[15]
Amplification of 1p36.33	+	–	[15]
Amplification of 2p11.2	+	–	[15]
Amplification of 7q35	+	–	[15]
Amplification of 14q32.32	+	–	[15]
Deletion of 15q11.2	+	–	[15]
Amplification of 21q11.2	+	–	[15]
Amplification of 22q11.22	+	–	[15]
Deletion of 22q11.21	+	–	[15]
Trisomy 8	+	+	[33]
Trisomy 20q (isochromosome 20q)	+	+	[33]
Amplification of 1q31.3	–	+	[35]
Deletion of 17q21.1	–	+	[35]
Deletion of 8q24.3	–	+	[35]

Abbreviations: hESC, human embryonic stem cells; hiPSC, human induced pluripotent stem cells.

Stem cells, 2012, Kristen et. al,

Timelines & resources dedicated to project

a. HPC screening of genome instable factors using 2,000 WGSs – June to October, 2014
b. Generation of 50~100 WGS(s) of instable Stem Cells - June to October, 2014
c. HPC screening of genome instable factors using 2,000 WGSs – October to December, 2014
d. Functional/Physical Mapping of Instable Factors between Cancer and Stem Cell by March 2015

Motivation:

Similarity between stem cell instability and cancer genomic alteration can be easily observed in many labs [2 & 3]. For examples, a chromosome aberration is either an incorrect number of chromosomes (that can occur as a consequence of an error during cell division) or a structural abnormality in one or more chromosomes. There are many types of chromosome anomalies, which can be organized into two groups: numerical or structural. An abnormal number of chromosomes is called aneuploidy and occurs when either one or more chromosomes are missing or gained. A structural abnormality is defined when the normal chromosome structure is altered (e.g., deletion, duplication, translocation etc). Without the terms, stem cell and cancer, there is no difference between the two topics, meaning that they both are involved in chromosome aberration.

Therefore, one way to understand cancer mechanism could be a direct comparison between cancer caused by environmental interrogation and pluripotent stem cells [embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs)] which are naturally tumorigenic.

ICGC Pan-Cancer Project:

First of all, normalization of depth and cellularity for all cancer genomes can be performed and then, a full extraction of variants such as SNV, indel, CNV, and SV by the published various NGS tools using 2,000 pairs of cancer genomes in ICGC data centre. Since most of tools give differently formatted outputs, we have to re-extract the original depth and basic information from the selected candidate markers using in-house ADISCAN(the below) which can take a list of candidate markers and multiple BAM files as an input data for validation and rescaling of the candidate markers. And, then by the given significant cutoff , we could extract the valid markers for common features between cancer and pluripotent stem cells, proving and mimicking some of possible cancer mechanisms.

Bioinformatics Lab:

We have a full manipulation of BAM files for extracting all different types of variants using ADISCAN (allelic depth and imbalance scanning) that is based on SAMTOOLS Engine developed by Heng Li from the Sanger Institute. In addition, we have a full control of 5,000 whole-exome sequencing data from PGM21 groups in Korea, some of which are deposited to ICGC.

Stem Cell Research Group:

Current studies investigated genetic and epigenetic instability in human pluripotent stem cells using NGS tool [4]. Prof. Kinarm Ko's group is fully equipped with manipulation of stem cell lines, human mesenchymal stem cells, human neural stem cells, human ESCs, and human iPSCs reprogrammed from somatic cells described by the methods [5, 6, 7 & 8]. They are planning to sequence many genomes (over 100 genomes) from different types of stem cells including human iPSCs generated with/without viral integration and investigate quality and safety of them for basic research and stem cell therapy. And then, they are going to generate NGS database with regard to a genetic instability of various stem cells.

[1] Breast Cancer Res Treat. 2006 May;97(1):107-10. Genomic instability in invasive breast carcinoma measured by inter-Simple Sequence Repeat PCR.

[2]High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. Nature Biotechnology, 2010, 372,

[3]Dynamic Changes in the Copy Number of Pluripotency and Cell Proliferation Genes in Human ESCs and iPSCs during Reprogramming and Time in Culture. 2011, Cell 8, 106–118

[4] Genetic and epigenetic stability of human pluripotent stem cells. 2012, Nature review genetic 13, 732-744

[5] Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. 2006, Cell 126, 663-676

[6]Breast Cancer Res Treat. 2006 May;97(1):107-10. Genomic instability in invasive breast carcinoma measured by inter-Simple Sequence Repeat PCR.

[7]Genome-Wide Interrogation of Mammalian Stem Cell Fate Determinants by Nested Chromosome Deletions. PLoS Genet 6(12): e1001241.

[8]A retroviral strategy that efficiently creates chromosomal deletions in mammalian cells. Nat Methods. 2007. 4(3):263-8

Legacy plans

We expect that the computational steps and stem cell instability database used to produce publication-ready results on the pan-cancer data set will be embodied in executable code, web-accessible and sufficiently well documented to enable replication by third parties.

Curriculum Vitae

Hyung-Lae Kim, MD, Ph.D

School of Medicine
Ewha Wonans Uiversity
Mok-5-Dong, Yangcheon-Gu,
Seoul , Korea
hyung@ewha.ac.kr
+82-2-2650-5727

CAREER & EDUCATION

2012 ~, Executive Committee Member, International Scientific Steering Member, ICGC

2011 ~, Director General, The National Project for Personalized Genomic Medicine, Korea

2013 ~, President, Korean Society of Biochemistry and Molecular Biology

2011, President, Korea Genome Organization

2008 ~2010, Director General, Korea National Institute of Heath, Korea

2006 ~ 2008, Director, Center for Genome Science, Korea National Institute of Heath, Korea

1995 ~, Professor, Dept. Biochemistry, School of Medicine, Ewha Womans University, Korea

1989 ~ 2002, Visiting Fellow, Lab Mol Biology, NINDS, NIH, USA

1986, PhD in Biochemistry, Graduate School, Seoul National University, Korea

PUBLICATION (5 years)

[1] Kim et al. A genome-wide association study of a coronary artery disease risk variant
J Hum Genet. 58(3):120-6 (2013).

[2] Kim et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. (2012) Mar 27;14(2):R56

[3] Cho et al. Meta-analysis of genome-wide association studies identifies 8 new loci for type 2 diabetes in East AsiansNat Genet. 44(1): (2011) 67-72

[4] The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret et al.Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367) (2011):103-9.

[5] Kim et al. Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. Nat Genet. 43(10): (2011); 990-5.

[6] Fox et al. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study.Hum Mol Genet. 20(2011): 2273-84.

[7] Jung et al. Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 5(7) (2010): e11855

[8] HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 326 (2009):1541-5.

[9] Cho et al. A large scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics 41(2009):527-34.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012
[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7
[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7
[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea
1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea
1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.
1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.
1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7
[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.
[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.
[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

Curriculum Vitae

Youngil Koh, MD

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
Go01@chol.com
+82-1091175012

CAREER & EDUCATION

2013.5 - Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea
2010.3 - 2013.4 Public Service Doctor, Kkotdongnae, Gapyeong, Korea (Military service)
2010.3 - PhD. Candidate, Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Masters in Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
2005.2, MD, Seoul National University College of Medicine, Seoul, Korea

AWARDS

2012 Best Oral Presentation Award, Korean Association for Clinical Oncology Annual Meeting
2011 Best Oral Presentation Award, Korea Cancer Association Annaul Meeting
2008 Travel Award, American Society of Hematology Annual Meeting
2008 Best doctor for patients, Seoul National University Hospital
1998 Bronze medal, 39th International Mathematics Olympiad, Taiwan
1997 Silver medal, 38th International Mathematics Olympiad, Argentina

PUBLICATION (recent 2 years, 1st author only, including co-first author)

[1] Park S, Koh Y, Jung SH, Chung YJ. Application of array comparative genomic hybridization in chronic myeloid leukemia. Methods in molecular biology 2013;973:55-68.
[2] Kim I, Koh Y, Yoon SS, et al. Fludarabine, cytarabine, and attenuated-dose idarubicin (m-FLAI) combination therapy for elderly acute myeloid leukemia patients. American journal of hematology 2013;88:10-5.
[3] Koh Y, Lim HY, Ahn JH, et al. Phase II trial of everolimus for the treatment of nonclear-cell renal cell carcinoma. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2013;24:1026-31.
[4] Koh Y, Kim I, Shin DY, et al. Polymorphisms in genes that regulate cyclosporine metabolism affect cyclosporine blood levels and clinical outcomes in patients who receive allogeneic hematopoietic stem cell transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation 2012;18:37-43.
[5] Koh Y, Lee HE, Oh DY, et al. The lack of CD34 expression in gastrointestinal stromal tumors is related to cystic degeneration following imatinib use. Japanese journal of clinical oncology 2012;42:1020-7.

Curriculum Vitae

Kwang-Sung Ahn, Ph.D

PDxen, Inc
Functioal Genome Institute,
Junggook-dong, Gwangjin-gu
and Technology,
Seoul , Korea
Kwangsung.ahn@gmail.com
+82-10-7722-2460

PROFESSIONAL EXPERIENCE

1] Cancer Research Center, Seoul National University, Seoul, Korea. Lab manger (Research Professor) [2004 – present]
Identification of prognostic markers and Functional analysis of drug responsive genes in Multiple myeloma and Acute Myeloid Leukemia.
2] Genome Research Center, Samsung Biomedical Research Center, Samsung Seoul Hospital, Seoul, Korea. (Research Professor) – Form Oct. 2001 to Oct. 2005
Functional analysis of metastatic genes in metastatic animal model.
3] Center for Health Science, School of Dentistry, University of California Los Angeles, CA, USA

CAREER & EDUCATION

1996 – 1997: CENTER FOR HEALTH SCIENCE, SCHOOL OF DENTISTRY, UNIVERSITY OF CALIFORNIA LOS ANGELES, LOS ANGELES, CALIFORNIA
1990 – 1996: DEPARTMENT OF BIOLOGY, GRADUATE SCHOOL OF ART & SCIENCES, UNIVERSITY OF HOUSTON, HOUSTON, TEXAS
1987 – 1989: DEPARTMENT OF BIOLOGY, GRADUATE SCHOOL OF ART & SCIENCES, LONG ISLAND UNIVERSITY AT C.W. POST, NEW YORK. MS, Department of microbiology,
1977 -1985, DEPARTMENT OF BIOLOGY, SUNGKYUNKWAN UNIVERSITY- KYUNGGI, KOREA. BS, Biology,

PUBLICATION (5 years)

[1] Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2] Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3] Park et al, RNA interference-directed caveolin-1 knockdown sensitizes SN12CPM6 cells to doxorubicin-induced apoptosis and reduces lung metastasis. Tumour Biol. 2010 6:643-50.
[4] Park et al, Establishment of a new Glivec-resistant chronic myeloid leukemia cell line, SNUCML-02, using an in vivo model. Exp Hematol. 2010 38(9):773-81.
[5] Cha et al Slug suppression induces apoptosis via Puma transactivation in rheumatoid arthritis fibroblast-like synoviocytes treated with hydrogen peroxide. Exp Mol Med. 2010 30;42(6):428-36.
[6] Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science
and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

CURRICULUM VITAE

Kinarm Ko Ph.D.

Associate Professor
Center for Stem Cell Research, Institute of Biomedical Science and Technology, Konkuk University
Department of Stem Cell Biology, School of Medicine, Konkuk University
Tel: 02-2030-7888
E-mail: knko@kku.ac.kr

EDUCATION

Ph.D. 2003 Endocrinology and Reproductive Physiology, University of Wisconsin, Madison, WI, USA
M.S. 1998 Animal Science, University of Wisconsin, Madison, WI, USA
B.S. 1991 Animal Science, College of Agriculture and Life Sciences, Seoul National University, Seoul, Korea

EMPLOYMENT

- Research Associate: (November 2008 – Present) Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Muenster, Germany
- Postdoctoral Fellow: (October 2004 - October 2008): Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Muenster, Germany
- Visiting Fellow (September 2003- September 2004): Lab of Comparative Carcinogenesis, National Cancer Institute, Frederick, MD, USA
- Research Assistant (January 1998 - June 2003): Lab of Developmental Toxicology, School of Pharmacy, University of Wisconsin, Madison, WI, USA
- Research Assistant (January 1995 - January 1997): Aquaculture Program in Department of Food Sciences, University of Wisconsin, Madison, WI, USA

AWARDS

2010 International Society of Stem Cell Research (ISSCR) 7th Meeting (Poster Travel Awards)
2008 International Society of Stem Cell Research (ISSCR) 6th Meeting (Junior Investigator Awards)
2007 Stem Cell Network North Rhine Westphalia (NRW) 4th Congress(Award of the best poster presentation)
2003 Society of Toxicology (SOT) 42th Meeting (Carl Smith Awards for Mechanisms Specialty Section)
2002 Vilas Travel Grant Award, University of Wisconsin-Madison
2001 Graduate Student Travel Grant Award, Graduate School, University of Wisconsin-Madison
1990 Student Education Travel Grant Award, Seoul National University, Seoul, Korea.

Publications (2010-2013)

[1]Lee JH, Park DY, Lee KJ, Kim YK, So YK, Ryu JS, Oh SH, Han YS, **Ko K**, Choo YK, Park SJ, Brodzik R, Lee KK, Oh DB, Hwang KA, Koprowski H, Lee YS, Ko K. Intracellular reprogramming of expression, glycosylation, and function of a plant-derived antiviral therapeutic monoclonal antibody. PLoS One. 2013; 8(8)

[2]Ko K, Wu G, Araúzo-Bravo MJ, Kim J, Francine J, Greber B, Mühlisch J, Joo JY, Sabour D, Frühwald MC, Tapia N, Schöler HR.Autologous pluripotent stem cells generated from adult mouse testicular biopsy. Stem Cell Rev. 2012 Jun; 8(2): 435-44

[3]Ko K, Wu G, Araúzo-Bravo MJ, Kim J, Francine J, Greber B, Mühlisch J, Joo JY, Sabour D, Frühwald MC, Tapia N, Schöler HR.Autologous Pluripotent Stem Cells Generated from Adult Mouse Testicular Biopsy. *Stem Cell Rev*. 2011 Aug 20.

[4]Ko K, Reinhardt P, Tapia N, Schneider RK, Araúzo-Bravo MJ, Han DW, Greber B, Kim J, Kliesch S, Zenke M, Schöler HR. (2011) Evaluating the potential of putative pluripotent cells derived from human testis. *Stem Cells*. Aug; 29(8): 1304-9.

[5]Ko K, Bravo MJ, Tapia N, Kim J, Lin Q, Berneman C, Han DW, Gentile L, Reinhardt P, Greber B, Schneider RK, Kliesch S, Zenke M, Schˆler HR. (2010) Human Adult Germline Stem Cells in Question. *Nature*. 465(7301)

[6]Ko K, Bravo MJ, Kim J, Stehling M, Schöler HR. (2010) Conversion of Adult Mouse Unipotent Germline Stem Cells into Pluripotent Stem Cells. *Nature Protocol*. 5(5): 921-8.

[7]Ko K, Huebner K, Mueller-Keuker J, Schˆler HR. (2010). *In vitro* derivation of germ cells from embryonic stem cells. *Frontiers in Bioscience* (15): 46-56.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Deciphering co-occurrence/exclusivity patterns between cancer elements from pan-cancer genome data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Keunchil Park, Samsung Medical Center
Hyunglae Kim, PGM21, Director

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Youngwook Kim, Samsung Medical Center	Sinho Jung, Samsung Medical Center

Background and preliminary data

Advances in high-throughput technologies have enabled a comprehensive, genomic characterization of hundreds of samples in a given cancer type. These analyses have revealed major cancer genome aberrations with functional significance, as well as a long tail of candidate genetic events that might contribute to the pathogenesis of human cancer (Nature 499:214-8, 2013; Nature 500:415-21, 2013; Nature 502:333-9, 2013). However, many challenges still remain in gaining insights into the nature of these minor candidate oncogenes/tumor-suppressors.

Compared to passenger mutations, important somatic-genetic events tend to leave their functional or physical signatures in various forms in the genome, including expression and mutation patterns, as they interact with other genetic or clinical factors in mechanistic or functional ways. The interaction between these cancer factors is often manifested in the form of co-occurrence of seemingly independent cancer elements in cancer samples, independent of their origin of tissue, or of mutual exclusivity between interconnected components of a functional signaling module (e.g. *EGFR* and *KRAS* do not occur simultaneously in lung adenocarcinoma, and smoking leaves their characteristic mutational signature in the smokers' genome). Thus, assessing the co-occurrence and mutual exclusivity of cancer elements in a statistically coherent setting from a vast collection of human cancer samples will help to delineate functionalities of some still uncharacterized cancer-associated genes.

We have previously explored the integrated analysis of interaction between cancer factors in a given type of tumor sample (JCO *in press*, 2014) and successfully observed previously reported interactions between cancer elements. It will be of great scientific and clinical importance to expand this type of analysis to larger data sets to further refine the stratification of many uncharacterized cancer-associated genes.

Timelines & resources dedicated to project

~ 2014. Mar. Integration of ICGC pan-cancer data set phase I
~ 2014. Jun. Integration of ICGC pan-cancer data set phase II
Reduction of sequencing data for statistical testing
~ 2014. Dec. Determination of statistical interaction between cancer elements
Validation of key findings

WGS, transcriptome, DNA methylome of ICGC pan cancer data

Research proposal

Specific Aim 1) **Integrated representation of pan-cancer genome data**

About 2000 whole genome data, in conjunction with transcriptome, exome and methylation data, will be integrated. The extremely large volume of pan-cancer ICGC genome data requires us to reduce the data set in statistically assessable ways. Catalogues of genomic mutations, including somatic single-nucleotide variations, small indels, and structural variations, along with simple mutational signatures, will be extracted from whole genome/exome data. Statistically significant recurrent cancer variations, including mutation and structural variations, will be digitized per gene unit, reflecting the presence/absence of mutational events in the corresponding gene. Other data set, including clinical annotation or mutation signature, will be included as independent factors in the integrated representation of genomic data.

Specific Aim 2) **Determination of statistical interaction between cancer elements**

Since mechanistically or functionally important cancer elements interact with other cancer elements in promoting tumorigenesis, whether extracted cancer elements display patterns of co-occurrence or mutual exclusivity with others will be tested using advanced statistical models. Given the vast amount of data set expected, we anticipate that problems associated with high false discovery rate may emerge. Knowledge obtained from previously characterized exclusivity within intracluster functional signaling module and co-occurring mutation patterns in each tumor type will be applied to the statistical model to control false-discovery rate.

Specific Aim 3) **Prioritization and functional validation of statistical interaction between cancer elements**

Identification obtained by statistical inference will be prioritized and combined with knowledge of biological mechanisms. Key findings obtained from the above described statistical inference will be functionally validated using cell-based biological model systems.

Legacy plans

For mutation calling, we will mainly use standard open-source pipelines available on line. In case that we newly develop algorithms/softwares for deciphering interactions between cancer elements, we will produce and disclose publication-ready executable source code in accessible public domains with proper user-guide documentation.

Curriculum Vitae

Keunchil Park, M.D., Ph.D.

Keunchil Park is Professor of the Division of Hematology-Oncology, Sungkyunkwan University School of Medicine in Seoul, Korea. Prof. Park is Director of the Medical Nano Element Development Center, and is the Principal Investigator of the 'Identification of Novel Therapeutic Targets in Lung Cancer with Unmet Need' of the National Project for Personalized Genomic Medicine(PGM21), both of which are funded by the Ministry of Health and Welfare, Korea.

Professor Park has served many domestic academic societies, e.g., Chair of the Scientific Committee of the Korean Cancer Association, Chair of the Lung Cancer Committee of the Korean Cancer Study Group(KCSG). Prof. Park also served as Chairman of the Board of Directors, Korean Association for Clinical Oncology (KACO) since June 2010 until May 2012.

Prof. Park has been also very actively involved in and served many international activities, such as the Scientific Secretary of the 12th WCLC (Sept, 2007), and the Chairman of the 4th Asia Pacific Lung Cancer Conference (Dec, 2010). He was elected as the Board of Directors of the IASLC and is serving as Associate Editor for the Journal of Thoracic Oncology (JTO) and on editorial board of the Asia-Pacific Journal of Clinical Oncology.

Prof. Park's main interests include the translational and early clinical researches for the treatments of upper aero-digestive tract cancers, especially lung cancer. Recently Dr. Park is leading several early clinical trials of the targeted agents as well as many pre-clinical development programs internationally. Prof. Park has several book chapters and authored more than 200 peer-reviewed publications in national and international journals.

Curriculum Vitae

Hyung-Lae Kim, MD, Ph.D

School of Medicine
Ewha Wonans Uiversity
Mok-5-Dong, Yangcheon-Gu,
Seoul , Korea
hyung@ewha.ac.kr
+82-2-2650-5727

CAREER & EDUCATION

2012 ~, Executive Committee Member, International Scientific Steering Member, ICGC

2011 ~, Director General, The National Project for Personalized Genomic Medicine, Korea

2013 ~, President, Korean Society of Biochemistry and Molecular Biology

2011, President, Korea Genome Organization

2008 ~2010, Director General, Korea National Institute of Heath, Korea

2006 ~ 2008, Director, Center for Genome Science, Korea National Institute of Heath, Korea

1995 ~, Professor, Dept. Biochemistry, School of Medicine, Ewha Womans University, Korea

1989 ~ 2002, Visiting Fellow, Lab Mol Biology, NINDS, NIH, USA

1986, PhD in Biochemistry, Graduate School, Seoul National University, Korea

PUBLICATION (5 years)

[1] Kim et al. A genome-wide association study of a coronary artery disease risk variant J Hum Genet. 58(3):120-6 (2013).

[2] Kim et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. (2012) Mar 27;14(2):R56

[3] Cho et al. Meta-analysis of genome-wide association studies identifies 8 new loci for type 2 diabetes in East AsiansNat Genet. 44(1): (2011) 67-72

[4] The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret et al.Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367) (2011):103-9.

[5] Kim et al. Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. Nat Genet. 43(10): (2011); 990-5.

[6] Fox et al. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study.Hum Mol Genet. 20(2011): 2273-84.

[7] Jung et al. Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 5(7) (2010): e11855

[8] HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 326 (2009):1541-5.

[9] Cho et al. A large scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics 41(2009):527-34.

Curriculum Vitae

Youngwook Kim

Samsung Biomedical Research Institute

Senior Researcher

Rm.188 B4 Cancer Center

50 Irwon-dong Gangnam-gu

Seoul, Korea zip: 135-710

Office: 82-2-2148-7349

Cell: 82-10-2300-9856

Publications

1. **Kim Y**, Hammerman P, Kim JG, Yoon J, Lee Y, Sun J, Wilkerson M, Pedamallu C, Cibulskis K, Yoo Y, Lawrence M, Stojanov P, Carter S, Hayes N, Getz G, Meyerson M, Park K Integrative and comparative genomic analysis of lung squamous cell carcinomas in East-Asians Journal of Clinical Oncology *in press (2014)*

2. **Kim Y**, Kim J, Lee J, Bae K, Min J, Park T, Lee J, Nam Y, Park K *Tumor-Targeted Delivery of Paclitaxel using Solid Lipid Nanoparticles* Nature Communications *in review*

3. **Kim Y**, Ko J, Cui Z, Abolhoda A, Ahn JS, Ou SH, Ahn MJ, Park K *The EGFR T790M mutation in acquired resistance to an irreversible second-generation EGFR inhibitor* Mol Cancer Ther. (2012) Mar;11(3):784-91.

4. Lee S*, **Kim Y***, Sun JM, Choi YL, Kim JG, Shim YM, Park YH, Ahn JS, Park K, Han JH, Ahn MJ *Molecular profiles of EGFR, K-ras, c-met, and FGFR in pulmonary pleomorphic carcinoma, a rare lung malignancy*. J Cancer Res Clin Oncol. (2011) Aug;137(8):1203-11. * equally contributing authors

5. Oh YH*, **Kim Y***, Kim YP, Seo SW, Mitsudomi T, Ahn MJ, Park K, Kim HS *Rapid detection of the epidermal growth factor receptor mutation in non-small-cell lung cancer for analysis of acquired resistance using molecular beacons.* J Mol Diagn. (2010) Sep;12(5):644-52. *equally contributing authors

Curriculum Vitae

Name: Sin-Ho Jung, Ph.D.

Education:	Institution	Date [year]	Degree
College	Seoul National University	1982	B.A.
Graduate or professional school	Seoul National University	1984	M.S.
	University of Wisconsin-Madison	1992	Ph.D.

Scholarly societies:

American Statistical Society	Elected fellow
Institute of Mathematical Statistics	Member
Biometric Society (ENAR)	Member
Korean Statistical Society	Member

Professional training and academic career: [chronologically, beginning with first postgraduate position]

1. Academic Appointments

Institution	Position/Title	Dates
Mayo Medical/Graduate School	Assistant Professor	1994-1995
Hallym University	Assistant Professor	1995-3/99
University of Wisconsin-Madison	Visiting Assistant Professor	1/98-7/98
Indiana University (IU) School of Medicine	Associate Professor	8/98-9/01 (with tenure)
IU School of Public Health	Adjunct Faculty	8/99-9/00
Duke University	Associate Professor	10/1/01-5/31/06
	Professor	6/1/06-present
SAIHST, Sungkyunkwan U	Professor	8/10/2013-present

2. Administrative Positions

Institution	Position/Title	Dates
IU Cancer Center	Biostatistics Core Director	8/98-9/01
American College of Surgeons Oncology Group	Acting Group Statistician	4/02-4/03
Cancer & Leukemia Group B	Director, Biostatistics Unit	1/08-present
Duke Cancer Institute	Interim Director, Biostatistics	4/11-present
Center for Biostatistics & Clinical Epidemiology		
Samsung Medical Center	Director	12/1/2013-present

3. Collaboration/Committee

Institution	Position/Title	Dates
Mayo/North Central Cancer Treatment Group	Faculty Statistician	1992-1995
Hoosier Oncology Group	Faculty Statistician	1999-2001
IU Cancer Center, Scientific Review Committee		8/98-9/01
Mary Margaret Walther Program, Internal Advisory Committee		5/99-9/01
IU General Clinical Research Center, Advisory Committee		5/01-9/01
IU School of Medicine, IRB-02 and IRB-04		7/01-9/01
American College of Surgeons Oncology Group	Faculty Statistician	10/01-4/04
Cancer and Leukemia Group B	Faculty Statistician	5/04-present
Cancer Protocol Committee, Duke Comprehensive Cancer Center		

4. Professional Activities

NCI Lymphoma Steering Committee, Member, 3/2010-present.
NCI DLBCL Working Group, Member, 12/2010-present.

Peer reviewed publications: 148 papers published or accepted for publication

Books:

1. Ko ER, Park BJ, Jung SH. Statistical Methods for Clinical Trials, 2nd ed, 1997 (in Korean).
2. Jung SH. Randomized Phase II Cancer Clinical Trials. Chapman & Hall/CRC, 2013.

Areas of research interest:

Statistical methods for cancer clinical trials, Bioinformatics, Survival analysis, Longitudinal data analysis, Design and analysis of phase II clinical trials, Clustered data analysis.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The HER2 pathway and Pan-Cancer Analysis

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Keunchil Park Samsung Medical Center, Sungkyunkwan University
Sungsoo Yoon Seoul National University Hospital

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Seung Tae Kim, Youngwook Kim Samsung Medical Center, Sungkyunkwan University	Jeeyun Lee, Samsung Medical Center Jongsun Jung, SyntecaBio

Background and preliminary data

Increased numbers of tumor sample data sets enhance the ability to detect and analyze molecular aberrations in cancers. These steps are able to categorize connecting subsets of tumors from different tissues in terms of molecular signature.

ERBB2-HER2 has an established role as a prognostic and predictive factor in breast cancer. Furthermore, ERBB2-HER2 is mutated and/or amplified in subsets of glioblastoma, serous endometrial, gastric, bladder, colon and lung cancer. Although HER2 deregulation, including overexpression, amplification, and mutation has been described in these cancer, its role as therapy biomarker remains undefined. Tumors also share complex interactions of various molecular signatures. For examples, co-activation of HER2/MET or HER2/PIK3CA pathway, HER2 amplification/mutation, and MET amplification/activated HER2 pathway are observed in same tumor tissue. Maybe, these molecular aberrations will cause the different response to same molecular targeted agent.

Our group has studied the jungle of HER2 pathway in gastric cancer. Concomitant overexpression of EGFR and MET in $HER2^{+}$ and $HER2^{-}$ GCs (1589 patients) were evaluated. Also, using patients-derived tumor cell model of $HER2^{+}$ and MET^{+} GC , we demonstrated that the combination of HER2 inhibitor (lapatinib) and MET inhibitor demonstrated more profound inhibition of ERK/AKT pathway than lapatinib alone (unpublished. Submitted). Recently, using whole-genome sequencing (WGS) results of 50 GC sample, we investigated genetic aberrations including the status of HER2. We found 6 patients with HER2 amplification and 2 with HER2 mutation. We also observed FLG-, TYROS-, ALK-, and IGF1R-gene mutations (unpublished. Submitted).

From these findings, we need to evaluate whether aberrations of HER2 pathway are drivers or passengers in various tumor types. The solution to these questions will ultimately inform clinical decision making.

Timelines & resources dedicated to project

Timeline:

Period 1. December 2013 – January 2014 :
Arrangement of project-team for pan-cancer analysis
Period 2. January 2014 – May 2014 :
Analysis of pan-cancer data from ICGC
Period 3. May 2014 – August 2014 :
Summary for Co-occurrence, mutual exclusivity and integrating event such as mutation, copy number changes, and epigenetic changes
Period 4. August 2014 – October 2014 :
Make Map for the jungle of HER2 pathway across tumor types
Period 5. November 2014 – December 2014 :
Perform validation if novel findings are founded
Period 6. ~ 20th March 2015 ; Manuscript preparation and submission

Research proposal

* From our preliminary researches and current knowledge, we find followings ;
1. Although HER2 deregulation, including overexpression, amplification, and mutation has been described in various cancers, its role as therapy biomarker remains undefined.
2. Tumors also share complex interactions of various molecular signatures.
For examples, co-activation of HER2/MET or HER2/PIK3CA pathway, HER2 amplification/mutation, and MET amplification/activated HER2 pathway are observed in tumor tissues.
3. These molecular aberrations will cause the different response to same molecular targeted agent.
4. For drug targetable HER2, it is important to know whether aberrations of HER2 pathway are drivers or passengers in various cancer types. That will give us accurate information for overcoming drug-resistance and increasing treatment-efficacy.
5. We must get the map for the jungle of HER2 pathway. The map will be made by comprehensive analysis of HER2 pathway through pan-cancer analysis.

* For research, we propose followings ;
1. We will analyze and interpret the WGS data in various 13 tumor-types (pancreatic cancer, gastric cancer, liver cancer, breast cancer, bone cancer, oesophageal cancer, glioma, endometrial cancer, head & neck cancer, non-small cell lung cancer, colorectal cancer, bladder cancer and cervical cancer). Especially, we focus the HER2 pathway including mutation, amplification, copy number changes, epigenetic changes, co-occurrence of genetic aberrations and mutual exclusivity of gene.
2. In order to interpret connections of these genetic aberrations, we will develop the algorithms for mapping based on the bio-informatics.
3. We will make a genomic map for HER2 pathway across tumor types by step1 and step2.
4. We try to interpret and analyze the outcome of clinical trials with HER2 directed therapies based on the developed genomic map for HER2 pathway.

* Validations
Perform validation if novel findings are founded
1. If we get the genomic map for the jungle of HER2 pathway, we will apply it to patients who already received HER2-directed therapies
2. We will prospectively confirm roles of novel biomarkers which were generated from this project.

* Meanings of this proposal
1. To inform clinical decision making ;
A> Discovery of drugable targets
B> Overcome the acquired-resistance of targeted agents
C> Develop companion diagnostic tools
D> Design novel adaptive, biomarker based clinical trials irrespective of tumor types
2. The development of clinical strategies for connecting subsets of tumors from different tissues in term of molecular or genetic signatures.

Legacy plans

For mutation calling, we will mainly use standard open-source pipelines available on line. In case that we newly develop algorithms/softwares for deciphering interactions between cancer elements, we will produce and disclose publication-ready executable source code in accessible public domains with proper user-guide documentation.

Curriculum Vitae

Keunchil Park, M.D., Ph.D.

Keunchil Park is Professor of the Division of Hematology-Oncology, Sungkyunkwan University School of Medicine in Seoul, Korea. Prof. Park is Director of the Medical Nano Element Development Center, and is the Principal Investigator of the 'Identification of Novel Therapeutic Targets in Lung Cancer with Unmet Need' of the National Project for Personalized Genomic Medicine(PGM21), both of which are funded by the Ministry of Health and Welfare, Korea.

Professor Park has served many domestic academic societies, e.g., Chair of the Scientific Committee of the Korean Cancer Association, Chair of the Lung Cancer Committee of the Korean Cancer Study Group(KCSG). Prof. Park also served as Chairman of the Board of Directors, Korean Association for Clinical Oncology (KACO) since June 2010 until May 2012.

Prof. Park has been also very actively involved in and served many international activities, such as the Scientific Secretary of the 12th WCLC (Sept, 2007), and the Chairman of the 4th Asia Pacific Lung Cancer Conference (Dec, 2010). He was elected as the Board of Directors of the IASLC and is serving as Associate Editor for the Journal of Thoracic Oncology (JTO) and on editorial board of the Asia-Pacific Journal of Clinical Oncology.

Prof. Park's main interests include the translational and early clinical researches for the treatments of upper aero-digestive tract cancers, especially lung cancer. Recently Dr. Park is leading several early clinical trials of the targeted agents as well as many pre-clinical development programs internationally. Prof. Park has several book chapters and authored more than 200 peer-reviewed publications in national and international journals.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012

[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7

[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7

[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea

1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea

1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.

1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.

1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea

1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press

[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.

[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7

[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.

[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.

[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

CURRICULUM VITAE

Full Name:	Kim	Seung-Tae	
	Last Name	First Name	Middle Initial

Study Role

☒ **Investigator (Principal Investigator or any study site staff designated as Sub investigator)**

Other **may be selected if needed per local country requirements** ☐ **Other, specify:** _____

Professional Mailing Address: (Include institution name.)

Samsung Medical Center,
Sungkyunkwan Univ School of Medicine,
Div. of Hematology-Oncology, Dep. of Medicine,
Seoul, Korea

Telephone: 82-2-3410-0297, 010-8782-2153

Academic Qualifications (most current date first)

Degree/Certification	Date (YYYY)	Institution, Country
Ph.D.	2011	Samsung Medical Center, Sungkyunkwan Univ School of Medicine

Current and Previous 4 Relevant Positions Including Academic Appointments (most current date first):

Start and End Dates	Title	Institution or Company, State/Province/Country
2013.11~Current	Professor (assistant)	Samsung Medical Center, Sungkyunkwan Univ School of Medicine, Div. of Hematology-Oncology, Dep. of Medicine, Seoul, Korea
2010~2013	Professor (assistant)	Korea University Anam Hospital, Korea Div. of Hematology-Oncology, Dep. of Medicine,
2009~2010	Fellow	Samsung Medical Center, Sungkyunkwan Univ School of Medicine, Div. of Hematology-Oncology, Dep. of Medicine, Seoul, Korea
2005~2009	Resident	Samsung Medical Center, Sungkyunkwan Univ School of Medicine
2004~2005	Intern	Samsung Medical Center, Sungkyunkwan Univ School of Medicine

Medical License/ID Number:	Medical Doctor/73470	Licensed in State/Province/Country:	KOREA, REPUBLIC OF

Signature: 김승태 **Signature Date: 2013.12.30**

(Signature required for ALL Investigators)

I will update and resubmit my one page CV if there are changes and particularly if there is any change in status which would adversely affect the assessment of my suitability to conduct/participate in clinical studies.

NOTE: CV MUST BE LIMITED TO ONE-PAGE FOR INCLUSION IN THE ICH-E3 COMPLIANT CLINICAL STUDY REPORT. PLEASE NO ATTACHMENTS, AND NO TEXT ON THE REVERSE SIDE.

Curriculum Vitae

Youngwook Kim

Samsung Biomedical Research Institute

Senior Researcher

Rm.188 B4 Cancer Center

50 Irwon-dong Gangnam-gu

Seoul, Korea zip: 135-710

Office: 82-2-2148-7349

Cell: 82-10-2300-9856

Publications

1. **Kim Y**, Hammerman P, Kim JG, Yoon J, Lee Y, Sun J, Wilkerson M, Pedamallu C, Cibulskis K, Yoo Y, Lawrence M, Stojanov P, Carter S, Hayes N, Getz G, Meyerson M, Park K Integrative and comparative genomic analysis of lung squamous cell carcinomas in East-Asians Journal of Clinical Oncology *in press (2014)*

2. **Kim Y**, Kim J, Lee J, Bae K, Min J, Park T, Lee J, Nam Y, Park K *Tumor-Targeted Delivery of Paclitaxel using Solid Lipid Nanoparticles* Nature Communications *in review*

3. **Kim Y**, Ko J, Cui Z, Abolhoda A, Ahn JS, Ou SH, Ahn MJ, Park K *The EGFR T790M mutation in acquired resistance to an irreversible second-generation EGFR inhibitor* Mol Cancer Ther. (2012) Mar;11(3):784-91.

4. Lee S*, **Kim Y***, Sun JM, Choi YL, Kim JG, Shim YM, Park YH, Ahn JS, Park K, Han JH, Ahn MJ *Molecular profiles of EGFR, K-ras, c-met, and FGFR in pulmonary pleomorphic carcinoma, a rare lung malignancy*. J Cancer Res Clin Oncol. (2011) Aug;137(8):1203-11. * equally contributing authors

5. Oh YH*, **Kim Y***, Kim YP, Seo SW, Mitsudomi T, Ahn MJ, Park K, Kim HS *Rapid detection of the epidermal growth factor receptor mutation in non-small-cell lung cancer for analysis of acquired resistance using molecular beacons.* J Mol Diagn. (2010) Sep;12(5):644-52. *equally contributing authors

CURRICULUM VITAE

Name: Jeeyun Lee, M.D.
50 Irwondong Gangnamgu Seoul 135-710 Korea

CITI completed 2010/June

Education

2002.9 – 2005.2 Ph.D. (internal medicine), Graduate School, Sungkyunkwan University School of Medicine, Seoul, Korea

1999.8 – 2002.6 M.S. (internal medicine), Graduate School, Sungkyunkwan University School of Medicine, Seoul, Korea

1995.2 – 1999.2 M.D., School of Medicine, Ewha University, Seoul, Korea

Experience and Positions Held:

2010.2 – present ***Assistant Professor***, Division of Hem/Oncology, Samsung Medical Center
2006.3 – 2010.2 ***Clinical Assistant Professor***, Division of Hem/Oncology, Samsung Medical Center
2006.7 – 2008. 1 ***Visiting Scientist***, Sidney Kimmel Cancer Center, San Diego, CA, USA
2005.3 – 2006. 2 ***Senior Research Fellow***, Clinical Trial Center, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
2005.7 completed AACR Workshop in Cancer Research: Molecular Biology in Clinical Oncology
2004.8 completed Australia and Asia Pacific Clinical Oncology Research Development Workshop: A Workshop in Effective Clinical Trials Design
2004.3 – 2006.2 ***Clinical Fellowship in Hematology-Oncology,*** Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
2000.3 – 2004.2 ***Residency in Internal Medicine***, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
1999.3 – 2000.2 ***Internship***, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

License/Board Certifications
Sub-specialty in Hematology-Oncology (# 67819), Korea

Awards

Merit Award, GI ASCO Symposium, 2009

Young Investigator's Award, Asian Pacific Cancer Conference, Fall of 2005

Young Investigator's Award, ESMO 2004 for a poster presentation entitled "Postoperative adjuvant chemoradiotherapy in resected gastric cancer patients"

Best Fellow Research Award (SMC), 2005

Best Resident Award (SMC), 2002, 2003

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science
and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

Defining genomic alterations underlying "immunologic tumor" using whole genome sequencing data of various tumor types

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Sung-Soo Yoon, Seoul National University Hospital (SNUH), member of ICGC
Hyung-Lae Kim, Ewha Womans University / ICGC member

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Youngil Koh, SNUH, Seoul, Korea Kwang-Sung Ahn, PDxen, Seoul, Korea	Jong-Sun Jung, Syntekabio, Seoul, Korea Ji-Wan Park, Hallym University, Chuncheon, Korea

Background and preliminary data

Conventional immunotherapy of cancer includes cytokine treatments (especially in renal cell carcinoma (RCC), and melanoma), and cell therapies (allogeneic hematopoietic stem cell transplantation (HSCT) and adoptive allogeneic cell transfer). However, not much is known about the genetics of these "immunologically treated tumors" yet. Considering that cancer is a genetic disease, we think genomics technology would also reveal interesting findings in the immunologic treatment of cancer. We especially expect to reveal the followings.

1) We expect to understand the genetic pathways or germline characteristics that are responsible to make cancer "immunologic".
2) We expect to select cancer subgroups that respond to cytokine treatment. This is of utmost clinical importance because cytokine treatment is capable of drawing durable remission in RCC and melanoma. We would find out the genomic pathway that is responsible for the response to cytokine.

We have sufficient clinical experience and data regarding the immunologic treatment of cancer. For cytokine treatment, we have an interest in the analysis of RCC. As for allogeneic HSCT treatment, we investigated the role of allogeneic HSCT in chemo-refractory acute leukemias. Overall survival (OS) of acute myeloid leukemia (AML) and acute biphenotypic leukemia (ABL) were better than OS of acute lymphoblastic leukemia (ALL). Considering that the clinical result of ALL and AML/ABL in this setting is different, we think immunologic response of tumor is defined by somatic genetic alterations. However, we do not exactly know the genetic alteration that is contributing to this phenomenon. Second, because only 14% of AML/ABL patients were salvaged by allogeneic HSCT, we are trying to find predictive markers for response in this setting. This may be related to either HLA haplotype of stem cell donor/recipient, or related to the somatic genetic variation that a cancer harbors. From our clinical experience, we think genetics underlying tumor response to immunotherapy is not simple and composed of at least two factors. That is, 1) HLA of patients and donors (in case of cell therapy) and 2) somatic mutation of cancer.

1) Germline DNA of immunologic tumors (HLA)

Immune clearance of tumor starts from the antigen presentation by antigen presenting cells (APC). Hence, it is no doubt that HLA may be important in tumor immunotherapy. Hence, we conjecture that there could be a HLA cluster that is responsible for either immunologic cancer development or to response of tumor to immunotherapy. We have a technical base (**https://www.syntekabio.com**) that enables easy HLA genotype analysis using NGS data. As a preliminary study, we utilized a manipulation tool package, HMAP, for IMGT/HLA database to analyze a set of genotypes from diploid whole exome sequencing (WES) data of 84 AML and 180 NSCLC Korean patients. It was successful and we are going to publish the data regarding this study soon. When NGS data of various cancers are analyzed for HLA clustering, we also expect to generate similarity score among various cancer subtypes.

2) Somatic alteration of cancer DNA

Somatic alteration of cancer DNA might be a very important factor for tumor response to immunotherapy. And among these, genetic changes involving genes that are known to have relationship with immune system may be of utmost importance. Public database such as NCBI provide genes related to immune reaction. We are going to investigate the somatic genetic changes in these genes. Because little is known about tumor immunology in genetic aspect, cross-comparison of these somatic genetic changes across various tumor types is mandatory. Comparative analysis should be performed between immunologic tumors and non-immunologic tumors to reveal causative genetic changes.

Timelines & resources dedicated to project

Timeline:

Dec 2013 – Jan 2014: Arrangement of task-force team for this project.
Jan 2014 – Apr 2014:
1) Compile genes that are important in immune response
2) Perform HLA clustering using open sources and our WGS data set.
Jun 2014 – July 2014: HLA clustering analysis of pan-cancer germline data from ICGC
Jun 2014 – Sep 2014: Variant call of somatic changes of samples
Sep 2014 – Oct 2014: Make cross-tumor comparison and construct immunology genomic map across tumors
Sep 2014 – Jan 2015: Perform functional validation if novel findings are found [Verification of comprehensive network of cytokine and their receptor between immunologic tumor and non-immunologic tumor]
Jan 2015 - Feb 2015: Manuscript preparation
Mar 2015: Manuscript submission

Resources:

Reference for HLA clustering analysis: WGS data of 1000 genome project
Immunologic cancer: WGS, WES and WTS data of AML, kidney cancer, melanoma, and lymphoma
Non-Immunologic cancer: WGS, WES and WTS data of , non-small cell lung cancer and gastric cancer

Research proposal

We suggest looking at the two factors using NGS data to understand cancer immunobiology.
1) HLA genotype of cancer patients using and,
2) Somatic genetic changes in genes involved in immunology
Cross-tumor comparison is mandatory to effectively reveal causative genetic changes. And to perform this research using WGS data of ICGC, we propose the followings.

1. First of all, our group will compile genes that are known to be important in immune response. These genes may include cytokine genes, genes involved in antigen-presentation (such as TLR, and Myd88).
2. We will analyze the WGS, WES and WTS data of AML, kidney cancer, melanoma, lymphoma, non-small cell lung cancer, and gastric cancer. In order to reveal genetic alterations that have close relationship with immune response, we selected the tumor types according to the followings criteria.
   1) Included cancers that are known to respond to cytokine therapy. (RCC and melanoma)
   2) Included cancers that can be benefited from allogeneic HSCT. (AML and lymphoma)
   3) Included cancers that are known to be resistant to immunotherapy. (NSCLC and gastric cancer)
   4) Excluded cancers with squamous histology because they have many immunogens related to DNA damage caused by environmental pathogens, and those immunogens are assumed to alter immune response (such as in case of PD-1 inhibitor treatment).
3. We will look at the germline data in WGS of the cancer samples listed above to perform HLA clustering.
4. We will perform somatic variant calling using WGS data of these samples.
5. Finally, we will make a genomic map of both germline and somatic DNA across tumor types using data generated by step 3 and step 4.
6. In the analysis, survival data will be used, because immunologic tumors responsive to cytokine treatment or allogeneic HSCT have long term remission. These patients with excellent prognosis will be additionally analyzed in the aspect of HLA clustering and somatic genetic changes in immunologically relevant genes.
7. Functional validation of novel findings found by step 3-5 will be followed.

Legacy plans
For the calling of somatic mutation, HLA-clustering and gene-gene interaction analysis, we will mainly use utilities in open-source. If we design and develop calling algorithm for somatic mutations detection and/or HLA-clustering analysis, and/or gene-gene interaction analysis during this pan-cancer project, we will disclose and produce publication-ready source in ETRI/Syntekabio located in Seoul Korea. A documented virtual machine will be opened in a cloud computing system and will be embodied in executable code.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012
[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7
[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7
[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea
1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea
1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.
1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.
1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7
[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.
[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.
[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

Curriculum Vitae

Hyung-Lae Kim, MD, Ph.D

School of Medicine
Ewha Wonans Uiversity
Mok-5-Dong, Yangcheon-Gu,
Seoul , Korea
hyung@ewha.ac.kr
+82-2-2650-5727

CAREER & EDUCATION

2012 ~, Executive Committee Member, International Scientific Steering Member, ICGC

2011 ~, Director General, The National Project for Personalized Genomic Medicine, Korea

2013 ~, President, Korean Society of Biochemistry and Molecular Biology

2011, President, Korea Genome Organization

2008 ~2010, Director General, Korea National Institute of Heath, Korea

2006 ~ 2008, Director, Center for Genome Science, Korea National Institute of Heath, Korea

1995 ~, Professor, Dept. Biochemistry, School of Medicine, Ewha Womans University, Korea

1989 ~ 2002, Visiting Fellow, Lab Mol Biology, NINDS, NIH, USA

1986, PhD in Biochemistry, Graduate School, Seoul National University, Korea

PUBLICATION (5 years)

[1] Kim et al. A genome-wide association study of a coronary artery disease risk variant J Hum Genet. 58(3):120-6 (2013).

[2] Kim et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. (2012) Mar 27;14(2):R56

[3] Cho et al. Meta-analysis of genome-wide association studies identifies 8 new loci for type 2 diabetes in East AsiansNat Genet. 44(1): (2011) 67-72

[4] The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret et al.Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367) (2011):103-9.

[5] Kim et al. Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. Nat Genet. 43(10): (2011); 990-5.

[6] Fox et al. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study.Hum Mol Genet. 20(2011): 2273-84.

[7] Jung et al. Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 5(7) (2010): e11855

[8] HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 326 (2009):1541-5.

[9] Cho et al. A large scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics 41(2009):527-34.

Curriculum Vitae

Youngil Koh, MD

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
Go01@chol.com
+82-1091175012

CAREER & EDUCATION

2013.5 - Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea
2010.3 - 2013.4 Public Service Doctor, Kkotdongnae, Gapyeong, Korea (Military service)
2010.3 - PhD. Candidate, Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Masters in Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
2005.2, MD, Seoul National University College of Medicine, Seoul, Korea

AWARDS

2012 Best Oral Presentation Award, Korean Association for Clinical Oncology Annual Meeting
2011 Best Oral Presentation Award, Korea Cancer Association Annaul Meeting
2008 Travel Award, American Society of Hematology Annual Meeting
2008 Best doctor for patients, Seoul National University Hospital
1998 Bronze medal, 39th International Mathematics Olympiad, Taiwan
1997 Silver medal, 38th International Mathematics Olympiad, Argentina

PUBLICATION (recent 2 years, 1st author only, including co-first author)

[1] Park S, Koh Y, Jung SH, Chung YJ. Application of array comparative genomic hybridization in chronic myeloid leukemia. Methods in molecular biology 2013;973:55-68.
[2] Kim I, Koh Y, Yoon SS, et al. Fludarabine, cytarabine, and attenuated-dose idarubicin (m-FLAI) combination therapy for elderly acute myeloid leukemia patients. American journal of hematology 2013;88:10-5.
[3] Koh Y, Lim HY, Ahn JH, et al. Phase II trial of everolimus for the treatment of nonclear-cell renal cell carcinoma. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2013;24:1026-31.
[4] Koh Y, Kim I, Shin DY, et al. Polymorphisms in genes that regulate cyclosporine metabolism affect cyclosporine blood levels and clinical outcomes in patients who receive allogeneic hematopoietic stem cell transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation 2012;18:37-43.
[5] Koh Y, Lee HE, Oh DY, et al. The lack of CD34 expression in gastrointestinal stromal tumors is related to cystic degeneration following imatinib use. Japanese journal of clinical oncology 2012;42:1020-7.

Curriculum Vitae

Kwang-Sung Ahn, Ph.D

PDxen, Inc
Functioal Genome Institute,
Junggook-dong, Gwangjin-gu
and Technology,
Seoul , Korea
Kwangsung.ahn@gmail.com
+82-10-7722-2460

PROFESSIONAL EXPERIENCE

1] Cancer Research Center, Seoul National University, Seoul, Korea. Lab manger (Research Professor) [2004 – present]
Identification of prognostic markers and Functional analysis of drug responsive genes in Multiple myeloma and Acute Myeloid Leukemia.
2] Genome Research Center, Samsung Biomedical Research Center, Samsung Seoul Hospital, Seoul, Korea. (Research Professor) – Form Oct. 2001 to Oct. 2005
Functional analysis of metastatic genes in metastatic animal model.
3] Center for Health Science, School of Dentistry, University of California Los Angeles, CA, USA

CAREER & EDUCATION

1996 – 1997: CENTER FOR HEALTH SCIENCE, SCHOOL OF DENTISTRY, UNIVERSITY OF CALIFORNIA LOS ANGELES, LOS ANGELES, CALIFORNIA
1990 – 1996: DEPARTMENT OF BIOLOGY, GRADUATE SCHOOL OF ART & SCIENCES, UNIVERSITY OF HOUSTON, HOUSTON, TEXAS
1987 – 1989: DEPARTMENT OF BIOLOGY, GRADUATE SCHOOL OF ART & SCIENCES, LONG ISLAND UNIVERSITY AT C.W. POST, NEW YORK. MS, Department of microbiology,
1977 -1985, DEPARTMENT OF BIOLOGY, SUNGKYUNKWAN UNIVERSITY- KYUNGGI, KOREA. BS, Biology,

PUBLICATION (5 years)

[1] Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2] Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3] Park et al, RNA interference-directed caveolin-1 knockdown sensitizes SN12CPM6 cells to doxorubicin-induced apoptosis and reduces lung metastasis. Tumour Biol. 2010 6:643-50.
[4] Park et al, Establishment of a new Glivec-resistant chronic myeloid leukemia cell line, SNUCML-02, using an in vivo model. Exp Hematol. 2010 38(9):773-81.
[5] Cha et al Slug suppression induces apoptosis via Puma transactivation in rheumatoid arthritis fibroblast-like synoviocytes treated with hydrogen peroxide. Exp Mol Med. 2010 30;42(6):428-36.
[6] Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

Curriculum Vitae

Ji Wan Park, MPH, PhD

Hallym University Colledge of Medicine
1 Hallymdaehak-gil,
200-702, Chuncheon, Korea
jwpark@hallym.ac.kr
+82- 33-248-2691

PROFESSIONAL EXPERIENCE

2010-present Committee member, Korean Genome Organization
2007-present Primary Instructor, Asian Institute in Statistical Genetics and Genomics
2006 Travel Award, the 2nd North American Congress of Epidemiology

CAREER & EDUCATION

2009–present Associate professor, Dept. of Medical Genetics, Hallym Univ. College of Medicine
2006-2009 Senior Scientist, Samsung Biomedical Research Institute, Seoul, Korea
2005-2006 Post-doctoral fellow/Instructor, Dept. of Epidemiology, Johns Hopkins University, Baltimore, MD, USA
2000–2005 Ph.D. in Epidemiology (Human Genetics/Genetic Epidemiology Track), The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
1987–1989 MPH in Epidemiology, Seoul National University, School of Public Health, Seoul, Korea
1983–1987 BS, Korea University, Seoul, Korea
1990-1992 Epidemiologist, US Army Medical Research Center, Seoul, Korea

PUBLICATIONS (Selected)

[1]Heo SG, Hwang JY, Uhmn S, Go MJ, Oh B, Lee JY, Park JW. Male-specific genetic effect on hypertension and metabolic disorders. Hum. Genet. 2013 October;doi:10.1007 (Epub ahead of print).
[2]Heo SG, Hong EP, Park JW. Genetic Risk Prediction for Normal-Karyotype Acute Myeloid Leukemia Using Whole Exome Sequencing. Genomics Inform. 2013;11:46-51.
[3]Yoon D, Park SK, Kang D, Park T, Park JW. Meta-analysis of homogeneous sub-groups reveals association between PDE4D gene variants and ischemic stroke. Neuroepidemiology. 2011;36:213-22.
[4]Jee SH, Sull JW, Lee JE, Shin C, Park J, Kimm H, Cho EY, Shin ES, YUN JE, Park JW, et al. Adiponectin concentrations: a genome-wide association study. Am J Hum Genet. 2010;87:545-52.
[5]Cho YS, Go MJ, Kim YJ,…, Park JW, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet. 2009;41:527-34.
[6]Park JW, et al. Family history of diabetes and risk of atherosclerotic cardiovascular disease in Korean men and women. Atherosclerosis 2008;197:224-31.
[7]Park JW, et al. BMI and stroke risk in Korean women. Obesity 2008;16:396-401.
[8]Park JW, et al. Association between IRF6 and nonsyndromic cleft lip with or without cleft palate in 4 populations. Genet Med. 2007;9:219-27.
[9]Park JW, et al. High throughput SNP and expression analyses of candidate genes for nonsyndromic oral clefts. J Med Genet. 2006;43:598-608.
[10]Park JW, et al. Comparing whole genome amplification methods and sources of biological sample for single-nucleotide polymorphism genotyping. Clin Chem. 2005;51:1520-3.

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

Defining the role of Epstein-Barr Virus (EBV) in cancer: Exploration of DNA integration pattern and oncogenesis mechanism in various tumors

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Sung-Soo Yoon, Seoul National University Hospital (SNUH), Seoul, Korea / member of ICGC
Keun-Chil Park, Samsung Medical Center (SMC), Seoul, Korea / Member of ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Youngil Koh, SNUH, Seoul, Korea Hyun Sub Cheong, SNP Genetics Inc, Seoul, Korea	Jong-Sun Jung, Syntekabio Inc, Seoul, Korea Jung Kyoon Choi, KAIST, Daejeon, Korea

Background and preliminary data

Contribution of viral infection to cancer development and progression is well-known. However, the precise role of virus in the carcinogenic progress is not fully understood. Research using next generation sequencing (NGS) technique in cancer recently succeeded in the detection of viral infection in cancer tissues. Also, recently, genetic alterations caused by viral infection were discovered by NGS technique. For example, the study of hepatocellular carcinoma (HCC) & hepatitis B virus (HBV) using whole genome sequencing (WGS) and whole transcriptome sequencing (WTS) identified 255 HBV integration sites. Also the researchers observed a diverse collection of genomic perturbations near viral integration sites, including direct gene disruption, viral promoter-driven human transcription, viral-human transcript fusion, and DNA copy number alteration.

Among viruses, Epstein-Barr virus (EBV) is known to have close correlation with various cancers including stomach, lymphoma, nasopharyngeal cancer, and some leukemias. Also, the role of EBV is suggested in prostate and breast cancers. However, the precise role of **EBV** in the carcinogenic progress is not fully understood. Hence, we wish to identify the role of EBV in the carcinogenesis of various cancers using NGS technology.

Owing to the high prevalence of NK-T cell lymphoma in Korea, EBV is of main interest to our group. Our institution (Seoul National University Hospital) has an outstanding experience regarding NK-T cell lymphoma treatment and we published several articles regarding the clinical characteristics and treatment strategy of this disease. EBV-associated malignancies are hard to treat, and we believe that they can be conquered only if the genetic changes caused by EBV infection are well delineated.

Recently, Khoury et al (J Virol, 2013, 87(16)) reported no EBV genome integration in the transcriptome of the TCGA AML patients. However, our preliminary WGS analysis suggested EBV integration in AML genomes. We mapped our WGS data of 10 AML patients to the EBV genome. Interestingly, one of the ten samples had a significant number of mapped reads, suggesting a role for EBV in the carcinogenesis of AML in this patient. More interestingly, four of the ten AML genomes showed putative EBV integration in their germline DNA. Hence, we think it will be valuable to analyze viral genome integration based on both WGS and WTS so as to reveal the contribution of viral infection in various types of cancer carcinogenesis. In particular, the investigation of both germline and tumor DNA will be profitable. The 1000 Genomes Project data can be used to assess the frequency of germline EBV integration in normal subjects. Although EBV will be of our primary interest, it will be valuable to investigate the integration of other viruses such as HTLV-1, HBV, HHV-8, HBV and HCV.

Timelines & resources dedicated to project

Timeline:

Dec 2013 – Jan 2014: Arrangement of task-force team for this project
Jan 2014 – Apr 2014: Perform viral genome dissection for analysis and perform viral genome integration analysis using open source and our WGS data
Jun 2014 – Sep 2014: Perform viral genome integration analysis using ICGC WGS and WTS data
Jun 2014 – Sep 2014: Make cross-tumor comparison and construct genomic map across tumors
Sep2014 – Dec 2014: Validate specific viral gene integration in other cancer tissues
Sep 2014 – Jan 2015: Perform functional validation if novel findings are found
Jan 2015 - Feb 2015: Manuscript preparation
Mar 2015: Manuscript submission

Resources:

Reference samples: WGS data of 1000 genome project
Tumor samples: WGS and WTS of pancreatic cancer, gastric cancer, head and neck cancer, lymphoma, acute myeloid leukemia, bone cancer, melanoma, bladder cancer, prostate cancer, non-small cell lung cancer, and kidney cancer

Research proposal

We suggest looking at two factors using NGS data to evaluate viral genome integration in cancers

1) Viral genome integration in cancer DNA
2) Viral genome integration in germline DNA (both in cancer patients and in healthy controls)

Cancers where EBV is known to have a definite role in carcinogenesis should be used as a reference in the analysis of viral genome integration in cancer DNA. So, NK/T-cell lymphoma and nasopharyngeal cancer should be included in this analysis. And to perform this research using WGS data of ICGC, we propose the followings.

1. First, we will dissect EBV genome so as to delineate precise role of EBV genome integration in cancer
2. We will analyze the WGS, and WTS data of pancreatic cancer, gastric cancer, head and neck cancer, lymphoma, acute myeloid leukemia, bone cancer, liver cancer, melanoma, bladder cancer, prostate cancer, non-small cell lung cancer, and kidney cancer. In order to reveal EBV integration relevantly, we selected the tumor types according to the followings criteria.
    1) Include cancers that are known to have strong association with EBV: Head and neck cancer, lymphoma
    2) Include cancers that are known to have close relationship with EBV: Gastric cancer, prostate cancer
    3) Include cancers that do have epidemiologic difference but do not have causative agents: Pancreatic cancer, acute myeloid leukemia, bone cancer, melanoma (some subtype does not have association with sun-exposure), non-small cell lung cancer, and kidney cancer
3. We will analyze integration of EBV genome in cancer DNA in these tumors
4. We will analyze integration of EBV genome in gremline DNA in these patients
5. We will analyze integration of EBV genome in healthy controls (using 1000 genome project data)
6. Finally, we will make a viral integration map of both germline and somatic DNA across tumor types using data generated by step 5 through step 7.
7. We will compare EBV genome integration rate between healthy controls and cancer patients
8. We will check the specific viral integration found by step 5 through step 7 in our validation cohorts composed of AML, ALL, and NSCLC patients.
9. Functional validation of novel findings found by step 5-7 will be followed.

Legacy plans

For the viral DNA integration analysis, we will mainly use utilities in open-source. If we design and develop analysis algorithm for viral genome integration analysis using WGS and/or WTS data during this pan-cancer project, we will disclose and produce publication-ready source in ETRI/Syntekabio located in Seoul Korea. A documented virtual machine will be opened in a cloud computing system and will be embodied in executable code.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012
[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7
[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7
[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea
1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea
1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.
1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.
1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7
[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.
[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.
[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

Keunchil Park, M.D., Ph.D.

Keunchil Park is Professor of the Division of Hematology-Oncology, Sungkyunkwan University School of Medicine in Seoul, Korea. Prof. Park is Director of the Medical Nano Element Development Center, and is the Principal Investigator of the 'Identification of Novel Therapeutic Targets in Lung Cancer with Unmet Need' of the National Project for Personalized Genomic Medicine(PGM21), both of which are funded by the Ministry of Health and Welfare, Korea.
Professor Park has served many domestic academic societies, e.g., Chair of the Scientific Committee of the Korean Cancer Association, Chair of the Lung Cancer Committee of the Korean Cancer Study Group(KCSG). Prof. Park also served as Chairman of the Board of Directors, Korean Association for Clinical Oncology (KACO) since June 2010 until May 2012.
Prof. Park has been also very actively involved in and served many international activities, such as the Scientific Secretary of the 12th WCLC (Sept, 2007), and the Chairman of the 4th Asia Pacific Lung Cancer Conference (Dec, 2010). He was elected as the Board of Directors of the IASLC and is serving as Associate Editor for the Journal of Thoracic Oncology (JTO) and on editorial board of the Asia-Pacific Journal of Clinical Oncology.
Prof. Park's main interests include the translational and early clinical researches for the treatments of upper aero-digestive tract cancers, especially lung cancer. Recently Dr. Park is leading several early clinical trials of the targeted agents as well as many pre-clinical development programs internationally. Prof. Park has several book chapters and authored more than 200 peer-reviewed publications in national and international journals.

Curriculum Vitae

Youngil Koh, MD

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
Go01@chol.com
+82-1091175012

CAREER & EDUCATION

2013.5 - Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea
2010.3 - 2013.4 Public Service Doctor, Kkotdongnae, Gapyeong, Korea (Military service)
2010.3 - PhD. Candidate, Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Masters in Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
2005.2, MD, Seoul National University College of Medicine, Seoul, Korea

AWARDS

2012 Best Oral Presentation Award, Korean Association for Clinical Oncology Annual Meeting
2011 Best Oral Presentation Award, Korea Cancer Association Annaul Meeting
2008 Travel Award, American Society of Hematology Annual Meeting
2008 Best doctor for patients, Seoul National University Hospital
1998 Bronze medal, 39th International Mathematics Olympiad, Taiwan
1997 Silver medal, 38th International Mathematics Olympiad, Argentina

PUBLICATION (recent 2 years, 1st author only, including co-first author)

[1] Park S, Koh Y, Jung SH, Chung YJ. Application of array comparative genomic hybridization in chronic myeloid leukemia. Methods in molecular biology 2013;973:55-68.
[2] Kim I, Koh Y, Yoon SS, et al. Fludarabine, cytarabine, and attenuated-dose idarubicin (m-FLAI) combination therapy for elderly acute myeloid leukemia patients. American journal of hematology 2013;88:10-5.
[3] Koh Y, Lim HY, Ahn JH, et al. Phase II trial of everolimus for the treatment of nonclear-cell renal cell carcinoma. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2013;24:1026-31.
[4] Koh Y, Kim I, Shin DY, et al. Polymorphisms in genes that regulate cyclosporine metabolism affect cyclosporine blood levels and clinical outcomes in patients who receive allogeneic hematopoietic stem cell transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation 2012;18:37-43.
[5] Koh Y, Lee HE, Oh DY, et al. The lack of CD34 expression in gastrointestinal stromal tumors is related to cystic degeneration following imatinib use. Japanese journal of clinical oncology 2012;42:1020-7.

Curriculum Vitae

Hyun Sub Cheong, PhD

SNP Genetics, Inc.
#TE1007, Teihard Hall, Sogang Univ.,
Shinsu-dong, Mapo-gu,
121-742, Seoul, Korea
chhs@snp-genetics.com
+82-1090680268

CAREER & EDUCATION

2013.5 Analysis Team Manager, SNP Genetics, Inc. National Research Lab

2012.2 Seoul National University, PhD in Tumor Biology, College of Medicine

2004.9 Seoul National University, Completed Doctorate Course in School of Agricultural Biotechnology

PUBLICATION (recent 5 years, 1st author only)

[1] ADFP promoter polymorphism associated with marbling score in Korean cattle. Cheong HS, Yoon DH, Bae JS, Kim LH, Kim EM, Kim JO, Hong J, Kim N, Shin HD. BMB Rep. 2009 Aug 31;42(8):529-34.

[2] Common CYP7A1 Promoter Polymorphism Associated With Risk of Neuromyelitis Optica. Kim HJ, Park HY, Kim E, Lee KS, Kim KK, Choi BO, Kim SM, Bae JS, Lee SO, Chun JY, Park TJ, Cheong HS, Jo I, Shin HD. Neurobiol Dis. 2010 Feb;37(2):349-55.

[3] Identification of genetic polymorphisms in bovine mitochondrial deoxyribonucleic acid. Kim E, Cheong HS, Bae JS, Chun JY, Park TJ, Lee K, Yun Y, Shin HD. J Anim Sci. 2010 Aug 88(8):2551-5.

[4] Association of RANBP1 haplotype with smooth pursuit eye movement abnormality. Cheong HS, Park BL, Kim EM, Park CS, Sohn JW, Kim BJ, Kim JW, Kim KH, Shin TM, Choi IG, Han SW, Hwang J, Koh I, Shin HD, Woo SI. Am J Med Genet B Neuropsychiatr Genet. 2011 Jan; 156(1):67-71.

[5] Genome-wide methylation profile of nasal polyps: relation to aspirin hypersensitivity in asthmatics. Cheong HS, Park SM, Kim MO, Park JS, Lee JY, Byun JY, Park BL, Shin HD, Park CS. Allergy. 2011. 66(5):637-644.

[6] Epigenetic modification of retinoic acid-treated human embryonic stem cells. Cheong HS, Lee HC, Park BL, Kim H, Jang MJ, Han YM, Kim SY, Kim YS, Shin HD. BMB Rep. 2010 Dec;43(12):830-5.

[7] Screening of genetic variations of SLC15A2, SLC22A1, SLC22A2 and SLC22A6 genes. Cheong HS, Kim HD, Na HS, Kim JO, Kim LH, Kim SH, Bae JS, Chung MW, Shin HD. J Hum Genet. 2011 Sep;56(9):666-70.

[8] Development of discrimination SNP markers for Hanwoo (Korean native cattle). Cheong HS, Kim LH, Namgoong S, Shin HD. Meat Sci. 2013 Mar 16;94(3):355-359.

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

JUNG KYOON CHOI

Dept. Bio and Brain Engineering
KAIST
335 Gwahak-ro, Yooseong-goo
Daejeon 305-701
Republic of Korea
T +82-42-350-4327
F +82-42-350-8834
jungkyoon@kaist.ac.kr
http://omics.kaist.ac.kr

ASSISTANT PROFESSOR — OCT 2009 ~ PRESENT
Dept. Bio and Brain Engineering, KAIST

PRINCIPAL INVESTIGATOR — JAN 2010 ~ DEC 2012
Genome Institute of Singapore (Joint appointment)

Publications - As corresponding or first author

Genome-wide reorganization of histone h2AX toward particular fragile sites on cell activation. *Nucleic Acids Res.* **in press.**

Regulation of the boundaries of accessible chromatin. *PLoS Genet.* **9**, e1003778 (2013).

Genetic landscape of open chromatin in yeast. *PLoS Genet.* **9**, e1003229 (2013).

Genome-wide profiles of H2AX and gamma-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. *Nucleic Acids Res.* **40**, 5965-5974 (2012).

Controlling transcriptional programs for cellular adaptation by chromatin regulation. *Mol. BioSyst.* **7**, 1713-1719 (2011).

Genetic and metabolic characterization of insomnia. *PLoS ONE* **6**, e18455 (2011).

Systems biology and epigenetic gene regulation. *IET Syst. Biol.* **4**, 289-295 (2010).

Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol.* **11**, R70 (2010).

Nucleosome deposition and DNA methylation at coding region boundaries. *Genome Biol.* **10**, R89 (2009).

Implications of the nucleosome code in regulatory variation, adaptation and evolution. *Epigenetics* **4**, 291-295 (2009).

Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat. Genet.* **41**, 498-503 (2009).

Stochastic and regulatory role of chromatin silencing in genomics response to environmental changes. *PLoS ONE* **3**, e3002 (2008).

Epigenetic regulation and the variability of gene expression. *Nat. Genet.* **40**, 141-147 (2008).

Environmental effects on gene expression phenotype have regional biases in the human genome. *Genetics* **175**, 1607–1613 (2007).

Impact of transcriptional properties on essentiality and evolutionary rate. *Genetics* **175**, 199–206 (2007).

Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21**, 4348-4355 (2005).

Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett.* **565**, 93-100 (2004).

Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19** Suppl. 1, i84-i90 (2003).

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

Investigation of crosstalk between germline and tumor DNA in patients with germline cancer-hotspot alterations

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Sung-Soo Yoon, Seoul National University Hospital (SNUH), Seoul, Korea / Member of ICGC
Keunchil Park, Samsung Medical Center (SMC), Seoul, Korea / Member of ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)
Youngil Koh, SNUH, Seoul, Korea Youngwook Kim, SMC, Seoul, Korea	Jong-Sun Jung, Syntekabio Inc, Seoul, Korea Kwang-Sung Ahn, PDxen Inc, Seoul, Korea

Background and preliminary data

Current cancer research using massively parallel sequencing technologies primarily focuses on somatic changes in the cancer genome. However, it is anticipated that individual germline variations might contribute to tumorigenesis, as exemplified by several well-known examples, such as BRCA2 variations in breast cancer. Thus, it is also important to explore crosstalk between germline variation and somatic mutation in cancer.

Since it is impossible to survey all kinds of known germline variations including SNP's in whole genome sequencing (WGS) data, we think it would be practical to focus on the "famous and causative" hotspots. That is, we think it is important to look at whether "genetic alteration of somatic cancer hotspots" is also present in germline of cancer patients. (in fact, a couple of studies are currently ongoing)

As a preliminary study, using whole-exome sequencing (WES) results of 80 acute myeloid leukemia (AML) samples and 180 non-small cell lung cancer (NSCLC) samples, we searched for cancer hotspots in germline DNA of these samples. Interestingly, many non-synonymous changes including MLH1 (Chr3;37067240, T>A), KIT (Chr4;55593464, A>T), and MET (Chr 7;116340262, A>C) were observed in germline of both AML and NSCLC patients. On the contrary, MLH1 germline change was not found in 18 healthy control samples. Frequency of MLH1, KIT, and MET germline changes were similar in AML and NSCLC patients. In contrast to AML, some of NSCLC patients had PTEN and TP53 germline changes. Furthermore, STK11 germline change was found only in NSCLC patients. Hence, we conjecture that patterns of germline changes in cancer hotspots might be different according to the cancer types. Moreover, to focus on the crosstalk between germline change and somatic mutation, we analyzed the somatic mutations of "patients with cancer hotspot germline mutations". When this was performed, we made some interesting observations: Among 3 AML patients with MLH1 germline changes, only one patient whose disease was very aggressive had NBPF gene non-synonymous somatic mutations, while the other two patients whose disease was indolent did not harbor somatic NBPF mutation.

To summarize, in the preliminary study of WES data of AML and NSCLC samples, the following was discovered:

1. Cancer hotspot variations, which were traditionally known as somatic changes, might be also present in germline of cancer patients, and their frequency might be different across various cancer subtypes.
2. Even in patients with same germline mutations, disease characteristics, including prognosis are vastly different according to the somatic mutations.

Hence we think it is valuable to make cross-tumor comparison of both germline mutations and somatic mutations using NGS data to understand cancer pathophysiology.

Timelines & resources dedicated to project

Timeline:

Dec 2013 – Jan 2014: Arrangement of task-force team for this project
Jan 2014 – Apr 2014: Select cancer hotspots to see in germline data. This includes SNV's, small indels, structural variants, and copy number changes. Perform germline calling using our WGS test data set (10 AML samples).
Jun 2014 – Jul 2014: Analysis of pan-cancer germline data from ICGC
Jun 2014 – Sep 2014: Variant call of somatic changes of samples with germline variation
Sep 2014 – Oct 2014: Make cross-tumor comparison and construct genomic map across tumors
Sep 2014 – Jan 2015: Perform functional validation if novel findings are found
Jan - Feb 2015: Manuscript preparation
Mar 2015:Manuscript submission

Resources:

Cancer Genome: WGS, WES and WTS of acute myeloid leukemia, non-small cell lung cancer (Excluding squamous), triple negative breast cancer, colon cancer, stomach cancer, kidney cancer, bone cancer, glioma and pancreatic cancer.
Reference Genome (to exclude insignificant germline variation): WGS data from 1000 genome project

Research proposal

We think it is valuable to make cross-tumor comparison of both germline mutations and somatic mutations at the same time using NGS data to understand cancer pathophysiology.. To perform this research using WGS data of ICGC, we propose the following:

8. Our group will compile a list of cancer-associated genes, known to be statistically enriched in each type of cancer. The genetic variations we consider include SNV's, small indels and structural variations.
9. We will analyze the WGS data of AML, lung cancer (excluding squamous histology), triple negative breast cancer, colon cancer, stomach cancer, kidney cancer, bone cancer, glioma and pancreatic cancer. In order to see cross-talks between germline changes and somatic changes, we selected the tumor types according to the followings criteria.
    5) Exclude cancers that are hormone dependent (such as ovarian cancer, prostate cancer, hormone-receptor positive breast cancer).
    6) Exclude cancers that are known to have definite pathogen, including smoking and viruses. (such as head and neck cancer, esophageal cancer, oral cancer, cervical cancer and liver cancer).
10. Next, we would like to look at the normal WGS of the cancer types listed above to select samples that have potential germline changes in cancer-associated genes. These germline changes include variations in the list of genes compiled by step 1. (remove this sentence) We will utilize 1000 genome data to filter cancer-enriched germ line variations of cancer-associated genes.
11. We will also perform somatic variant calling using WGS data of these samples selected by step 3.
12. Finally, we will make a genomic map of both germline and somatic DNA across tumor types using data generated by step 3 and step 4.
13. External/functional validation of findings found by step 3-5 will be followed.

Legacy plans

For the calling of somatic mutation, we will mainly use utilities in open-source. If we design and develop calling algorithm for somatic mutations detection during this pan-cancer project, we will disclose and produce publication-ready source in ETRI/Syntekabio located in Seoul Korea. A documented virtual machine will be opened in a cloud computing system and will be embodied in executable code.

Curriculum Vitae

Sung-Soo Yoon, MD, Ph.D

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
ssysmc@snu.ac.kr
+82-1047546706

PROFESSIONAL EXPERIENCE

< Basic and Translational Research in Hematologic Malignancies >

[1] Chairperson, Korean Multiple Myeloma Working Party (KMMWP) under the auspice of Korean Society of Hematology (KSH), since 2012
[2] Director, Division of Hematology/Medical Oncology, SNUH, since 2012.7
[3] Director, Center for Hematologic Malignancy, SNUH, Cancer Hospital, since 2012.7
[4] Principal investigator in various clinical trials (from phase I through phase III)

CAREER & EDUCATION

2006.4-present, Professor of Medicine, SNUH, Seoul, Korea
1996.2, PhD, Seoul National University College of Medicine, Seoul, Korea
1992.8-1994.12, Visiting Scientist, Department of Cell Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.
1991.5-1992.4, Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea.
1988. 2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
1984. 2, MD, Seoul National University College of Medicine, Seoul, Korea

PUBLICATION (recent years, Corresponding author only)

[1]Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2]Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3]Kim et al, Hepatic sinusoidal obstruction syndrome after allogeneic hematopoietic stem cell transplantation in adult patients with idiopathic aplastic anemia. Leuk Res. 2013 Oct;37(10):1241-7
[4]Kim et al, Recombinant human epidermal growth factor on oral mucositis induced by intensive chemotherapy with stem cell transplantation. Am J Hematol. 2013 Feb;88(2):107-12.
[5]Yhim et al, Matched-pair analysis to compare the outcomes of a second salvage auto-SCT to systemic chemotherapy alone in patients with multiple myeloma who relapsed after front-line auto-SCT. Bone Marrow Transplant. 2013 Mar;48(3):425-32.
[6]Kim et al, Mitoxantrone, etoposide, cytarabine, and melphalan (NEAM) followed by autologous stem cell transplantation for patients with chemosensitive aggressive non-Hodgkin lymphoma. Am J Hematol. 2012 May;87(5):479-83.

Keunchil Park, M.D., Ph.D.

Keunchil Park is Professor of the Division of Hematology-Oncology, Sungkyunkwan University School of Medicine in Seoul, Korea. Prof. Park is Director of the Medical Nano Element Development Center, and is the Principal Investigator of the 'Identification of Novel Therapeutic Targets in Lung Cancer with Unmet Need' of the National Project for Personalized Genomic Medicine(PGM21), both of which are funded by the Ministry of Health and Welfare, Korea.
Professor Park has served many domestic academic societies, e.g., Chair of the Scientific Committee of the Korean Cancer Association, Chair of the Lung Cancer Committee of the Korean Cancer Study Group(KCSG). Prof. Park also served as Chairman of the Board of Directors, Korean Association for Clinical Oncology (KACO) since June 2010 until May 2012.
Prof. Park has been also very actively involved in and served many international activities, such as the Scientific Secretary of the 12th WCLC (Sept, 2007), and the Chairman of the 4th Asia Pacific Lung Cancer Conference (Dec, 2010). He was elected as the Board of Directors of the IASLC and is serving as Associate Editor for the Journal of Thoracic Oncology (JTO) and on editorial board of the Asia-Pacific Journal of Clinical Oncology.
Prof. Park's main interests include the translational and early clinical researches for the treatments of upper aero-digestive tract cancers, especially lung cancer. Recently Dr. Park is leading several early clinical trials of the targeted agents as well as many pre-clinical development programs internationally. Prof. Park has several book chapters and authored more than 200 peer-reviewed publications in national and international journals.

Curriculum Vitae

Youngil Koh, MD

Seoul National University Hospital (SNUH)
101 Daehak-ro, Jongro-gu,
110-744, Seoul, Korea
Go01@chol.com
+82-1091175012

CAREER & EDUCATION

2013.5 - Clinical Fellow in Hematology/Medical Oncology, SNUH, Seoul, Korea
2010.3 - 2013.4 Public Service Doctor, Kkotdongnae, Gapyeong, Korea (Military service)
2010.3 - PhD. Candidate, Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Masters in Molecular and Clinical Oncology, Seoul National University, Seoul, Korea
2010.2, Board Certified in Internal Medicine, SNUH, Seoul, Korea
2005.2, MD, Seoul National University College of Medicine, Seoul, Korea

AWARDS

2012 Best Oral Presentation Award, Korean Association for Clinical Oncology Annual Meeting
2011 Best Oral Presentation Award, Korea Cancer Association Annaul Meeting
2008 Travel Award, American Society of Hematology Annual Meeting
2008 Best doctor for patients, Seoul National University Hospital
1998 Bronze medal, 39th International Mathematics Olympiad, Taiwan
1997 Silver medal, 38th International Mathematics Olympiad, Argentina

PUBLICATION (recent 2 years, 1st author only, including co-first author)

[1] Park S, Koh Y, Jung SH, Chung YJ. Application of array comparative genomic hybridization in chronic myeloid leukemia. Methods in molecular biology 2013;973:55-68.
[2] Kim I, Koh Y, Yoon SS, et al. Fludarabine, cytarabine, and attenuated-dose idarubicin (m-FLAI) combination therapy for elderly acute myeloid leukemia patients. American journal of hematology 2013;88:10-5.
[3] Koh Y, Lim HY, Ahn JH, et al. Phase II trial of everolimus for the treatment of nonclear-cell renal cell carcinoma. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2013;24:1026-31.
[4] Koh Y, Kim I, Shin DY, et al. Polymorphisms in genes that regulate cyclosporine metabolism affect cyclosporine blood levels and clinical outcomes in patients who receive allogeneic hematopoietic stem cell transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation 2012;18:37-43.
[5] Koh Y, Lee HE, Oh DY, et al. The lack of CD34 expression in gastrointestinal stromal tumors is related to cystic degeneration following imatinib use. Japanese journal of clinical oncology 2012;42:1020-7.

Youngwook Kim

Samsung Biomedical Research Institute
Senior Researcher
Rm. 188 B4 Cancer Center
50 Irwon-dong Gangnam-gu
Seoul, Korea zip: 135-710
Office: 82-2-2148-7349
Cell: 82-10-2300-9856

Publications

1. **Kim Y**, Hammerman P, Kim JG, Yoon J, Lee Y, Sun J, Wilkerson M, Pedamallu C, Cibulskis K, Yoo Y, Lawrence M, Stojanov P, Carter S, Hayes N, Getz G, Meyerson M, Park K Integrative and comparative genomic analysis of lung squamous cell carcinomas in East-Asians Journal of Clinical Oncology *in press (2014)*

2. **Kim Y**, Kim J, Lee J, Bae K, Min J, Park T, Lee J, Nam Y, Park K *Tumor-Targeted Delivery of Paclitaxel using Solid Lipid Nanoparticles* Nature Communications *in review*

3. **Kim Y**, Ko J, Cui Z, Abolhoda A, Ahn JS, Ou SH, Ahn MJ, Park K *The EGFR T790M mutation in acquired resistance to an irreversible second-generation EGFR inhibitor* Mol Cancer Ther. (2012) Mar;11(3):784-91.

4. Lee S*, **Kim Y***, Sun JM, Choi YL, Kim JG, Shim YM, Park YH, Ahn JS, Park K, Han JH, Ahn MJ *Molecular profiles of EGFR, K-ras, c-met, and FGFR in pulmonary pleomorphic carcinoma, a rare lung malignancy*. J Cancer Res Clin Oncol. (2011) Aug;137(8):1203-11. * equally contributing authors

5. Oh YH*, **Kim Y***, Kim YP, Seo SW, Mitsudomi T, Ahn MJ, Park K, Kim HS *Rapid detection of the epidermal growth factor receptor mutation in non-small-cell lung cancer for analysis of acquired resistance using molecular beacons.* J Mol Diagn. (2010) Sep;12(5):644-52. *equally contributing authors

Curriculum Vitae

Jongsun Jung, Ph.D

Syntekabio, Inc
992 VentureTown,
Korea Institute of Science
and Technology,
Seoul , Korea
jung@syntekabio.com
+82-107123-9104

PROFESSIONAL EXPERIENCE

< BI tool Development in C/C++ >

[1]ADISCAN: Allelic Depth Imbalance Scanning for NGS data, 2013
[2]IGA: Indexed Genome Analysis & Integration for Genomic Data, 2006-2010
[3]RVR: Records Virtual Rack, a Tool Package for Indexing Bio Big Data, 2002-2006,
[4]LSHEBA: Local Alignment Based Protein Circular Permutation Scanning, Protein Science, 2001
[5]SHEBA: Structural Homology based Alignment, Protein Engineering, 2000
[6]PASSC: Pair to Pair Alignment of Sequence Structure Correlation, Protein Science, 2000

CAREER & EDUCATION

2009 ~, CEO/CTO, Syntekabio, Inc., Korea
2004~2007, Principal Researcher, KCDC, Korea
1996~2002, NIH/NCI, Visiting Fellow, Bethesda, MD USA
1996~1999, PH.D, Biochemistry/Bioinformatics, American Uni., Washington DC USA

PUBLICATION (5 years)

[1]Hong et al, Application of variant calling algorithms for Mendelian disorders: lessons from whole-exome sequencing in Charcot–Marie–Tooth disease. Clinical Genomics, 2013
[2]Park et al, Differential expression of MicroRNAs in patients with glioblastoma after concomitant chemoradiotherapy. OMICS. 2013 May;17(5):259-68.
[3]Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210
[4]Jung et al, Gene flow between the Korean peninsula and its neighboring countries. PLoS One. 2010 Jul 29;5(7):e11855.
[5]Hong et al, Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. J Hum Hypertens. 2010 Nov; 24(11):763-74. PMID: 20147969
[6]The HUGO Pan-Asian SNP Consortium et al., Mapping Human Genetic Diversity in Asia, Science. 2009, 326:1541-5.
[7]Jeon et al, A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med, 2009
[8]Kaput et al, Planning the human variome project: the Spain report. Hum Mutat. 2009
[9]Park et al, Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. J Hum Genet. 2008

Curriculum Vitae

Kwang-Sung Ahn, Ph.D

PDxen, Inc
Functioal Genome Institute,
Junggook-dong, Gwangjin-gu
and Technology,
Seoul , Korea
Kwangsung.ahn@gmail.com
+82-10-7722-2460

PROFESSIONAL EXPERIENCE

1] Cancer Research Center, Seoul National University, Seoul, Korea. Lab manger (Research Professor) [2004 – present]
Identification of prognostic markers and Functional analysis of drug responsive genes in Multiple myeloma and Acute Myeloid Leukemia.
2] Genome Research Center, Samsung Biomedical Research Center, Samsung Seoul Hospital, Seoul, Korea. (Research Professor) – Form Oct. 2001 to Oct. 2005
Functional analysis of metastatic genes in metastatic animal model.
3] Center for Health Science, School of Dentistry, University of California Los Angeles, CA, USA

CAREER & EDUCATION

1996 – 1997: CENTER FOR HEALTH SCIENCE, SCHOOL OF DENTISTRY, UNIVERSITY OF CALIFORNIA LOS ANGELES, LOS ANGELES, CALIFORNIA
1990 – 1996: DEPARTMENT OF BIOLOGY, GRADUATE SCHOOL OF ART & SCIENCES, UNIVERSITY OF HOUSTON, HOUSTON, TEXAS
1987 – 1989: DEPARTMENT OF BIOLOGY, GRADUATE SCHOOL OF ART & SCIENCES, LONG ISLAND UNIVERSITY AT C.W. POST, NEW YORK. MS, Department of microbiology,
1977 -1985, DEPARTMENT OF BIOLOGY, SUNGKYUNKWAN UNIVERSITY- KYUNGGI, KOREA. BS, Biology,

PUBLICATION (5 years)

[1] Park et al, Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. BMB Rep. In Press
[2] Lee et al, TNF α mediated IL-6 secretion is regulated by JAK/STAT pathway but not by MEK phosphorylation and AKT phosphorylation in U266 multiple myeloma cells. Biomed Res Int. In Press.
[3] Park et al, RNA interference-directed caveolin-1 knockdown sensitizes SN12CPM6 cells to doxorubicin-induced apoptosis and reduces lung metastasis. Tumour Biol. 2010 6:643-50.
[4] Park et al, Establishment of a new Glivec-resistant chronic myeloid leukemia cell line, SNUCML-02, using an in vivo model. Exp Hematol. 2010 38(9):773-81.
[5] Cha et al Slug suppression induces apoptosis via Puma transactivation in rheumatoid arthritis fibroblast-like synoviocytes treated with hydrogen peroxide. Exp Mol Med. 2010 30;42(6):428-36.
[6] Kim et al, Proteomic and bioinformatic analysis of membrane proteome in type 2 diabetic mouse liver. Proteomics. 2013 Jan 24. doi: 10.1002/pmic.201200210

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The germline component of common cancer: defining common genes and pathways of cancer susceptibility

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Xavier Estivill, MD, PhD, Center for Genomic Regulation (CRG), Barcelona, Spanish CLL-Genomics Consortium
Stephan Ossowski, PhD, Center for Genomic Regulation (CRG), Barcelona, Spanish CLL-Genomics Consortium

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Aparna Prasad, PhD, CRG Oliver Drechsel, PhD, CRG	Manolis Kogevinas, MD, CREAL Gemma Castaño, PhD, CREAL

Background and preliminary data

Most genomic studies in human cancers have so far been focused on identifying somatic aberrations driving the cancerous state of the tumor cells. While the landscape of somatic mutations that occur in human cancers provide crucial information about tumor development and evolution, the initial triggers of tumor development can only be evaluated by analyzing the genetic variants that are present in the germline, before tumor development. Cancer susceptibility studies have been performed through linkage and genome wide association studies (GWAS), leading to the identification of many loci harboring genes and variants potentially involved in hereditary risk to cancer. However, the loci identified to date represent a small proportion of the heritability of cancer, particularly of those that are defined as sporadic, suggesting that the susceptibility to cancer likely resides on a large number of rare variants undetectable by GWAS. We speculate that cancer patients have a convergence of rare variants in genes of cancer related pathways like DNA damage repair, cell cycle control, apoptosis etc., and that susceptibility genes and pathways highly overlap between different cancer types. Thus we propose to analyze enrichment of germline rare variants in genes and pathways across multiple cancer types in order to identify the genes, pathways protein interactions, and their interplay, involved in cancer risk.

Timelines & resources dedicated to project

We have previously developed a novel method for identification of genes or gene networks harboring a significantly increased number of rare variants in cancer cases. We have successfully applied our new method termed REWAS (Rare variant Exome-wide Association Study) to over 350 cases of CLL, sequenced as part of the International Cancer Genome Consortium (ICGC) and have identified accumulation of mutations in genes that are also involved in breast cancer susceptibility, e.g. *BRCA1*, *ATM*, *CHEK2,* etc. (in preparation). With the pan-cancer data set, we will have excellent resources to interrogate the landscape of germline mutations in 23 different cancer types. The large number of samples will substantially increase the statistical power of our REWAS method to identify rare variant enrichments in common cancer susceptibility genes. As an added benefit our proposed project will produce a comprehensive catalogue of germline variants, including SNPs, indels and CNVs, which will facilitate the discovery of combinations of genetic changes in susceptibility genes that predispose to cancer. To perform our project we need access to whole exome sequencing (WES) data from ~8,000 matched tumor and normal samples, optimally including deep clinical information for each patient. As the germline variant enrichment analysis pipeline has successfully been set up for the ICGC-CLL project, analysis with the whole data set could start immediately after getting access to the data.

Given pre-existing sequence alignments (in BAM format) the variant analysis (SNPs, indels) of 8,000 WES datasets requires approximately 20 days on 24 compute nodes with at least 8 cores and 64GB memory each. The procedure uses GATK multi-sample by chromosome. GATK-MS has a linear increase in computation time by sample and has performed a similar analysis on 1000 in-house samples in 2 days. Variant annotation,

filtering and quality control require approx. two weeks. CNV prediction using Conifer and ClinCNV (unpublished in-house tool) will require 2 weeks each on a single high memory (128GB) server.

REWAS on the complete data set can be performed in less then one week on a single server, however multiple runs will be performed to: a) test different combinations of cancers, b) perform REWAS on genes, pathways and protein-protein interaction networks, and c) optimize the method for the large number of samples. As prototypes of all algorithms are in place we would thus expect initial results in about 2 to 3 month after the start of the project. Time required for follow up statistic and functional analysis as well as replication of implicated genes depends on the results of the REWAS and cannot be estimated precisely. However replication of a set of candidate genes by targeted sequencing of several thousand cancer cases from the population-based multicase-control study (MCC-SPAIN) (colorectal cancer, breast cancer, gastroesophageal cancer, prostate cancer and chronic lymphocytic leukemia) could be performed within 3 to 6 month.

Research proposal

Research aim and objectives:

The aim of this project is to identify the landscape of (rare) germline that predispose an individual to a certain cancer type as well as susceptibility genes and pathways increasing risk for multiple cancer types.

Objective 1: Defining the landscape of germline mutations in several cancer types

We will utilize ~8,000 matched tumor and normal exome pairs aligned to hg19 to produce a catalogue of genic germline variants using e.g. GATK-MS (GATK UnifiedGenotyper) for SNP and indel prediction and Conifer and ClinCNV (manuscript in preparation) for predicting copy number variants (CNVs). As an intrinsic quality control we will require germline variants to be found in both normal and tumor tissue. We assume that alignments in BAM format are already available, otherwise we propose to use bwa-mem to align all samples. We further assume that somatic mutation predictions are available at the time of analysis and otherwise propose the use of e.g. Mutect, indelocator, Clindel and ClinCNV (latter two are in-house tools, manuscript in preparation) to predict somatic SNPs, indels and CNVs, which are required for correlation to germline mutations. Variants are annotated using the in-house annotation pipeline eDIVA (manuscript in preparation) that builds on the snpeff tool for functional annotation, sift, polyphen2, mutation assessor and condel for damage estimation, several population allele frequency datasets (1000genomes, EVS, dbSNP, GEEVS) for identification of rare variants, cancer specific variant data from Cosmic and Intogen as well as many other OMICs and comparative genomics resources retrieved from UCSC Genome Browser, NCBI or Ensembl.

Objective 2: Identification of potential risk genes enriched in rare and damaging germline mutations

Rare variants exome-wide association study (REWAS): We have previously developed a novel approach for identification of cancer susceptibility genes and pathways that builds on the hypothesis that these genes and pathways are enriched for rare and highly damaging mutations. The approach termed rare variant exome-wide association study (REWAS) is designed similar to a case-control mutation enrichment analysis, however variants with a population allele frequency above 1% or low functional impact are removed and the remaining variants are accumulated by gene (or by pathway) in cases and controls, respectively. Genes overlapping rare CNVs are considered as highly damaged. A Fisher exact test followed by Benjamini-Hochberg correction is applied to identify genes significantly enriched for rare and damaging mutations in cases vs. control.

We will use annotated variant calls of 8,000 cancer samples from multiple cancer types to define genes and pathways enriched in rare mutations a) across all cancers, b) across subsets of cancers, and c) in single cancer types. To improve the flexibility of REWAS we will adjust our approach to work on protein-interaction sub-networks (PPI-SNs), build by integrating knowledge from PPIs (String PPI database) and identify recurrently mutated genes. In comparison to REWAS on pathways we expect this new approach to more specifically identify protein interactions, regulatory mechanisms or protein complexes involved in cancer development. Candidate cancer susceptibility genes identified by REWAS and PPI-SNs will be used for replication by targeted sequencing of samples from the population-based multicase-control study (MCC-SPAIN), containing at least 1,000 cases of different cancer types (CLL, breast, prostate, stomach and colon), and control samples.

Objective 3: Functional studies of susceptibility genes and pathways

We expect to obtain a set of genes that potentially involved in the susceptibility to develop different cancer types. These genes will be ranked on the basis of significant effect on the susceptibility to cancer, integrating knowledge about somatic driver mutations. We will use *Saccharomyces cerevisiae* (yeast) for functional analyses of the selected genes to evaluate convergent data on mutations that give susceptibility to cancer.

Legacy plans
We are developing a statistical method called “Rare variants exome-wide association study” (REWAS). The project will provide a repertoire of the genes and variants that are involved in cancer susceptibility. We expect that the number of genes will be large and that mutations in several genes will combine to provide the susceptibility to cancer in general. The knowledge should have important consequences in prevention actions.

Xavier Estivill Curriculum Vitae Sketch

Name: **Xavier Estivill** Position title: Senior Group Leader Genomics and Disease
Bioinformatics and Genomics Programme
Birth date: 28/09/1955 Citizenship: Spanish Gender (M/F): M
Office address: Doctor Aiguader 88, 08003 Barcelona Office telephone: +3493 316 0138
E-mail: xavier.estivill@crg.eu

1. Education / Training

Degree	Institution	Year(s)
Medicine	Universitat Autònoma de Barcelona	1979
Haematology	Universitat Autònoma de Barcelona	1985
MD in Genetics	Universitat Autònoma de Barcelona	1987
PhD in Molecular Genetics	University of London, UK	1995

2. Professional academic positions

Dates (from-until)	**Position**	**Department & Institution**
1980-1981	Medical Intern Resident	Residencia Sanitaria de Bellvitge, Spain
1982-1985	Medical Intern Resident	Hospital de Sant Pau, Spain
1985	Research Fellow	Università degli Estudi di Torino, Italy
1986-1988	Research Fellow	St. Mary's Hospital, I College, London, UK
1988-1990	Research Professor	Fundació Investigació Sant Pau, Spain
1991-1997	Head of the Genetics Service	Hospital Clínico de Barcelona, Spain
1990-2001	Head of the Molecular Genetics Deptment	Instituto Investigación Oncológica, Spain
2001-2002	Visiting Scientist	Hospital for Sick Children, Toronto, Canada
2002-2012	Program Coordinator Genes and Disease	Center Genomic Regulation, Spain
2002-2012	Senior Scientist Genetic Causes of Disease	Center Genomic Regulation, Spain
2002-	Associate Professor	Pompeu Fabra University, Spain
2013-	Director Genomics Medicine Unit	Institut Universitary Quirón-Dexeus, Spain

3. Publications (10 selected publications, last 3 years, 2011-2013)

1. Molina-Vila MA, Nabau-Moretó N, Tornador C, Sabnis AJ, Rosell R, Estivill X, Bivona T, Marino-Buslje C. Activating Mutations Cluster in the "Molecular Brake" Regions of Protein Kinases and do not Associate with Conserved or Catalytic Residues. **Hum Mutat**. 2013 Dec 9.
2. Lappalainen T, Sammeth M, Friedländer MR, …, Estivill X, Dermitzakis ET; Geuvadis Consortium. Transcriptome and genome sequencing uncovers functional variation in humans. **Nature**. 2013 Sep 26;501(7468):506-11.
3. 't Hoen PA, Friedländer MR, …, Estivill X, Syvänen AC, Dermitzakis ET, Lappalainen T. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. **Nat Biotechnol**. 2013 Nov;31(11):1015-22.
4. Bassaganyas L, …, Estivill X. Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis. **Leukemia**. 2013 Dec;27(12):2376-9.
5. Horikoshi M, …, Estivill X, …; Early Growth Genetics (EGG) Consortium. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. **Nat Genet**. 2013 Jan;45(1):76-82.
6. Tsoi LC, …, Estivill X, …, Trembath RC. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. **Nat Genet**. 2012 Dec;44(12):1341-8.
7. Taal HR, et al. (includes Estivill X); Early Growth Genetics Consortium. (2012) Common variants at 12q15 and 12q24 are associated with infant head circumference. **Nat Genet** 44(5): 532-8.
8. Louis-Dit-Picard H, Barc J, Trujillano D, …, Estivill X, Froguel P, Hadchouel J, Schott JJ, Jeunemaitre X. (2012) KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. **Nat Genet** 44(4): 456-60.
9. Quesada V, …, (includes Estivill X). (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. **Nat Genet** 44(1): 47-52.
10. Puente XS, …, Estivill X, Montserrat E, López-Otín C, Campo E. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. **Nature** 2011 475(7354): 101-5.

Stephan Ossowski Curriculum Vitae Sketch

Name: **Stephan Ossowski** Position title: Junior Group Leader Genomic & Epigenomic Variation in Disease Bioinformatics and Genomics Programme
Birth date: 19/01/1977 Citizenship: German Gender (M/F): M
Office address: Doctor Aiguader 88, 08003 Barcelona Office telephone: +3493 316 0100
E-mail: stephan.ossowski@crg.eu

1. Education / Training

Degree	Institution	Year(s)
Master in Computer Science	Eberhard Karls Universität Tübingen Germany	2004
PhD in Computer Science	Eberhard Karls Universität Tübingen Germany	2010

2. Professional academic positions

Dates (from-until)	Position	Department & Institution
1999-2003	Lecturing tutor in: Algorithms, Bioinformatics,	University of Tübingen
2003-2004	Undergraduate Assistant	University of Tübingen
2004-2004	Research Fellow	Southwest Medical Center
2005-2010	Graduate student	Max Planck Institute
2010-2011	Guest scientist (part-time)	Massachusetts Institute of Technology (MIT), Boston
2010-ongoing	Group Leader	Center for Genomic Regulation (CRG), Barcelona, Spain

3. Publications (10 selected publications, last 2 years, 2012-2013)

1. The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. Wijnker E, …, Ossowski S, Weigel D, Koornneef M, Keurentjes JJ, Schneeberger K. **Elife**. 2(0): e01426. 2013.
2. A genomic-scale artificial microRNA library as a tool to investigate the functionally redundant gene space in Arabidopsis. Hauser F, Chen W, Deinlein U, Chang K, Ossowski S, Fitz J, Hannon GJ, Schroeder **JI.Plant Cell**. 25(8): 2848-63. 2013.
3. Accurate molecular diagnosis of phenylketonuria and tetrahydrobiopterin-deficient hyperphenylalaninemias using high-throughput targeted sequencing.Trujillano D, …, Ossowski S, Armengol L, Cornejo V, Desviat LR, Ugarte M, Estivill X. **Eur J Hum Genet.** [Epub ahead of print] 2013.
4. Extensive sequence analysis of CFTR, SCNN1A, SCNN1B, SCNN1G and SERPINA1 suggests an oligogenic basis for cystic fibrosis-like phenotypes. Ramos M, Trujillano D, Olivar R, Sotillo F, Ossowski S, …, Estivill X, Casals T. **Clin Genet**. [Epub ahead of print] 2013.
5. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Koenig D, …, Ossowski S, Lanz C, Xiong G, Taylor-Teeples M, Brady SM, Pauly M, Weigel D, Usadel B, Fernie AR, Peng J, Sinha NR, Maloof JN. **Proc Natl Acad Sci USA**. 110(28): E2655-62. 2013.
6. PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. Escaramís G, Tornador C, Bassaganyas L, Rabionet R, Tubio JM, Martínez-Fundichely A, Cáceres M, Gut M, Ossowski S, Estivill X**. PLoS One**. 8(5): e63377. 2013.
7. Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. Trujillano D, Ramos MD, González J, Tornador C, Sotillo F, Escaramis G, **Ossowski S**, Armengol L, Casals T, Estivill X. **J Med Genet**. 50(7): 455-62. 2013
8. Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis. Bassaganyas L, …, Ossowski S, López-Otín C, Campo E, Estivill X. **Leukemia**. 27(12): 2376-9. 2013
9. The cis-regulatory code of Hox function in Drosophila. Sorge S*, Ha N*, Polychronidou M.*, Friedrich J*, Bezdan D*, Kaspar P, Schaefer MH, Ossowski S, Henz SR, Mundorf J, Ratzer J, Papagiannouli F and Lohmann I. **EMBO J.** 31(15): 3323-33. 2012.
10. KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. Louis-Dit-Picard H, .., Ossowski S, Caulfield M; International Consortium for Blood Pressure (ICBP), Bruneval P, Estivill X, Froguel P, Hadchouel J, Schott JJ, Jeunemaitre X. **Nat Genet**. 44(4): 456-60. 2012.

Aparna Prasad CV Sketch

Name: Aparna Prasad, PhD
Email: aparna.prasad@crg.eu Phone: +34-673385665 (cell)
Postdoctoral research fellow
Centre for Genomic Regulation (CRG) Room 529.02, Dr. Aiguader, 88
08003 Barcelona, Spain

Education and Research Experience:

1. *November 2012 – To date:* ***Postdoctoral Research Fellow*** *in CLL Genetics*
   Institution: **Centre for Genomic Regulation, Barcelona, Spain**

2. *April 2009 – October 2012:* ***Postdoctoral Research Fellow*** *in Autism Genetics*
   Institution: **The Hospital for Sick Children, Toronto, Canada**

3. *Sep 2004 – July 2009:* ***PhD*** *in Bovine Genomics*
   Institution: Department of Agricultural, Food and Nutritional Science, **University of Alberta, Edmonton, Canada**

4. *Aug 2002- Sep 2004:* ***Research Assistant***
   Institution: **National Institute of Immunology (NII), New Delhi, India**

5. *July 2000- Aug 2002:* ***Masters*** *in Microbiology*
   Institution: **Barkatullah University, Bhopal, India**

Peer-Reviewed Publications:

1. **Prasad A.** et al. 2012. **G3: Genes, Genomes, Genetics** 2(12):1665-85.
2. **Prasad A.,** et al. 2008. **Animal Genetics** 39(6): 597-605.
3. **Prasad A.,** et al. 2007. **BMC Genomics** 8: 310.
4. Lionel A.C. et al. 2013. **Human Molecular Genetics** 22(10):2055-66
5. Sato D. et al. 2012. **American Journal of Human Genetics** 90(5):879-87.
6. Vaags A.K. et al. 2011. **American Journal of Human Genetics** 90(1):133-41.
7. Pinto D. et al. 2011. **Nature Biotechnology** 29 (6): 512-520.
8. Carter M.T. et al. 2011. **Clinical Genetics** 80(5):435-43.
9. Pinto D et al. 2010. **Nature** 466(7304):368-72.
10. Anney R. et al. 2010. **Human Molecular Genetics** 19(20):4072-82.
11. Murdoch B.M. et al. 2010. **BMC Genetics** 11: 20.
12. Kolbehdari D. et al. 2009. **Journal of Animal Breeding and Genetics** 126(3): 216-227.
13. Kolbehdari D. et al. 2008. **Journal of Dairy Science** 91: 2844-2856.
14. Amaral M.E.J. et al. 2008. **BMC Genomics** 9:631.
15. Jann O.C. et al. 2006. **BMC Genomics** 7: 283.
16. Bashamboo A. et al. 2005. **Molecular and Human Reproduction** 11: 117-127.
17. Rahman M.M. et al. 2004. **DNA and Cell Biology** 23: 561-571.

Oliver Drechsel CV Sketch

Name: **Oliver Drechsel**, PhD
Position: Postdoctoral Research Fellow
Date of Birth: 01.11.1978
Nationality: German
Office Address: Doctor Aiguader 88, ES-08003 Barcelona
E-Mail: oliver.drechsel@crg.eu
Phone: +34 933160203

Education and Research Experience:

Since September 2011: Postdoctoral Research Fellow in "Genomic and Epigenomic variants in Disease" at Center for Genomic Research (CRG), Barcelona, Spain

January 2011 – July 2011: Postdoctoral Research Fellow in "Integrative carbon biology" at Max-Planck-Institute for Molecular Plant Physiology (MPI-MP), Potsdam, Germany

April 2009 – December 2010: Postdoctoral Research Fellow in "Organellar biology, biotechnology and molecular ecophysiology" at MPI-MP

March 2005 – April 2009: PhD in "Organellar biology, biotechnology and molecular ecophysiology" at MPI-MP

October 2003 – September 2004: Master in "Plant – Environment Interaction" at MPI-MP

October 2002 – September 2003: Student Worker in "Plant – Environment Interaction" at MPI-MP

August 2000 – July 2002: Student Worker in "Biotechnology" at the University of Potsdam, Potsdam, Germany

Publications

L. Bassaganyas, S. Bea, G. Escaramis, C. Tornador, I. Salaverria, L. Zapata, O. Drechsel, P.G. Ferreira, B. Rodriguez-Santiago, J.M. Tubio, A. Navarro, D. Martin-Garcia, C. Lopez, A. Martinez-Trillos, A. Lopez-Guillermo, M. Gut, S. Ossowski, C. Lopez-Otin, E. Campo, X. Estivill
Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis,
Leukemia 2013 online publication

M. Lohse*, O. Drechsel*, S. Kahlau, R. Bock
OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets
Nucleic Acids Research 41(Web Server Issue). 2013

O. Drechsel, R. Bock
Selection of Shine-Dalgarno sequences in plastids",
Nucleic Acid Research 39(4) 1427-38. 2011

M. Walter, K. Piepenburg, M.A. Schöttler, K. Petersen, S. Kahlau, N. Tiller, O. Drechsel, J. Kudla, R. Bock
Knockout of the plastid RNase E leads to chloroplast ribosome deficiency caused by defective RNA processing,
Plant Journal 64(5) 851-63. 2010

M. Lohse, O. Drechsel, Ralph Bock
OrganellarGenomeDRAW (OGDRAW) – a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes,
Current Genetics 52(5-6) 267-74. 2007

F. Sellrie, J.A. Schenk, O. Behrsing, O. Drechsel and B. Micheel
Cloning and Characterization of a Single Chain Antibody to Glucose Oxidase from a Murine Hybridoma,
Journal of Biochemistry and Molecular Biology 40(6) 875-880. 2007

*equal contribution

Manolis Kogevinas Curriculum Vitae Sketch

Name: **Manolis Kogevinas** Position title: Joint Scientific Director and Researcher
Office address: CREAL. Doctor Aiguader 88, 08003 Barcelona Office telephone: +3493 214 73 30
E-mail: kogevinas@creal.cat

1. Education / Training

Degree	Institution	Year(s)
PhD, Medicine	University of London	1989
PhD, Medicine	University of Athens	1984
Master of Science in Epidemiology	University of London	1984
MD, Medicine	University of Athens	1982

2. Professional academic positions

Dates (from-until)	**Position**	**Department & Institution**
2006 - Present	Co-Director	Center for Research in Environmental
2009 - Present	Professor (part time)	National School of Public Health, Athens,
2000 – Present	Senior Researcher	Institut Municipal d'Investigació Mèdica
2005 – 2009	Professor (part time)	Medical School, University of Crete, Greece
1995 – 2003	Associate professor	Universitat Autònoma de Barcelona
1989 – 1994	Epidemiologist/Technical Officer	International Agency Research Cancer
1987	Temporal "Clinical lecture"	University College London
1985 – 1986	Course tutor	Open University, London Region
1985	Specialization in radiotherapy	Metaxa Hospital, Pireaus, Greece
1982 - 1983	Specialization in radiotherapy	Gral. Laico Hospital, Athens, Greece

3. Publications (the last ones)

1. Siroux V, González JR, Bouzigon E, Curjuric I, Boudier A, Imboden M, Anto JM, Gut I, Jarvis D, Lathrop M, Omenaas ER, Pin I, Wjst M, Demenais F, Probst-Hensch N, Kogevinas M, Kauffmann F. Genetic heterogeneity of asthma phenotypes identified by a clustering approach. **Eur Respir J**. 2013 Dec 5. [Epub ahead of print].
2. Balbás-Martínez C, Sagrera A, Carrillo-de-Santa-Pau E, Earl J, Márquez M, Vazquez M, Lapi E, Castro-Giner F, Beltran S, Bayés M, Carrato A, Cigudosa JC, Domínguez O, Gut M, Herranz J, Juanpere N, Kogevinas M, Langa X, López-Knowles E, Lorente JA, Lloreta J, Pisano DG, Richart L, Rico D, Salgado RN, Tardón A, Chanock S, Heath S, Valencia A, Losada A, Gut I, Malats N, Real FX. Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. **Nat Genet**. 2013 Dec;45(12):1464-9. doi: 10.1038/ng.2799. Epub 2013 Oct 13.
3. Stayner LT, Pedersen M, Patelarou E, Decordier I, Vande Loock K, Chatzi L, Espinosa A, Fthenou E, Nieuwenhuijsen MJ, Gracia-Lavedan E, Stephanou EG, Kirsch-Volders M, Kogevinas M. Exposure to Brominated Trihalomethanes in Water During Pregnancy and Micronuclei Frequency in Maternal and Cord Blood Lymphocytes. **Environ Health Perspect**. 2013 Nov 1. [Epub ahead of print].
4. Leventakou V, Roumeliotaki T, Koutra K, Vassilaki M, Mantzouranis E, Bitsios P, Kogevinas M, Chatzi L. Breastfeeding duration and cognitive, language and motor development at 18 months of age: Rhea mother-child cohort in Crete, Greece. **J Epidemiol Community Health**. 2013 Dec 13. doi: 10.1136/jech-2013-202500. [Epub ahead of print].
5. Leventakou V, Roumeliotaki T, Martinez D, Barros H, Brantsaeter AL, Casas M, Charles MA, Cordier S, Eggesbø M, van Eijsden M, Forastiere F, Gehring U, Govarts E, Halldórsson TI, Hanke W, Haugen M, Heppe DH, Heude B, Inskip HM, Jaddoe VW, Jansen M, Kelleher C, Meltzer HM, Merletti F, Moltó-Puigmartí C, Mommers M, Murcia M, Oliveira A, Olsen SF, Pele F, Polanska K, Porta D, Richiardi L, Robinson SM, Stigum H, Strøm M, Sunyer J, Thijs C, Viljoen K, Vrijkotte TG, Wijga AH, Kogevinas M, Vrijheid M, Chatzi L. Fish intake during pregnancy, fetal growth, and gestational length in 19 European birth cohort studies. **Am J Clin Nutr**. 2013 Dec 11. [Epub ahead of print].
6. Cohen G, Vardavas C, Patelarou E, Kogevinas M, Katz-Salamon M. Adverse circulatory effects of passive smoking during infancy: surprisingly strong, manifest early, easily avoided. **Acta Paediatr**. 2013 Dec 12. doi: 10.1111/apa.12538. [Epub ahead of print].
7. Papantoniou K, Kogevinas M. Shift work and breast cancer: do we need more evidence and what should this be? **Occup Environ Med**. 2013 Dec;70(12):825-6. doi: 10.1136/oemed-2013-101630. Epub 2013 Oct 9.
8. Fucic A, Katic J, Fthenou E, Kogevinas M, Plavec D, Koppe J, Batinic D, Chalkiadaki G, Chatzi L, Lasan R, Kleinjans J, Kirsch-Volders M. Increased frequency of micronuclei in mononucleated lymphocytes and cytome analysis in healthy newborns as an early warning biomarkers of possible future health risks. **Reprod Toxicol**. 2013 Dec;42:110-5. doi: 10.1016/j.reprotox.2013.08.004. Epub 2013 Aug 28.

Gemma Castaño-Vinyals CV Sketch

Name: Gemma Castaño-Vinyals, PhD
Email: gcastano@creal.cat Office phone: 93 214 7303
Centre de Recerca en Epidemiologia Ambiental (CREAL)
Doctor Aiguader, 88 1st floor; 08003 Barcelona; Spain

EDUCATION

2007. PhD in Health and Life Sciences at the Universitat Pompeu Fabra, Spain.
2003. MSc, Department of Health and Experimental Sciences, Universitat Pompeu Fabra, Spain.
2000. Bachelor in Environmental Sciences, Universitat Autònoma de Barcelona, Spain.

PROFESSIONAL POSITIONS

2008-current: Project Manager/Post-doctoral research fellow MCC-Spain study at CREAL.
2008 – 2008: Research Technician Hi-Wate Study at CREAL.
2005 – 2007: Research technician at the Municipal Institue of Medical Research (IMIM), Barcelona, Spain.
2000 - 2005: PhD student at the IMIM.
1999 – 2000: undergraduate fellowship at IMIM.

PUBLICATIONS (selected from the last 3 years)

- Zock JP, …, **Castaño-Vinyals G**, Antó JM, Barbera JA. Evaluation of the persistence of functional and biological respiratory health effects in clean-up workers 6years after the Prestige oil spill. Environ Int 2014; 62: 72-77.
- Monyarch G, …, **Castaño-Vinyals G**, …, Fuster C. Chromosomal bands affected by acute oil exposure and DNA repair errors. PLoS ONE 2013; 8(11): e81276.
- **Castaño-Vinyals G**, Carrasco E, Lorente JA, Sabaté A, Cirac J, Pollán M, Kogevinas M. Anogenital distance and the risk of prostate cancer. BJU International 2012; 110(11b):E707-710.
- Villanueva CM, **Castaño-Vinyals G**, …, Kogevinas M. Concentrations and correlations of disinfection by-products in municipal drinking water in Spain. Environ Res 2012; 114: 1-11.
- Perea MD, **Castaño-Vinyals G**, …, Sala M. Prácticas de cribado de cáncer y estilos de vida asociados en la población de controles del estudio español multi-caso control (MCC-Spain). Gac Sanit 2012; 26: 301-310.
- Zock JP, …, **Castaño-Vinyals G**, Antó JM, Barberà JA. Persistent respiratory symptoms in clean-up workers 5 years after the Prestige oil spill. Occup Environ Med 2012; 69:508-513.
- **Castaño-Vinyals G**, …, Villanueva CM. Participation rates of population controls in a case-control study of colorectal cancer using two recruitment methods. Gac Sanit 2011; 25(5):353-356.
- **Castaño-Vinyals G**, Cantor KP, Villanueva CM, Tardon A, Garcia-Closas R, Serra C, Carrato A, Malats N, Rothman N, Silverman D, Kogevinas M. Socioeconomic status and exposure to disinfection byproducts in Spain. Environ Health 2011; 10:18.
- Stevens RG, …, **Castaño-Vinyals G**, …, Straif K. Considerations of circadian impact for defining 'shift work' in cancer studies: IARC Working Group Report. Occup Environ Med 2011;68(2):154-62.
- Cantor KP, …, **Castaño-Vinyals G**, Samanic C, Rothman N, Kogevinas M. Polymorphisms in GSTT1, GSTZ1, and CYP2E1, Disinfection Byproducts, and Risk of Bladder Cancer in Spain. Environ Health Perspect 2010; 118 (11):1545-50.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Multidimensional data visualization of ICGC Pan-Cancer results

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Nuria Lopez-Bigas, ICREA and University Pompeu Fabra, Barcelona

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Jordi Deu-Pons and David Tamborero, University Pompeu Fabra, Barcelona	

Background and preliminary data

The ICGC Pan-Cancer project will collate highly valuable data and will produce a high number of important results for the cancer research community. The visual exploration of these type of data by cancer researchers is often hampered by the high complexity and amount of information generated. To speed up research using ICGC Pan-Cancer data it is crucial to minimize the barrier currently existing between the complex cancer genomic datasets and the effective exploration of these data by cancer researchers.

We have worked on this topic in the last years in two directions:

IntOGen web resource. On one side we have created a web-based resource to provide the results of the analysis of the somatic mutations detected in thousands of tumor genomes. We create the web of IntOGen-mutations using Onexus (http://www.onexus.org), a system developed in our lab to manage, query and visualize large and complex datasets. We plan to update IntOGen-mutations with information obtained from the ICGC Pan-Cancer project.

Gitools interactive heat-maps. Heat-maps are useful and intuitive graphical representations frequently used to describe transcriptomics and genomics data stored in the form of matrices. Usually heat-maps are represented as static images. Nevertheless, the complexity of oncogenomics data and the variety of questions to be addressed make static heat-maps unsuitable for efficient knowledge extraction. Instead, we have proposed the use of interactive heat-maps to explore multidimensional cancer genomics data enabling the user to explore the data interactively (Perez-Llamas and Lopez-Bigas 2011, Schroeder et al., Genome Medicine 2013, http://www.gitools.org/datasets). We have recently prepared multidimensional Pan-Cancer data from TCGA to be browsed in Gitools (see description and image in next page). We plan to prepare ICGC Pan-Cancer data to be browsed with Gitools interactive heat-maps.

Timelines & resources dedicated to project

Timeline:

- December 2013 – November 2014: Working in Gitools 3.0: New version of the software including new capabilities and improvements in user interface.
- November 2014 to December 2014: Reformatting ICGC Pan-Cancer data to prepare interactive heat-maps to be explored with Gitools. Preparing IntOGen web browser to host ICGC Pan-Cancer data.

Resources: One postdoc (David), who will help preparing the data and one software engineer (Jordi), who is in charge of the development of Gitools 3.0.

Research proposal

IntOGen-mutations

We plan to provide the results of analyzing ICGC Pan-Cancer somatic mutations data through our IntOGen portal.

Exploring ICGC Pan-Cancer results using Gitools Interactive Heatmaps

We plan to prepare multidimensional heat-maps from ICGC Pan-Cancer projects to be browsed using Gitools interactive heat-maps. We have recently prepared Gitools interactive heat-maps for exploring TCGA Pan-Cancer data, in which tumor samples are shown in columns and genes in rows. Each cell contains multiple layers of information, including mutation, copy number status and expression values. Samples are annotated with clinical information, which can be displayed as colored headers on top of the columns heat-map (see figure below). Similarly genes are annotated with multiple information, such as signals of positive selection detected pointing to driver genes, mutation frequency, etc.

Gitools has three unique features that make it especially suitable to knowledge discovery from ICGC Pan-Cancer data. First, it supports the work with big data matrices comprising information from tens of thousands of rows (genes or non-coding functional elements) and thousands of tumor samples through the use of desktop computers. Second, it allows the user to interact with heat-maps, by moving, sorting, filtering, clustering, hiding and adding annotations to columns and rows, as well as doing basic analyses like correlations, group comparisons, aggregations and sample level enrichment analysis. Third, via the use of multidimensional matrices, it allows the user to uncover correlations between different cancer alterations, such as the impact of mutations or copy number alterations on gene expression.

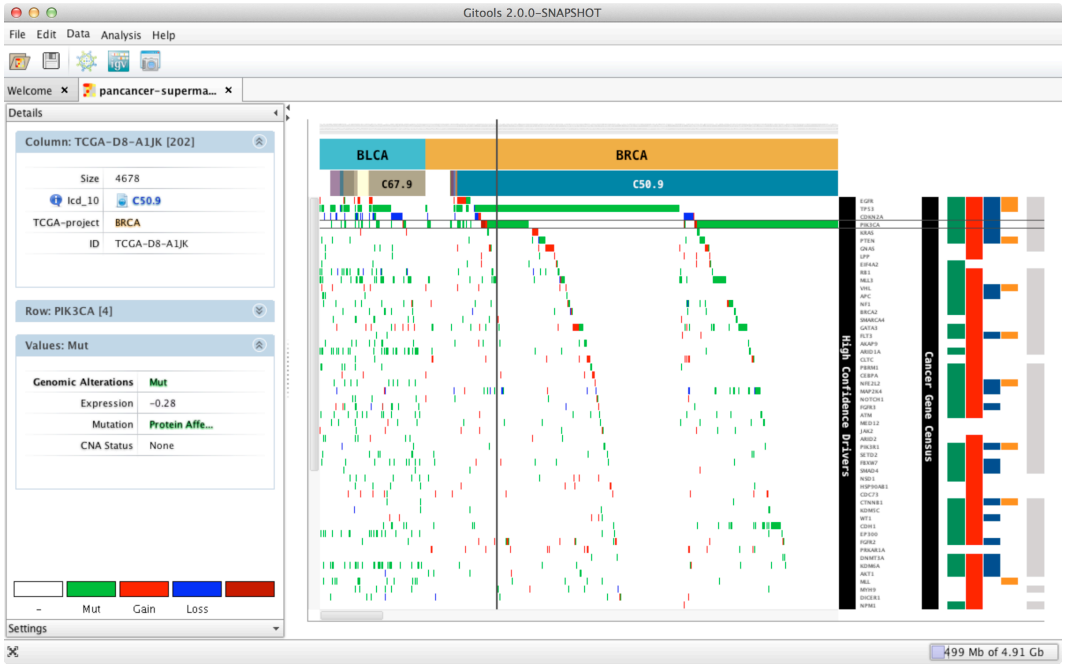


Figure. Example of one Gitools interactive heat-map with TCGA Pan-Cancer information. Columns represent samples and rows genes. Each cell is associated to multiple values that can be seen when the cell is selected in the details panel in the left. Interactive capabilities and analysis options allows us to explore the data and extract interesting knowledge from it.

Legacy plans

Gitools is an open source software (http://github.com/gitools), with extensive documentation. Onexus is also open source (http://github.com/onexus).

Nuria Lopez-Bigas, PhD - Curriculum Vitae

Biomedical Genomics Group, Research Unit on Biomedical Informatics, Experimental and Health Science Department
University Pompeu Fabra, 08003 - Barcelona. **T** 933260507**. F** 933160550
nuria.lopez@upf.edu - http://bg.upf.edu

Personal statement

I am an ICREA Research professor at the University Pompeu Fabra. I lead the Biomedical Genomics group (http://bg.upf.edu) since April 2006. My group works on Computational Cancer Genomics. We have developed orginal methods for the identification of cancer drivers and tools for the analysis and visualization of cancer genomics data.

Relevant achievements

- **IntOGen**: Platform that summarizes mutations, genes and pathways involved in cancer across thousands of tumor genomes/exomes from different cancer sites. http://www.intogen.org
- Methods to identify **cancer drivers**: OncodriveFM, OncodriveCLUST and OncodriveCIS
- Methods to assess the **functional impact** of coding variants: TransFIC and Condel
- **Gitools**: Visualization and analysis of genomics data using interactive heat-maps
- Participation in **ICGC**: Co-leading (with Lincoln Stein) the ICGC Mutation Consequences and Pathways Subgroup
- Participation in **TCGA** Pan-Cancer project

Selected publications (from 74 peer-reviewed publications)

Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, Gouin A, Djebali S, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M, Clot G, Aymerich M, Rozman M, Kulis M , Tamborero D, Papasaikas P, Blanc J, Gut M, Gut I, Puente XS, Pisano DG, Martin- Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, Guigo R. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. **Genome Research**. 2013 Nov 21

The Cancer Genome Atlas Pan-Cancer Analysis Project. The Cancer Genome Atlas Research Network (including Abel Gonzalez-Perez, David Tamborero and Nuria Lopez-Bigas). **Nature Genetics** 2013. 45, 1113–1120.

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N. Comprehensive identification of mutational cancer driver genes across 12 tumor types. **Scientific Reports**, 2013. 3:2650 doi:10.1038/srep02650

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. **Nature Methods**, 2013. doi:10.1038/nmeth.2642

Abel Gonzalez-Perez#, Ville Mustonen#, Boris Reva#, Graham R.S. Ritchie#, Pau Creixell, Rachel Karchin, Miguel Vazquez, J. Lynn Fink, Karin S. Kassahn, John V. Pearson, Gary Bader, Paul C. Boutros, Lakshmi Muthuswamy, B.F. Francis Ouellette, Jüri Reimand, Rune Linding, Tatsuhiro Shibata, Alfonso Valencia, Adam Butler, Serge Dronov, Paul Flicek, Nick B. Shannon, Hannah Carter, Li Ding, Chris Sander, Josh M. Stuart, Lincoln Stein and Nuria Lopez-Bigas for the ICGC Mutation Pathways and Consequences Subgroup. Computational approaches to identify functional genetic variants in cancer genomes. **Nature Methods**. 2013 10, 723-72.

OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. **Bioinformatics**, 2013; 29:2238-44.

Schroeder MP, Gonzalez-Perez A and Lopez-Bigas N. Visualizing multidimensional cancer genomics data. **Genome Medicine**. 2013, 5:9

Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. **Nucleic Acids Res**. 2012 40(21):e169.

Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, and Lopez-Bigas N. IntOGen: integration and data mining of multidimensional oncogenomic data. **Nat Methods** 2010 7(2):92-3.

Full list of publications can be found at:

Google Scholar: http://scholar.google.com/citations?user=I_QcK7oAAAAJ&hl=en&oi=ao

PubMed: http://www.ncbi.nlm.nih.gov/pubmed?term=Lopez-Bigas%5BAuthor%5D

Jordi Deu-Pons

I am a software engineer currently working on software development for the analysis and visualization of genomics data at the group of Nuria Lopez-Bigas at the Universitat Pompeu Fabra, Barcelona.

Education

Year 2007: Postgraduate course in Bioinfomatics, Genomics and Structural Biology (Universitat Oberta de Catalunya)
Year 2005: Bachelor of Science in Informatics Technical Engineering (Universitat Autonoma de Barcelona)

Developed Software

jHeatmap (http://jheatmap.github.io/jheatmap)
Description: Interactive web heatmap viewer (*contribution*: Main architect and developer)
Technology: HTML5, javascript, jQuery.

Onexus (http://www.onexus.org)
Description: modular framework to manage the complete life cycle of data analyses, that allows creating websites to browse interactively the result datasets. Example website http://www.intogen.org. (*contribution*: Main architect and developer)
Technology: Java, OSGi, blueprint, Apache Karaf (Apache Felix)

Gitools (http://www.gitools.org)
Description: framework for analysis and visualization of genomic data using interactive heatmaps (*contribution*: version 2.0 development)
Technology: Java, Swing

Expertise

- Software development. Deep knowledge and long experience developing server side and client side software with Java related technololgies (J2EE, OSGi, Hibernate, Spring, Swing, SWT, GWT, Karaf, Wicket, JSP, Blueprint, Guice, Maven, Ant, JAXB)
- Data management. Experience managing big databases (MySQL, SQL server).
- Web development. Experience developing websites integrating different technologies (HTML5, CSS3, PHP, ASP, AngularJS, jQuery)
- System administration. Experience working as a system administrator of networks using different operating systems like: RedHat, Fedora, CentOS, Debian, Ubuntu, Windows 2000 server, Windows 98, Windows XP.

Publications

Deu-Pons J, Schroeder MP and Lopez-Bigas N. **jHeatmap: an interactive heatmap viewer for the web.** Under review

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N. **Comprehensive identification of mutational cancer driver genes across 12 tumor types.** Scientific Reports, 2013. 3:2650 doi:10.1038/srep02650

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods, 2013. doi:10.1038/nmeth.2642

Abel Gonzalez-Perez, Jordi Deu-Pons, Nuria Lopez-Bigas. **Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation.** Genome Medicine 11/2012; 4(11):89.

Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, et al. **IntOGen: integration and data mining of multidimensional oncogenomic data**. Nature Methods. 2010 February;7(2):92–93.

David Tamborero Noguera

- Year 2012: **Master degree** in **Bioinformatics for Health Sciences** *(Universitat Pompeu Fabra)*. Final project *"Functional impact assessment of Somatic Aberrations to retrieve Genes involved in Cancer"*.

- Year 2009 : **PhD** in **Biopatology of Medicine** (*Facultat de Medicina, Universitat de Barcelona).* Doctoral thesis "*Atrial Fibrillation Radiofrequency Ablation: Techniques, Complications and Results*" qualified as Magna cum Laude.

- Year 2003: **Bachelor of Science** in **Telecommunications Superior Engineering** (*Escola d'Enginyeria la Salle Barcelona, Universitat Ramon Llull*). Final project "*Electroanatomic Navigation System for Radiofrequency Ablation Procedures*" qualified as Distinction.

- Year 2000: **Bachelor of Science** in **Electronics Technical Engineering** (*Escola d'Enginyeria la Salle Barcelona, Universitat Ramon Llull*). Final project *"Teleassistance System based on Lonworks Technology*" qualified as Distinction.

Currently working at the **Biomedical Genomics Group**, *Universitat Pompeu Fabra (Barcelona)*, in the analysis of next-generation sequencing data of cancer. According to Google Scholar, I have 100 manuscripts published in several peer-review journals accumulating more than 1,100 citations, with an *h-index* of 19 and *i-index* of 20. The following manuscripts are the latest ones related with the topic of the present project:

- Transcriptome characterization by RNA sequencing identifies a major molecular and clnical subdivision in chronic lymphocytic leukemia. *Ferreira P, Jares P, Rico D, Gómez G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles G, Monlong J, Johnson R, Quesada V, Djebali S, Papasaikas P, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M , Clot G, Aymerich M, Rozman M, Kulis M , **Tamborero D**, Gouin A, Blanc J, Gut M , Valcarcel J, Gut I, Puente X, Pisano D, Martin-Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, Guigó R.* **Genome Research 2013**, in press.

- The Cancer Genome Atlas Pan-Cancer analysis project. *Cancer Genome Atlas Research Network (including **Tamborero D**).* **Nat Genet 2013**; 45: 113-20.

- Comprehensive identification of mutational cancer driver genes across 12 tumor types. ***Tamborero D**, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N.* **Nat Sci Rep 2013**; 3: 2650.

- IntOGen-mutations identifies cancer drivers across tumor types. *Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, **Tamborero D**, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N.* **Nature Methods, 2013**; 10:1081-2.

- OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. ***Tamborero D**, Gonzalez-Perez A, Lopez-Bigas N.* **Bioinformatics, 2013**; 29: 2238-44.

- OncodriveCIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. ***Tamborero D,** Lopez-Bigas N, Gonzalez-Perez A.* **PLoSONE, 2013**: 8, e55489.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Whole-genome landscape of cancer driver mutations

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Nuria Lopez-Bigas, ICREA and University Pompeu Fabra, Barcelona and Roderic Guigó, Center for Regulatory Genomics, Barcelona

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Abel Gonzalez-Perez and David Tamborero, University Pompeu Fabra, Barcelona	**Rory Johnson, Center for Regulatory Genomics, Barcelona**

Background and preliminary data

One of the major challenges in cancer genomics is to identify which of the large number of somatic mutations occurring in a tumor are driving the tumorigenic process. In the last years we have addressed this issue by developing computational methods able to detect complementary signals of positive selection in the pattern of somatic mutations in protein-coding genes (OncodriveFM and OncodriveCLUST). We have integrated these methods within a framework named IntOGen-mutations (http://www.intogen.org/mutations) in which data collected from 31 independent re-sequencing projects (including ICGC, TCGA and others) accounting for 4621 samples have already been analysed. The analysis pipeline is also available to users. Therefore, this resource can be used both to explore putative drivers from the already available cancer projects as well as to analyse novel sample sets provided by the user.

We have recently participated in the TCGA Pan-Cancer initiative, a project that -taking advantage of the homogeneous analysis of the data generated- was able to detect putative driver genes across tissue boundaries in 3,205 samples from 12 cancer types. For this work, we have combined the results of several complementary methods to construct a list of 291 high-confidence driver genes acting in one or more of these tumor types, some of which are novel candidates not previously reported as implicated in cancer (Tamborero et al., 2013).

For the ICGC WG pan-cancer analysis we propose to study the whole-genome landscape of cancer by adapting the same principles to non-coding regions. In other words, we propose to develop methods that detect both coding and non-coding elements with clear signals of positive selection such as mutational frequency, functional impact bias, and regional clustering.

Timelines & resources dedicated to project

- December 2013 – November 2014: Development and validation of novel methodologies to identify signals of positive selection in non-coding functional elements. The methods will be developed and tested in the first instance using already available whole-genome sequencing data from individual projects (Alexandrov et al.).
- December 2013 – November 2014: In parallel, we will adapt our existing IntOGen-mutations pipeline (www.intogen.org/mutations) to incorporate new methods to detect drivers amongst both coding and non-coding regions. Once adapted, the pipeline will manage the most complete collection of methods to detect signals of positive selection across tumor samples. It will thus be prepared to fully uncover the landscape of mutational events driving cancer across all tumor types analyzed within the ICGC pan-cancer initiative.
- November 2013 to December 2014: Application of the methods to ~2000 WGs and ~8000 Exomes to fully uncover this landscape of mutational events driving cancer across all tumor types.

Resources: Three postdocs (two from Núria's lab and one from Roderic's lab) will be dedicated to the project. Both groups together possess the sufficient computational infrastructure to carry out the validation phase.

Research proposal

Identification of protein-coding driver mutations

We will use our pipeline to analyse the ~8000 exomes available for the ICGC Pan-Cancer project. We will refine this analysis by adding new versions of the methods aimed to identify signals of positive selection among protein-coding mutations. Furthermore, we will add new drivers detection methods to the pipeline to automatize the combinatorial approach we recently developed and tested as part of the TCGA pan-cancer. We will take advantage of the data quality provided by the ICGC to perform analyses aimed to detect drivers specific of certain tumor types as well as to disentangle common mutational patterns across several tissues in order to retrieve the most comprehensive list of cancer drivers acting in coding regions. We will also analyze the clonality of the drivers found in several tumor types, to differentiate between founder and late drivers.

Identification of non-coding driver mutations

We will develop novel methods to identify non-coding elements, both cis- and trans-acting which could drive tumorigenesis across different cancer types included within the ICGC pan-cancer. We will follow the same basic principles that we have successfully applied for protein-coding mutations, i.e.: a) mutations exhibiting signals of positive selection highlight events targeted by cancer, and b) the combination of methods based on complementary criteria allows to retrieve more comprehensive and reliable lists of putative drivers.

To do so, we will first define the non-coding elements of interest using the data provided by the ENCODE project. Thereafter, we will analyse the patterns of somatic mutations in these regions with the following rationales:

a) Regional Clustering of mutations. We plan to adapt our previously developed method OncodriveCLUST to the particular conditions of functional non-coding elements. The aim of this method will be to detect non-coding elements whose mutations tend to cluster in certain sites. This method will compute the expected distribution of mutations across these regions and will take into account, when available predictions on the distortion of non-coding RNAs 3D structure.

b) Functional impact bias. This method will detect non-coding regions accumulating functional mutations. It will follow the principles of OncodriveFM. We will first score the putative functional impact of mutations in several non-coding elements employing data such as conservation information and the allele frequency of observed germline variants taken from the 1000 genomes projects and the 2000 normal genomes included in the ICGC pan-cancer. The score will also incorporate particular features of each element. For instance, the functional impact of mutations in transcription factor binding sites will be evaluated through position weight matrices when available. We will also search for mutations falling in predicted splicing regulatory regions.

c) Expression impact bias. We plan to use the RNAseq information, when available to assess the impact of mutations in regulator elements on the RNA levels of the regulated genes. Here we will build upon the idea implemented in OncodriveCIS (Tamborero et al, Plos One 2012) method. We expect that regulatory mutations driving tumorigenesis will have higher impacts on the expression of regulated genes. We will use RNAseq data to test the effect of mutations on absolute RNA levels as well as individual splice variants.

d) Significant mutation frequency. We will assess the frequency of mutations in each non-coding element to identify elements deviating from the background rate. Here, we will take into account mutational signatures detected across tumor types, as well as covariates that may affect the background mutation rate. We will investigate whether we can discover cancer-driver mutations in ncRNA by exon-intron enrichment analysis.

Both protein coding and non-coding elements identified as putative drivers will be integrated within network analyses aimed at producing a global picture of the cellular pathways affected by tumorigenesis in each tumor type. To functionally test new candidate non-coding RNA oncogenes, we will investigate the possibility of performing loss of function experiments with short hairpin or CRISPRi technologies.

Legacy plans

We always make our code available through Github or Bitbucked (see for example https://bitbucket.org/bbglab and https://bitbucket.org/intogen), and try to thoroughly document our methods.

For IntOGen-mutations pipeline we are currently preparing a documented virtual machine image.

For the ICGC Pan-Cancer project we plan to release the code of the analysis pipeline created through Bitbucked and enable the use of the pipeline and replication of results by third parties through a documented virtual machine image.

Nuria Lopez-Bigas, PhD - Curriculum Vitae

Biomedical Genomics Group, Research Unit on Biomedical Informatics, Experimental and Health Science Department
University Pompeu Fabra, 08003 - Barcelona. **T** 933260507**. F** 933160550
nuria.lopez@upf.edu - http://bg.upf.edu

Personal statement

I am an ICREA Research professor at the University Pompeu Fabra. I lead the Biomedical Genomics group (http://bg.upf.edu) since April 2006. My group works on Computational Cancer Genomics. We have developed orginal methods for the identification of cancer drivers and tools for the analysis and visualization of cancer genomics data.

Relevant achievements

- **IntOGen**: Platform that summarizes mutations, genes and pathways involved in cancer across thousands of tumor genomes/exomes from different cancer sites. http://www.intogen.org
- Methods to identify **cancer drivers**: OncodriveFM, OncodriveCLUST and OncodriveCIS
- Methods to assess the **functional impact** of coding variants: TransFIC and Condel
- **Gitools**: Visualization and analysis of genomics data using interactive heat-maps
- Participation in **ICGC**: Co-leading (with Lincoln Stein) the ICGC Mutation Consequences and Pathways Subgroup
- Participation in **TCGA** Pan-Cancer project

Selected publications (from 74 peer-reviewed publications)

Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, Gouin A, Djebali S, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M, Clot G, Aymerich M, Rozman M, Kulis M , Tamborero D, Papasaikas P, Blanc J, Gut M, Gut I, Puente XS, Pisano DG, Martin- Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, Guigo R. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. **Genome Research**. 2013 Nov 21

The Cancer Genome Atlas Pan-Cancer Analysis Project. The Cancer Genome Atlas Research Network (including Abel Gonzalez-Perez, David Tamborero and Nuria Lopez-Bigas). **Nature Genetics** 2013. 45, 1113–1120.

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N. Comprehensive identification of mutational cancer driver genes across 12 tumor types. **Scientific Reports**, 2013. 3:2650 doi:10.1038/srep02650

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. **Nature Methods**, 2013. doi:10.1038/nmeth.2642

Abel Gonzalez-Perez#, Ville Mustonen#, Boris Reva#, Graham R.S. Ritchie#, Pau Creixell, Rachel Karchin, Miguel Vazquez, J. Lynn Fink, Karin S. Kassahn, John V. Pearson, Gary Bader, Paul C. Boutros, Lakshmi Muthuswamy, B.F. Francis Ouellette, Jüri Reimand, Rune Linding, Tatsuhiro Shibata, Alfonso Valencia, Adam Butler, Serge Dronov, Paul Flicek, Nick B. Shannon, Hannah Carter, Li Ding, Chris Sander, Josh M. Stuart, Lincoln Stein and Nuria Lopez-Bigas for the ICGC Mutation Pathways and Consequences Subgroup. Computational approaches to identify functional genetic variants in cancer genomes. **Nature Methods**. 2013 10, 723-72.

OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. **Bioinformatics**, 2013; 29:2238-44.

Schroeder MP, Gonzalez-Perez A and Lopez-Bigas N. Visualizing multidimensional cancer genomics data. **Genome Medicine**. 2013, 5:9

Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. **Nucleic Acids Res**. 2012 40(21):e169.

Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, and Lopez-Bigas N. IntOGen: integration and data mining of multidimensional oncogenomic data. **Nat Methods** 2010 7(2):92-3.

Full list of publications can be found at:

Google Scholar: http://scholar.google.com/citations?user=I_QcK7oAAAAJ&hl=en&oi=ao

PubMed: http://www.ncbi.nlm.nih.gov/pubmed?term=Lopez-Bigas%5BAuthor%5D

Dr. Roderic Guigó i Serra
Program Coordinator, Bioinformatics and Genomics, Centre de Regulació Genòmica
Professor of Bioinformatics (Catedràtic). Universitat Pompeu Fabra

Education and training. BS. Biology, 1981. Ms, Biology, 1983, phD Statistics, 1988; BS Phylosophy (96% completed)

Professional Experience. 1983-1988 Teaching Assistant, Universitat de Barcelona). 1988-1991 Postdoctoral Fellow Dana Farber Cancer Institute—Harvard University. 1991-1992 Postdoctoral Fellow Boston University, 1992-1993 Research Associate Los Alamos National Laboratory. 1994-1999 Assistant Professor Universitat de Barcelona. 1995 Visiting scientist, École Polytechnique Fédérale de Lausanne. 1998-2002 Consultant GlaxoSmithKline. 1999-2005 Associate Professor Universitat Pompeu Fabra.

Research Interest

The research in Guigo's lab focuses around gene prediction in eukarya. The group is interested in the mechanisms and evolution of the signals involved in the specification of the genes in the eukaryotic genome (splice sites, promotor elements, etc.). Furthermore, the group is interested in the development of software for the prediction of genes in DNA sequences.

Community activities

Roderic Guigó participates in numerous international conferences, often as invited speaker, or session chair, and he has been invited to deliver seminars at research institutes and universities around the world. His group has also produced highly regarded teaching materials, both electronic materials in the web (http://genome.imim.es/courses), and chapters in textbooks, dictionaries and encyclopedias.

Main active grants

- *Nodo de bioinformática y Genómica del Instituto Nacional de Bioinformática*. GENOMA ESPAÑA- Fundación para el Desarrollo de Investigación en Genómica y Proteómica. (2004-2014)
- *"A BLUEPRINT of Haematopoietic Epigenomes"*. European Comission (EU). BLUEPRINT_282510_RTD. (2011-2016)
- *"Investigacion del splicing mediante sequenciacion masivamente paralela del transcriptoma y del estatus de la cromatina"*. Ministerio de Economía y Competitividad. BIO2011-26205 (2012-2014)
- *"Uncovering and understanding RNA through Massively Parallel Sequencing"*. European Commission (EU). RNA-MAPS_294653 (2012-2017)
- *"Landscape of transcription in human and mouse"*. National Institutes of Health.(NIH), USA. 1U54HG007004. (2012-2016)
- *"ENCODE Data Analysis Center"*. National Institutes of Health.(NIH), USA. 1U41HG007000. (2012-2016).
- *"GENCODE –Comprehensive gene annotation for human and mouse"*. National Institutes of Health.(NIH), USA. U41HG007234. (2013-2017)
- *"GTEX – Methods for high-resolution analysis of genetic effects on gene expression"*. National Institutes of Health.(NIH), USA. R01MH101814. (2013-2016).

Five representative publications in the last years:

- Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, Guigó R. "Nucleosome positioning as a determinant of exon recognition". Nat Struct Mol Biol. 2009 Sep;16(9):996-1001.
- Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R.. "The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012 Sep;22(9):1775-89.
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs". Genome Res. 2012 Sep;22(9):1616-25.
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. "Landscape of transcription in human cells". Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233.
- Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, Gouin A, Djebali S, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M, Clot G, Aymerich M, Rozman M, Kulis M, Tamborero D, Papasaikas P, Blanc J, Gut M, Gut I, Puente XS, Pisano DG, Martin-Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, **Guigo R**. "Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia". Genome Res. 2013 Nov 21. [Epub ahead of print]

Abel Gonzalez-Perez

Education
2006: Ph.D., Bioinformatics, Universidad de la Habana (UH)
1998: B.Sc., Biochemistry, Universidad de la Habana (UH)

Professional experience
1998-1999: Scientific Trainee. Electrophysiology Laboratory, Center of Marine Bioactives
2000-2001: Research coordinator. HABITAT-CUBA, NGO
2002-2006: Ph.D. Student. National Bioinformatics Center
2006-2008: Head of Computational Genomics. National Bioinformatics Center
2008-2010: National Program Officer (50%) SDC (Swiss Development and Cooperation Agency) Office in Havana
2008-2010: Assistant Researcher (50%). National Bioinformatics Center
2010-: Postdoctoral Researcher. Research Group on Biomedical Informatics, University Pompeu Fabra, Barcelona Biomedical Research Park

Publications, presentations, patents and teaching
Eight first-author papers, three joint-first-author papers, three last-author papers, eight other papers
More than 20 posters in international meetings; four oral presentations in international papers
One patent
Lectures and seminars in eight international courses

Main scientific projects
2002-2005: Extraction of Biologically Relevant Information from Protein Families. Ministry of Science, Technology and the Environment, Cuba
2006-2009: Bioinformatics Tools to study Biological Networks. Ministry of Science, Technology and the Environment, Cuba
2010-2011: Automated System to manage and exploit genomics data for personalized medicine. Spanish Ministry of Science and Innovation (Innocash)
2011-2012: Integrating genomics data to study cancer. Spanish Ministry of Science and Innovation (Plan Nacional)
2010-: Mutations Consequence and Pathways subgroup (MUCOPA)Bioinformatics Working Group
International Cancer Genome Consortium (ICGC)
2012-2013: PAN-cancer analysis initiative. The Cancer Genome Atlas (TCGA)

Fellowships
Four fellowships in international research institutions

Awards
2012: Award to knowledge transfer to Genomed project. Social counsel of University Pompeu Fabra, Spain
2004, 2005, 2006: Three awards to best scientific result 2004, 2005, and 2006. Agency of Nuclear Energy and Advanced Technologies, Cuba
1995-1998: Six First Places and Six Second Places(Competition Exams). University of Havana, Cuba

David Tamborero Noguera

- Year 2012: **Master degree** in **Bioinformatics for Health Sciences** *(Universitat Pompeu Fabra)*. Final project *"Functional impact assessment of Somatic Aberrations to retrieve Genes involved in Cancer"*.

- Year 2009 : **PhD** in **Biopatology of Medicine** (*Facultat de Medicina, Universitat de Barcelona).* Doctoral thesis "*Atrial Fibrillation Radiofrequency Ablation: Techniques, Complications and Results*" qualified as Magna cum Laude.

- Year 2003: **Bachelor of Science** in **Telecommunications Superior Engineering** (*Escola d'Enginyeria la Salle Barcelona, Universitat Ramon Llull*). Final project "*Electroanatomic Navigation System for Radiofrequency Ablation Procedures*" qualified as Distinction.

- Year 2000: **Bachelor of Science** in **Electronics Technical Engineering** (*Escola d'Enginyeria la Salle Barcelona, Universitat Ramon Llull*). Final project *"Teleassistance System based on Lonworks Technology*" qualified as Distinction.

Currently working at the **Biomedical Genomics Group**, *Universitat Pompeu Fabra (Barcelona)*, in the analysis of next-generation sequencing data of cancer. According to Google Scholar, I have 100 manuscripts published in several peer-review journals accumulating more than 1,100 citations, with an *h-index* of 19 and *i-index* of 20. The following manuscripts are the latest ones related with the topic of the present project:

- Transcriptome characterization by RNA sequencing identifies a major molecular and clnical subdivision in chronic lymphocytic leukemia. *Ferreira P, Jares P, Rico D, Gómez G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles G, Monlong J, Johnson R, Quesada V, Djebali S, Papasaikas P, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M , Clot G, Aymerich M, Rozman M, Kulis M , **Tamborero D**, Gouin A, Blanc J, Gut M , Valcarcel J, Gut I, Puente X, Pisano D, Martin-Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, Guigó R.* **Genome Research 2013**, in press.

- The Cancer Genome Atlas Pan-Cancer analysis project. *Cancer Genome Atlas Research Network (including **Tamborero D**).* **Nat Genet 2013**; 45: 113-20.

- Comprehensive identification of mutational cancer driver genes across 12 tumor types. ***Tamborero D**, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N.* **Nat Sci Rep 2013**; 3: 2650.

- IntOGen-mutations identifies cancer drivers across tumor types. *Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, **Tamborero D**, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N.* **Nature Methods, 2013**; 10:1081-2.

- OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. ***Tamborero D**, Gonzalez-Perez A, Lopez-Bigas N.* **Bioinformatics, 2013**; 29: 2238-44.

- OncodriveCIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. ***Tamborero D,** Lopez-Bigas N, Gonzalez-Perez A.* **PLoSONE, 2013**: 8, e55489.

Dr. Roderic Guigó i Serra
Program Coordinator, Bioinformatics and Genomics, Centre de Regulació Genòmica
Professor of Bioinformatics (Catedràtic). Universitat Pompeu Fabra

Education and training. BS. Biology, 1981. Ms, Biology, 1983, phD Statistics, 1988; BS Phylosophy (96% completed)

Professional Experience. 1983-1988 Teaching Assistant, Universitat de Barcelona). 1988-1991 Postdoctoral Fellow Dana Farber Cancer Institute—Harvard University. 1991-1992 Postdoctoral Fellow Boston University, 1992-1993 Research Associate Los Alamos National Laboratory. 1994-1999 Assistant Professor Universitat de Barcelona. 1995 Visiting scientist, École Polytechnique Fédérale de Lausanne. 1998-2002 Consultant GlaxoSmithKline. 1999-2005 Associate Professor Universitat Pompeu Fabra.

Research Interest

The research in Guigo's lab focuses around gene prediction in eukarya. The group is interested in the mechanisms and evolution of the signals involved in the specification of the genes in the eukaryotic genome (splice sites, promotor elements, etc.). Furthermore, the group is interested in the development of software for the prediction of genes in DNA sequences.

Community activities

Roderic Guigó participates in numerous international conferences, often as invited speaker, or session chair, and he has been invited to deliver seminars at research institutes and universities around the world. His group has also produced highly regarded teaching materials, both electronic materials in the web (http://genome.imim.es/courses), and chapters in textbooks, dictionaries and encyclopedias.

Main active grants

- *Nodo de bioinformática y Genómica del Instituto Nacional de Bioinformática*. GENOMA ESPAÑA- Fundación para el Desarrollo de Investigación en Genómica y Proteómica. (2004-2014)
- *"A BLUEPRINT of Haematopoietic Epigenomes".* European Comission (EU). BLUEPRINT_282510_RTD. (2011-2016)
- *"Investigacion del splicing mediante sequenciacion masivamente paralela del transcriptoma y del estatus de la cromatina".* Ministerio de Economía y Competitividad. BIO2011-26205 (2012-2014)
- *"Uncovering and understanding RNA through Massively Parallel Sequencing".* European Commission (EU). RNA-MAPS_294653 (2012-2017)
- *"Landscape of transcription in human and mouse".* National Institutes of Health.(NIH), USA. 1U54HG007004. (2012-2016)
- *"ENCODE Data Analysis Center".* National Institutes of Health.(NIH), USA. 1U41HG007000. (2012-2016).
- *"GENCODE –Comprehensive gene annotation for human and mouse".* National Institutes of Health.(NIH), USA. U41HG007234. (2013-2017)
- *"GTEX – Methods for high-resolution analysis of genetic effects on gene expression"*. National Institutes of Health.(NIH), USA. R01MH101814. (2013-2016).

Five representative publications in the last years:

- Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, Guigó R. "Nucleosome positioning as a determinant of exon recognition". Nat Struct Mol Biol. 2009 Sep;16(9):996-1001.
- Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R.. "The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012 Sep;22(9):1775-89.
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs". Genome Res. 2012 Sep;22(9):1616-25.
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. "Landscape of transcription in human cells". Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233.
- Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, Gouin A, Djebali S, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M, Clot G, Aymerich M, Rozman M, Kulis M, Tamborero D, Papasaikas P, Blanc J, Gut M, Gut I, Puente XS, Pisano DG, Martin-Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, **Guigo R**. "Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia". Genome Res. 2013 Nov 21. [Epub ahead of print]

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)
Prof. Modesto Orozco Barcelona Supercomputing Center and Institute for Research in Biomedicine

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data
The group has been working for the last years in the CLL project, contributing to the detection of mutations and structural variants. The group has also a deep knowledge in rational drug-design projects and wishes to join both types of expertise in defining the universe of potential cancer-effective drugs. Our objective is to determine general, cancer-specific, subtype specific and individually tailored potential targets that can be tackled by small molecules. The long-term idea will be to transform academic genomic information derived from the project into "leads" for the treatment of cancer at different levels.

Timelines & resources dedicated to project
We will dedicate our local resources at the Barcelona Supercomputing Center, including all our local tools for target drugability validation, chemometry and rational drug design. We will rely on annotated mutational and expression data, so we believe we will be involved at advanced stages of the project .

Research proposal

We will analyze the potential drug-universe active against general cancer, particular cancer types and subtypes, moving to the individual level. We expect the most productive results in the two later levels of details, which have been not as widely studied. Potential targets will be: mutated proteins, proteins over-produced in specific cancer, proteins in pathways that have been altered in cancer. We do not plan to have a major role in genome analysis, so we will be dependent on data annotated by other groups. We will also benefit from the interaction with groups that surely will focus their interest in pathway analysis.

Once potential targets are detected we will discard those that are unlikely to be drugable, i.e. that can be targeted by small drugs. In a first approach we will select two types of targets: i) those with known drugs, or homologous to proteins with known ligands, ii) proteins with known structure or whose 3D structure can be reasonably modeled.

We will then use "in-house" chemometric tools as well as "in house" rational drug design strategies to determine potential drugs that can either affect a protein or an entire pathway. Our initial universe of drugs will be a locally-filtrated database with more than 1 million compounds, favoring those which have been already selected for clinical tests. In most cases we expect our strategy will recover drugs already in the market with other indications, but new chemical entities will surely emerge.

Validation on the ligand properties of suggested drugs, if necessary, can be done easily. Validation on patients in compassionate use would require close collaboration with clinical groups.

Legacy plans

As always in our group, all our software will be available as web applications and web services to the entire community, through the Barcelona Supercomputing Center and the Spanish National Institute of Bioinformatics.

CV Modesto Orozco (summary)

OROZCO Modesto
Born in Barcelona, Spain, October 12th 1962.
Barcelona Superomputing Center and Institute for Research in Biomedicine
e-mail: modesto.orozco@bsc.es; modesto.orozco@irbbarcelona.org

Appointments

Pre-Doctoral research fellow of the Spanish MEC. Departament de Bioquímica. Universitat de Barcelona. 1987-1989. Assistant Professor of Biochemistry. Departament de Bioquímica. Universitat de Barcelona. 1989-1990. Professor of Biochemistry and Molecular Biology. Departament de Bioquímica. Universitat de Barcelona. 1991-2001. Invited Scientist. Department of Chemistry. Yale University 1991-1993. Full Professor of Biochemistry and Molecular Biology. Departament de Bioquímica. Universitat de Barcelona 2002-present. Driector of the Structural Bioinformatics Node. Spanish National Bioinformatics Insitute. 2003-. Director Molecular Modeling and Bioinformatic Unit. Institute for Research in Biomedicine. 2001-. Director of the Life Sciences Deparment. Barcelona Supercomputer Center, 2005-. Director of the Joint IRB-BSC research program in computational biology 2007-.

Advisory Committees

Referee of projects for the European Union, Spanish Ministry of Education and Science, Spanish Ministry of Health. Referee of the Colombian Science Agency. Advisor of the Philip Morris External Research Program. Member of the advisorial board of the Argentina Science Foundation. Consultant of Boerhinger Ingelheim Pharmaceutical, Inc. USA. Consultant of Almirall Prodesfarma. Barcelona. Spain Consultant of Uriach & Cia. Barcelona. Spain. Consultant of Kraft, Inc. USA. Consultant of Pfizer Inc. Consultant of the "Subdirección General de Planificación y Seguimiento". Spanish Ministry of Science and Technology. 2002-2004. Member of the Panel of Experts of the Ministry of Science and Technology in the evaluation of Biomedicine projects. 2003. Member of the Evaluation Panel of the Program Ramon & Cajal. 2003-2006. Member of the Evaluation Panel of the Program Juan de la Cierva. 2004. Member of the Panel of Experts of the Ministry of Science and Technology in the design of the General Research Plan for 2004-2007. Area of Biotechnology. Member of the National Agency of Research Evaluation. Member of the Stearing and Scientific panel of European Supercomputer Innitiative PRACE. 2010-

Editorial experience

Associated Editor Willey Interdisciplinary Reviews. Advisory editor of Theoretical Chemistry Accounts: Theory, Computation and Modeling. Member of the Editorial board of Journal of Computational Chemistry. Member of the Editorial board of Journal of Chemical Theory and Computation. Referee for over 30 journals (including Science, Nature Methods, PNAS, Angew.Chem., Genome Res., JACS,…).

Awards

Fullbright fellowship. 1992, 1993. International award of the Chemical Structure Association for young scientists. 1992. Award of the Spanish Ministry for excellence in teaching in the period 1987-1992. Award of the Spanish Ministry for excellence in scientific research in the period 1987-1993. Award of the Spanish Ministry for excellence in scientific research in the period 1994-2000. Annual award "Sant Albert" of the Catalonian Chemical Association . 1995 National annual award for young scientists "Diaz de Santos". 1997 National annual award for young scientists of the Spanish Society of Biochemistry (SEBBM). 2000 Distinción Investigadora de la Generalitat de Catalunya. 2000 (Annual award of Science of the Catalan Science Ministry) FEBS Anniversary Prize of the Gesellschaft für Biochemie und Molekularbiologie. 2001. National Award of the Spanish Biophysics Society (Premio Bruker), 2010. Catalan Award of Excelence ICREA-Academia 2011. ERC- Advanced Grant 2012-2017

Research

Research activity focussed on the theoretical study of biological systems with emphasis in the rationalization by means of physical models of the behaviour of nucleic acids and proteins. Three of our designed drugs have reached clinical trials. Our Nearly 400 papers published in international peer-reviewed journals like Nature, Nature Genetics, Proc.Natl.Acad.Sci.USA, Chem.Rev., Angew. Chem.Int.Ed-, Chem.Soc.Rev., Acc.Chem.Res., J.Am.Chem.Soc., Genom. Biol,… More

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca

Title of abstract

Identification and characterization of signatures of structural variation across PanCancer genomes.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)

David Torrents, Head of the computational genomics group at the Barcelona Supercomputing Center. CLL-ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)

Santiago Gonzalez (PhD student)

Valentí Moncunill (Technician, Computer Scientists)

Name(s) & institute(s) of non-ICGC collaborators (Name no more than 2; append 1 page CV for each)

Background and preliminary data

The analysis of a mixed collection of tumors has recently allowed defining the first comprehensive catalogue of somatic variation in cancer (Nature 2013). This catalogue sets the basis for the diagnosis and classification of tumors that will guide the development of specific treatments. This study was based on the analysis of somatic single nucleotide substitutions and centered in coding regions. The present PanCancer initiative, which is centered in whole genomes, allows exploring, for the first time, the landscape of somatic structural variation in cancer beyond coding regions and can complement the current catalogue of variation signatures. During the past years new terminologies have been proposed to define complex scenarios of chromosomal rearrangements from the observation of a limited number of tumors that suffered bursts of non-independent translocations. To date, the terms chromotripsis and chromoplexy have been proposed to define two independent and different chromosomal catastrophes. Even though these two events differ in some key aspects (number of translocations, number of chimeric chromosomes, etc…) they overlap in others. Again, the PanCancer project offers the possibility to identify more of these events and to come up with more precise classification criteria.

But current strategies used to identify and define structural variation in cancer have important limitations. These complex karyotypes have been so far defined using a combination of experimental (SNP arrays) and expensive computational approaches, which are not possible to scale up and apply in a systematic way to all the genomes of the present PanCancer project.

In order to overcome these limitations and to obtain a comprehensive and unbiased collection of structural variants in these tumors, we propose to use SMUFIN, a novel reference-free strategy that we recently developed in the group for the characterization of structural variation in cancer genomes (manuscript submitted). As we describe in more detail in our Proposal 1, we have recently been able to identify and characterize at base pair resolution and with sensitivities and specificities above 90%, different forms of aggressive tumors with severe chromosomal rearrangements. The variants defined by SMUFIN and the posterior analysis of the breakpoints let the identification of three derivative chimeric chromosomes in a chromoplectic scenario that were confirmed by using M-FISH and PCR-sequencing techniques. The fact that SMUFIN has proved to be consistent in identifying SVs across solid and hematological tumors sequenced in different centers, makes it further ideal for the complete analysis of the heterogeneous collection of PanCancer genomes.

For these reasons, we propose (below) to address specific questions regarding the generation and nature of complex structural variation in cancer. These tasks can be part of a wider analysis of structural variation emerging from all the other proposals.

Timelines & resources dedicated to project

The resources devoted are:

- Marenostrum Supercomputer and other storage and analysis resources.
- Own experimental laboratory to perform verification when required.

Timeline:

Starting from the analysis phase (around February 2014) of the general strategy for PanCancer genome analysis (see proposal 1), genomes can be processed in parallel using SMUFIN and the results processed and derived to a database generated for later comparisons at multiple levels. In June 2014 we estimate that we can have all the definitions of structural variation in all available genomes.

Research proposal

We propose to contribute to the analysis of structural variation of PanCancer genomes by focusing in the following questions:

1) Search for translocation events significantly associated to a particular type of tumor, or to a group of tumors, i.e. structural signatures.
2) Understanding the mechanism of chromosomal rearrangements in cancer. The fact that SMUFIN provides base pair resolution, allows the analysis of elements proximal to the breakpoint, such as repeats or homologous regions, that might play a role in the underlying translocation mechanism. Here we propose to search and evaluate the possible correlation of breakpoints and close annotated elements in the genome (from gene coding and regulatory regions to repeats).
3) Provided that SMUFIN defines breakpoints independently of the reference genome, it is suitable to identify breaks affected by foreign DNA, such as viruses, or surrounded by unknown human genomic regions. A collection of regions flanking the breakpoints that cannot be located in the reference genome, will be blasted against non-redundant DNA databases.
4) Search for correlations between the type of translocation and the genomic shard produced during the event.
5) We recently found in the analysis of a MCL genome an overrepresentation of translocations between different chromosomes, resulting from the double strand breakage of DNA and the reciprocal rejoining of ends. This has also been pointed for prostate cancer recently. PanCancer genomes offer the opportunity to identify such cases, which point to regions in the chromatin that are probably very proximal or interacting in someway.

We will be happy to adapt any of these questions and procedures to the studies that are finally decided to be carried out within the PanCancer.

APPENDIX:

Short CV: David Torrents, PI

ICREA Research Professor at CNS - BSC (Centro Nacional de Supercomputación - Barcelona Supercomputing Center), Life Science Department.
Head of the Computational Genomics Group.

Research fields:
Genomics, Bioinformatics, Biomedicine, Gene Regulation, Evolution

Academic background:
PhD in Biological Sciences, Universitat de Barcelona, (2000).
Degree in Biological Sciences, Universitat de Barcelona, (1994).

Professional experience:
- ICREA Research Professor at BSC (Barcelona Supercomputing Center) - Life Sciences and Computational Biology (2006-Present).
- Staff at the EMBL (European Molecular Biology Laboratory) - Structure and computational biology (2002-2006).
- EMBO long term fellow. EMBL (European Molecular Biology Laboratory) - Structure and computational biology (2000-2002).
- PhD Student. UB (Universitat de Barcelona) - Biochemistry and Molecular Biology (1994-2000).

Selected Publications:

• Bønnelykke K, Sleiman P, Nielsen K, … & Bisgaard H. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. ***Nat Genet.*** 2013 Nov 17. doi: 10.1038/ng.2830.
• Mercader J, Puiggros M, Segrè AV... & Torrents D. Identification of novel type 2 diabetes candidate genes involved in the crosstalk between the mitochondrial and the insulin signaling systems. ***PLoS Genetics***, 2012 Dec 6; doi:10.1371/journal.pgen.1003046
• TomatoGenomeConsortium. The tomato genome sequence provides insights into fleshy fruit evolution. ***Nature***. 2012 May 30;485(7400):635-41. doi: 10.1038/nature11119.
• Adams D, Altucci L, Antonarakis SE ... & Willcocks S. BLUEPRINT to decode the epigenetic signature written in blood. ***Nat Biotechnol***. 2012 Mar 7;30(3):224-6. doi: 10.1038/nbt.2153.
• González S, Montserrat B, Sánchez F, … & Torrents D. ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. ***Bioinformatics***. 2012 Mar 15;28(6):763-70.
• Puente XS, Pinyol M, Quesada V, … Campo E. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. ***Nature.*** 2011 Jun 5;475(7354):101-5. doi: 10.1038/nature10113.
• Arumugam M, Raes J, Pelletier E, … & Bork. P. Enterotypes of the human gut microbiome.
Nature. 2011 May 12;473(7346):174-80.
• CarlosQuijano,PavelTomancak,JesusLopez-Marti & Manzanares M.* Selective maintenance of Drosophila tandemly-arranged duplicated genes during evolution ***Genome Biology***, 2008 Dec 16; 9(12):R176
• Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. ***Nature*** 2005 Sep 1;437(7055):69-87. PMID: 16136131
• Human chromosomes 2 and 4 sequencing and analysis group. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. ***Nature*** 2005 Apr 7;434(7034):724-31.

Santiago González (PhD Student)

Graduated in Biology in 2009

Graduated in Computer Sciences in 2012.

Santiago is in his third year of PhD. He has also played an important role in the development of SMUFIN and has done the analysis of chromotripsis and chromoplexy in several tumor genomes. He has also developed a set of scripts that help the definition, analysis and comparison of different scenarios of somatic chromosomal rearrangements.

Publications:

- Durán E, Djebali S, González S, … & Orozco M. Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. ***Nucleic Acids Res.*** 2013 Aug;41(15):7220-30.
- González S, Montserrat B, Sánchez F, … & Torrents D. ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. ***Bioinformatics***. 2012 Mar 15;28(6):763-70.

Valentí Moncunill (Computer Scientist)

Graduated in Computer Sciences in 2007 at the UPC, Barcelona.

His last two years have been devoted to Bioinformatics, where he has been developing software for genome analysis at different levels. He has developed SMUFIN from scratch and is applying this tool to several cancer genome projects.

He is fully capable of performing the mentioned proposal within the HPC environment of the BSC.

PLEASE TREAT THE FOLLOWING INFORMATION AS CONFIDENTIAL

Accurate characterization of complex structural variation in cancer by using a reference-free approach.

Valentí Moncunill[1,10], Santi Gonzalez[1,10], Lise Andrieux[1], Sílvia Beà[2], Itziar Salaverria[2], Cristina Royo[2], Laura Martinez[1], Montserrat Puiggròs[1,3], Maia Segura-Wang[4], Adrian M. Stütz[4], Alba Navarro[2], Romina Royo[1,3], Josep Ll. Gelpí[1,3,5], Ivo G. Gut[6], Carlos López-Otín[7], Modesto Orozco[1,5,8], Jan O. Korbel[4], Elias Campo[2], Xose S. Puente[7], David Torrents[1,9].

[1] Joint IRB-BSC Program in Computational Biology, Barcelona Supercomputing Center, 08034 Barcelona; [2] Department of Pathology, Hematopathology Unit, Hospital Clínic, August Pi I Sunyer Biomedical Research Institute (IDIBAPS), Hospital Clinic of Barcelona, University of Barcelona,08036 Barcelona, Spain; [3] Computational Bioinformatics, National Institute of Bioinformatics, Spain; [4] European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany; [5] Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain; [6] Centro Nacional de Analisis Genomico, PCB, C/Baldiri Reixac 4, 08028 Barcelona, Spain [7] Dpt. Bioquímica y Biología Molecular, Universidad de Oviedo - IUOPA, 33006 Oviedo, Spain; [8] Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Spain; [9] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona. [10]These authors contributed equally to this work
Correspondence should be addressed to DT: david.torrents@bsc.es

Abstract

The development of highthroughput sequencing technologies has changed our understanding of cancer. However, despite the increasing demand to identify the genetic alterations in tumor cells, the characterization of somatic structural variants in cancer still remains a challenge. Current strategies depend on the alignment of reads to a reference genome, a step that restricts the complete definition of structural variation. In this work we developed a reference-independent approach called SMUFIN (Somatic MUtation FINder), which is able to accurately identify all types of somatic variation, from substitutions to large structural variants (SVs), at base pair resolution. Performance tests showed average sensitivity of 92% and 74% for SNVs and SVs, with specificities of 95% and 91%, respectively. Analysis of two aggressive forms of solid and hematological tumors revealed that this procedure identifies breakpoints associated with chromothripsis and chromoplexy with high specificity. Taken together, SMUFIN constitutes the first reference-free and integrated solution for an accurate and complete characterization of somatic variation in cancer.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca

Title of abstract

Contribution to the identification of somatic variation in PanCancer genomes using SMUFIN, a reference-free approach.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

David Torrents, Head of the computational genomics group at the Barcelona Supercomputing Center. CLL-ICGC

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Valentí Moncunill (Technician, Computer Scientists)
Santiago Gonzalez (PhD student)

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Accurate identification and characterization of somatic structural variation from tumor genomes remains a big challenge. In particular, the definition of complex structural rearrangements and the resulting scenarios of tumor genomes involving chimeric chromosomes (as in chromothripsis or chromoplexia), still requires the combination of several experimental and complex computational approaches. This task is expected to be even more challenging in the context of the PanCancer project, where thousands of different tumor genomes sequenced at different sequencing centers must be described with different methods, as to their somatic variation.

In the frame of a collaboration among several ICGC groups, we have recently developed a novel methodology, called **SMUFIN** (for Somatic MUtation FINder), for the identification of all types of somatic variation in cancer genomes (manuscript submitted. Abstract attached). Our method takes FASTQ files as input and directly compares sequencing reads from normal and tumor genome samples, i.e. without the need of a reference genome, to identify SNV and SVs of all sizes in a single execution. Performance tests using in silico and real tumor data (at 60-fold depth of coverage), as well as orthogonal experimental approaches have shown average sensitivity and specificity values of 92% and 95% for SNVs, and 74% and 91% for SV respectively. SMUFIN was also able to identify, with a specificity of 92%, nearly all the breakpoints previously inferred using different experimental and computational approaches in aggressive forms of solid (Medulloblastoma) and blood (Mantle Cell Lymphoma) tumors. There, we were able to identify, at base pair resolution, hundreds of breakpoints that define chimeric forms of rearranged chromosomes that agree with the definitions of chromothripsis and chromoplexy. In addition, breakpoints corresponding to these large structural variants are provided by SMUFIN together with a reconstruction of the corresponding sequence in the tumor, including additional sequence shard or sequence not present in the reference genome.

Since the original version of SMUFIN did not allow a massive and parallel processing of tumor genomes, we have developed, in collaboration with the group of Jesús Labarta (head of the Computer Science Department), a new optimized version of SMUFIN that is able to process several genomes in parallel in cluster-based computing (or HPC) environments. This improved version of SMUFIN is based on OmpSs (http://pm.bsc.es/ompss), a programming model that enables an enhanced parallelization of programs. Although this work is still in progress, preliminary results provide estimations for the completion of a whole genome analysis (i.e. from FASTQ to somatic SNV and SVs of all sizes) of a normal and tumor pairs (at 60-fold coverage) of around 6 hours in 16 cores.

Timelines & resources dedicated to project

Once we finish the development of the parallel version of SMUFIN (estimated for February 2014), we are going to integrate the resulting program in a SLES 11 virtual machine image file ready to be executed in any cloud environment. We will test the performance on a set of genomes currently used for the benchmarking of variant calling methods within the ICGC.

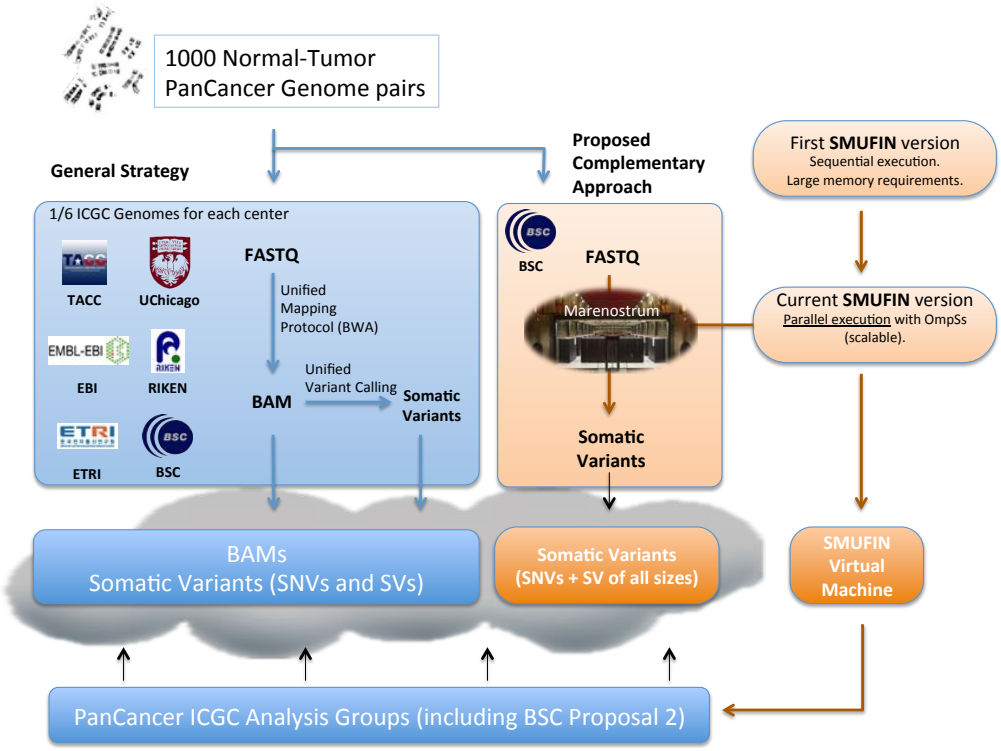
SMUFIN Requirements:
SMUFIN has been implemented using the OmpSs programming model (http://pm.bsc.es/ompss), which requires MPI (Message Passing Interface) communication between the different instances of the SMUFIN's virtual machine.
We have performed benchmark analysis in our cluster (Marenostrum III BSC) in order to know the execution time and the memory requirements for standard Normal-Tumor genome pairs, keeping the same prediction accuracy and sensitivity. For a whole genome (60x-coverage) normal-tumor comparison, the parallel version of SMUFIN running on OmpSs, requires a reservation of 16 virtual machine instances (nodes) with each one having at least 25GB of available RAM and needs approximately 6 hours of execution.
We will dedicate the necessary computing power in our center to complete this task within the estimated timeline below.

Timelines:

Feb.2014: Parallel version of SMUFIN finished. Ready to use VMs available for any ICGC-Pancancer member.
Feb-2014- June 2014: SMUFIN's Variant Calling for all ICGC PanCancer genomes at the BSC using the HPC Marenostrum. Distribution of variant lists to all ICGC-PanCancer members.

Research proposal

From the Barcelona Supercomputing Center, one of the six centers involved in the primary analysis of PanCancer genomes, we propose an additional approach for variant calling complementary to those already included in the general PanCancer strategy (BAMs + Variant Calling Pipelines). This approach consists in running SMUFIN on all the ICGC whole genomes at the BSC using Marenostrum Supercomputer. The proposed strategy is summarized in the scheme below.



We believe that the generation and distribution of SMUFINs calls of somatic variation can complement the results obtained with alignment-based methods at following levels:

1) SMUFIN can provide large structural variants at base pair resolution level with a sensitivity and specificity values above 90%. These include an important fraction of rearrangement events that remain undetectable to current methods, as an important part of reads that cover them cannot be unambiguously mapped onto the reference genome. The fact that the ICGC PanCancer project is centered in the analysis of whole genomes, makes necessary an accurate calling of large structural variants, not detectable in previous exome-based pan-analyses of tumors.

2) Our approach also provides the sequence of the tumor genome around breakpoints associated to intra or chromosomal translocations. This is key to study the underlying mechanism of chromosomal rearrangements in chromotriptic and chromoplexic scenarios.

3) In combination with the other methods, SMUFIN can aid the definition of complex landscapes of genomic rearrangements, including the identification of chimeric chromosomes, as we recently did on a MCL genome.

4) The fact that SMUFIN uses the same underlying search mechanisms to identify during the same run, small and large somatic variants, the performance values for each type of variant results similar. This allows the comparison of mutation rates within and among genomes without the need of correcting for the biases derived from the combination of different alignment-based methodologies.

On top of the proposed analysis of the 1000 ICGC genome pairs, we will also provide SMUFIN in the form of a Virtual Machine that users can use to analyze additional genomes not included in the original collection of samples, and that NIH trusted centers (e.g. University of Chicago) cap apply to TCGA genomes.

Our group we will be open to support the installation and deployment of the virtual machine at any PanCancer group.

APPENDIX:

Short CV: David Torrents, PI

ICREA Research Professor at CNS - BSC (Centro Nacional de Supercomputación - Barcelona Supercomputing Center), Life Science Department.
Head of the Computational Genomics Group.

Research fields:
Genomics, Bioinformatics, Biomedicine, Gene Regulation, Evolution

Academic background:
PhD in Biological Sciences, Universitat de Barcelona, (2000).
Degree in Biological Sciences, Universitat de Barcelona, (1994).

Professional experience:
- ICREA Research Professor at BSC (Barcelona Supercomputing Center) - Life Sciences and Computational Biology (2006-Present).
- Staff at the EMBL (European Molecular Biology Laboratory) - Structure and computational biology (2002-2006).
- EMBO long term fellow. EMBL (European Molecular Biology Laboratory) - Structure and computational biology (2000-2002).
- PhD Student. UB (Universitat de Barcelona) - Biochemistry and Molecular Biology (1994-2000).

Selected Publications:

• Bønnelykke K, Sleiman P, Nielsen K, … & Bisgaard H. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. ***Nat Genet.*** 2013 Nov 17. doi: 10.1038/ng.2830.
• Mercader J, Puiggros M, Segrè AV... & Torrents D. Identification of novel type 2 diabetes candidate genes involved in the crosstalk between the mitochondrial and the insulin signaling systems. ***PLoS Genetics***, 2012 Dec 6; doi:10.1371/journal.pgen.1003046
• TomatoGenomeConsortium. The tomato genome sequence provides insights into fleshy fruit evolution. ***Nature***. 2012 May 30;485(7400):635-41. doi: 10.1038/nature11119.
• Adams D, Altucci L, Antonarakis SE ... & Willcocks S. BLUEPRINT to decode the epigenetic signature written in blood. ***Nat Biotechnol***. 2012 Mar 7;30(3):224-6. doi: 10.1038/nbt.2153.
• González S, Montserrat B, Sánchez F, … & Torrents D. ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. ***Bioinformatics***. 2012 Mar 15;28(6):763-70.
• Puente XS, Pinyol M, Quesada V, … Campo E. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. ***Nature.*** 2011 Jun 5;475(7354):101-5. doi: 10.1038/nature10113.
• Arumugam M, Raes J, Pelletier E, … & Bork. P. Enterotypes of the human gut microbiome.
Nature. 2011 May 12;473(7346):174-80.
• CarlosQuijano,PavelTomancak,JesusLopez-Marti & Manzanares M.* Selective maintenance of Drosophila tandemly-arranged duplicated genes during evolution ***Genome Biology***, 2008 Dec 16; 9(12):R176
• Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. ***Nature*** 2005 Sep 1;437(7055):69-87. PMID: 16136131
• Human chromosomes 2 and 4 sequencing and analysis group. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. ***Nature*** 2005 Apr 7;434(7034):724-31.

Valentí Moncunill (Computer Scientist)

Graduated in Computer Sciences in 2007 at the UPC, Barcelona.

His last two years have been devoted to Bioinformatics, where he has been developing software for genome analysis at different levels. He has developed SMUFIN from scratch and is applying this tool to several cancer genome projects.

He is fully capable of performing the mentioned proposal within the HPC environment of the BSC.

Santiago González (PhD Student)

Graduated in Biology in 2009

Graduated in Computer Sciences in 2012.

Santiago is in his third year of PhD. He has also played an important role in the development of SMUFIN and has done the analysis of chromotripsis and chromoplexy in several tumor genomes. He has also developed a set of scripts that help the definition, analysis and comparison of different scenarios of somatic chromosomal rearrangements.

Publications:

- Durán E, Djebali S, González S, … & Orozco M. Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. ***Nucleic Acids Res.*** 2013 Aug;41(15):7220-30.
- González S, Montserrat B, Sánchez F, … & Torrents D. ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. ***Bioinformatics***. 2012 Mar 15;28(6):763-70.

PLEASE TREAT THE FOLLOWING INFORMATION AS CONFIDENTIAL

Accurate characterization of complex structural variation in cancer by using a reference-free approach.

Valentí Moncunill[1,10], Santi Gonzalez[1,10], Lise Andrieux[1], Sílvia Beà[2], Itziar Salaverria[2], Cristina Royo[2], Laura Martinez[1], Montserrat Puiggròs[1,3], Maia Segura-Wang[4], Adrian M. Stütz[4], Alba Navarro[2], Romina Royo[1,3], Josep Ll. Gelpí[1,3,5], Ivo G. Gut[6], Carlos López-Otín[7], Modesto Orozco[1,5,8], Jan O. Korbel[4], Elias Campo[2], Xose S. Puente[7], David Torrents[1,9].

[1] Joint IRB-BSC Program in Computational Biology, Barcelona Supercomputing Center, 08034 Barcelona; [2] Department of Pathology, Hematopathology Unit, Hospital Clínic, August Pi I Sunyer Biomedical Research Institute (IDIBAPS), Hospital Clinic of Barcelona, University of Barcelona,08036 Barcelona, Spain; [3] Computational Bioinformatics, National Institute of Bioinformatics, Spain; [4] European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany; [5] Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain; [6] Centro Nacional de Analisis Genomico, PCB, C/Baldiri Reixac 4, 08028 Barcelona, Spain [7] Dpt. Bioquímica y Biología Molecular, Universidad de Oviedo - IUOPA, 33006 Oviedo, Spain; [8] Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Spain; [9] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona. [10]These authors contributed equally to this work
Correspondence should be addressed to DT: david.torrents@bsc.es

Abstract

The development of highthroughput sequencing technologies has changed our understanding of cancer. However, despite the increasing demand to identify the genetic alterations in tumor cells, the characterization of somatic structural variants in cancer still remains a challenge. Current strategies depend on the alignment of reads to a reference genome, a step that restricts the complete definition of structural variation. In this work we developed a reference-independent approach called SMUFIN (Somatic MUtation FINder), which is able to accurately identify all types of somatic variation, from substitutions to large structural variants (SVs), at base pair resolution. Performance tests showed average sensitivity of 92% and 74% for SNVs and SVs, with specificities of 95% and 91%, respectively. Analysis of two aggressive forms of solid and hematological tumors revealed that this procedure identifies breakpoints associated with chromothripsis and chromoplexy with high specificity. Taken together, SMUFIN constitutes the first reference-free and integrated solution for an accurate and complete characterization of somatic variation in cancer.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Alien DNA, Garbagenomics

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Tyler Alioto, Centro Nacional de Analisis Genomico, Barcelona, Spain
Ivo Gut, Centro Nacional de Analisis Genomico, Barcelona, Spain

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Paolo Ribeca, Centro Nacional de Analisis Genomico, Barcelona, Spain	**Tomas Marques-Bonet, Universitat Pompeu Fabra, Barcelona, Spain**

Background and preliminary data

Viral genomes and retro-elements have been shown to be relevant to the etiology of several types of human and non-human cancer. However they intrinsically show a very high degree of polymorphism, which makes them elusive and very difficult to study systematically in a reference-based setup. In fact, most resequencing pipelines ignore both “alien” reads as those potentially stemming from viruses, and very repetitive reads as those produced by the sequencing of repetitive elements. Similar considerations can be formulated for viral integration sites in the human genome. They have the potential to disrupt genes, alter transcript expression or splicing and promote further genomic instability and rearrangements, but, being hard to study in a framework based on short sequencing reads, they are often utterly neglected.

As a genome sequencing center, the CNAG has a long-standing experience in the detection of genomic variants and rearrangements, through both a careful usage of existing software pipelines and the development of new ones. We have developed a novel very accurate short read mapper/split mapper particularly suited to conduct high-precision genomic studies (Marco-Sola et al. Nature Methods 2012). We have been central to the recent project of evaluating the quality of genome variant calling in the ICGC/CLL group. We have already participated in the detection of polymorphic elements in comparative genomics including cross species mapping (Hormozdiari et al. PNAS 2013). Although this study would require a somehow special angle, the center possesses all the technology necessary to successfully implement and carry it out.

Timelines & resources dedicated to project
To be able to perform the proposed research, we will need to have access to the sequencing reads which are discarded by the standard alignment pipelines used within the ICGC. This should happen in an unbiased way, i.e. we should gain access, independently of the details of the alignment parameters used by each pipeline, to the raw sequencing data such that one or two read ends turn out to be unmappable as they contain viral or very repetitive sequences. As a result, we might or might not need to perform realignment of a big amount of sequencing data to isolate such reads. Once we have obtained the reads interesting to our project, we estimate 3-4 months of computation for clustering algorithms and detection of potential integration sites, correcting for the estimated heterogeneity of cells based on SNP lists already available for each cancer type. We estimate we can have a full report for the group in 6 months. If necessary to complete the project on time, should cloud computing resources be limited, CNAG computational resources can be utilized to carry out the more intensive analyses.

Research proposal

Our proposal focuses on (1) the systematic comparison of the extent of presence/polymorphism of viral/repetitive sequences in human cancer genomes compared to the expectation from normal tissues (2) the systematic exploration of the features of integration sites of viral/repetitive sequence in human cancer genomes (3) the correlation between the nature of viral/repetitive sequences, the nature of their integration sites, and the phenotype of the corresponding tumors (4) determining whether the presence of viral/repetitive sequence in the germ-line can be related to the genome instability seen in the corresponding cancer cells.

Our proposal focuses on (1) the systematic comparison of the extent of presence/polymorphism of viral/repetitive sequences in human cancer genomes compared to the expectation from normal tissues (2) the systematic exploration of the features of integration sites of viral/repetitive sequence in human cancer genomes (3) the correlation between the nature of viral/repetitive sequences, the nature of their integration sites, and the phenotype of the corresponding tumors (4) determining whether the presence of viral/repetitive sequence in the germ-line can be related to the genome instability seen in the corresponding cancer cells.

Detection of genomic variants (as viral/repetitive sequence would look like with respect to the reference) from short sequencing data has received a lot of attention in recent times, and several workflows are possible. Although the present project will likely require some specialized analysis pipeline due to the amount and the complexity of the data to be processed, we can mention a couple of existing methods that might serve as inspiration. When considering retro-elements and their integration sites, one might use a modification of a previously published method VariationHunter (Hormozdiari et al. Bioinformatics 2010). In particular, one might detect both polymorphism of repetitive elements already existing in the human reference and novel integration sites (novel integration sites can be predicted for each sample based on aggregation of events and coverage estimates). The problem of detecting presence of viruses in sequencing data, both integrated into the genome or not, has also been previously tackled in the literature (for instance by program VirusFinder, Qingguo et al. PLos ONE 2013). Databases of viruses to classify reads unmapped to the reference are readily available, and several standard methods (based on discordant paired end data and fusion read detection) can be used to identify integration events.

The final goal will be to detect known or novel inserted sequences alien to the genome, together with a set of well characterized insertion sites. Recurrent locations and rates of integration among different types of cancer will then be analyzed. Obviously, annotation of sites together with the potential impact on nearby genes will be taken into account. Specifically, we would like to test whether genes related to recombination/repair machinery are more often affected. In the case of potential recurrent sites, we will proceed with a machine learning approach to detect potential motifs or patterns that may allow us to explain the non-randomness of integration. Finally, when such information is available we will correlate the nature of the event, both in germ-line and in cancer cells, with the clinical history of the patient.

Legacy plans

The outcome of this analysis will be written up for a companion publication with the Pan-Cancer flagship paper. It is not known at the time of this writing whether we will be entirely relying upon published methods, or the project will require a substantial development of new software. In the latter case, we might additionally consider the publication of a method paper.

Dr. Tyler Alioto leads the Genome Annotation and Assembly team of the Bioinformatics Development group at the Centro Nacional de Análisis Genómico in Barcelona, Spain. His team, currently made up of one postdoc and one Masters student, in addition to himself, assembles and annotates new eukaryotic genomes de novo. Current and past research carried out at the CNAG, the Centre for Genomic Regulation (postdoctoral), the University of California at Berkeley (doctoral), and Stanford University (predoctoral), has led to more than 20 peer-reviewed publications focused on gene family evolution, RNA splicing, RNA-seq analysis, gene prediction, genome annotation and genome sequence assembly. In 2010 and 2011 he organized the de novo Genome Assembly Assessment Project (dnGASP) and led the analysis team, which was tasked with evaluating submitted assemblies of a benchmark genome. Work on this project resulted in the dnGAAS server, an online genome assembly benchmarking service. Current projects include the sequencing, assembly and annotation of the Iberian lynx, the turbot, the Galician mussel, the Mediterranean pine, the olive tree, the carnation, the cedar aphid and a parasitic wasp. His team has also annotated the genomes of the melon and the common bean.

SELECTED PUBLICATIONS

1. **Alioto T**, Picardi E, Guigó R, Pesole G. ASPic-GeneID: A Lightweight Pipeline for Gene Prediction and Alternative Isoforms Detection. BioMed Research International. Article ID 502827, 2013.
2. Steijger T, Abril JF, Engström PG, Kokocinski F; The RGASP Consortium. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013 Nov 3.
3. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A; The RGASP Consortium. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013 Nov 3.
4. Piqué N, Aquilini E, **Alioto T**, Miñana-Galbis D, Tomás JM. Genome Sequence of Plesiomonas shigelloides Strain 302-73 (Serotype O1). Genome Announc. 2013 Jun 27;1(4).
5. Djebali S, et al. Landscape of transcription in human cells. Nature. 2012 Sep 6;489(7414):101-8.
6. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.
7. Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E, **Alioto T**, Capella-Gutiérrez S, Blanca J, Cañizares J, Ziarsolo P, Gonzalez-Ibeas D, Rodríguez-Moreno L, Droege M, Du L, Alvarez-Tejado M, Lorente-Galdos B, Melé M, Yang L, Weng Y, Navarro A, Marques-Bonet T, Aranda MA, Nuez F, Picó B, Gabaldón T, Roma G, Guigó R, Casacuberta JM, Arús P, Puigdomènech P. The genome of melon (Cucumis melo L.). Proc Natl Acad Sci USA. 2012 Jul 17;109(29):11872-7.
8. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012 May 30;485(7400):635-41.
9. **Alioto T**. Gene prediction. Methods Mol Biol. 2012;855:175-201.
10. Mudge JM, Frankish A, Fernandez-Banet J, **Alioto T**, Derrien T, Howald C, Reymond A, Guigó R, Hubbard T, Harrow J. The origins, evolution, and functional potential of alternative splicing in vertebrates. Mol Biol Evol. 2011 Oct;28(10):2949-59.
11. International Aphid Genomics Consortium. Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol. 2010 Feb 23;8(2):e1000313.

IVO GUT is **Director of the Centro Nacional de Análisis Genómico (CNAG)** in Barcelona, one of the largest European genome sequencing operations, which he established in 2010. He received his **PhD in Physical Chemistry** from the University of Basel in 1990. His post-doctoral work was at Harvard Medical School and the Imperial Cancer Research Foundation of London, he led a group in the Department for Vertebrate Genomics at Max-Planck-Institute for Molecular Genetics in Berlin.
For the 11 years prior to CNAG he was at the Centre National de Génotypage (CNG) – CEA as Associate Director and in charge of Technology Development. He initiated and was coordinator of the 16 MEuro EU-funded project READNA on nucleic acid technology development. His research interests are genomics, high-throughput nucleic acid analysis methods, proteomics, omics technologies, automation, bioinformatics, disease gene identification, cancer, and agrogenomics. He has more than 20 years experience and has authored more than 170 research papers, 11 reviews and 12 book chapters, cited over 15000 times. He is inventor of 25 patents or patent applications, founder of 4 biotechs, and serves on numerous international advisory boards. Significant achievements are the development of pioneering methods for DNA analysis by mass spectrometry (genotyping, sequencing, DNA methylation analysis and haplotyping), bisulphite Pyrosequencing, TAMSIM a Imaging Mass Spectrometry method, execution of many of the major GWAS and ICGC studies.

SELECTED PUBLICATIONS

1. Prado-Martinez J, Sudmantet P H, Kidd J M, … **Gut I G**, Eichler E E, Marques-Bonet T. (2013). Great ape genetic diversity and population history. Nature 499(7459): p. 471-5.
2. Kulis M., Heath S, Bibikova M, … **Gut I**, C. Lopez-Otin, E. Campo and J. I. Martin-Subero (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44(11): 1236-1242.
3. Narni-Mancinelli, Jaeger EBN, Bernat C, … **Gut I G**, E. Vivier and S. Ugolini (2012). Tuning of natural killer cell reactivity by NKp46 and Helios calibrates T cell responses. Science 335(6066): 344-348.
4. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, Eick D, **Gut I**, Ferrier P, Andrau JC. (2011). . Nat Struct Mol Biol. 17;18(8):956-63.
5. Puente XS, Pinyol M, Quesada V, … **Gut I**, López-Guillermo A, Estivill X, Montserrat E, López-Otín C, Campo E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 5;475(7354):101-5.
6. Moffatt MF, **Gut I G**, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. (2010). A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 363(13):1211-21.
7. Lambert JC, Heath S, … **Gut I**, Van Broeckhoven C, Alpérovitch A, Lathrop M, Amouyel P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 41(10):1094-9.
8. Bishop DT, Demenais F, … **Gut I**, …., Lathrop GM, Barrett JH, Bishop JA. (2009). Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 41(8):920-5.
9. Hung RJ, McKay JD, … **Gut I**, … Brennan P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452: 633-637.
10. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, **Gut I**, Lathrop M and Collins R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 359, 789-799.
11. Moffatt MF, Kabesch M, Liang L, … **Gut I G**, Lathrop GM, Cookson WO. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 448: 470-473.

Paolo Ribeca – Short CV

Research interests

A physicist by education, my research interests focus on the application of various disciplines of computer science, physics and mathematics to high-performance scientific computing in genomics and biology.

Since the inception of high-throughput sequencing techniques, I specialized on algorithms for short-read processing. I am the main architect of GEM, a software package for the analysis of short sequence reads produced by modern high-throughput sequencers (http://gemlibrary.sourceforge.net).

Current position

I currently lead the Algorithm Development unit at the Centro Nacional de Análisis Genómico (CNAG), the Spanish National sequencing center. My group provides bleeding-edge algorithmic methods, to make possible the precise and timely analysis of the ~1Tb of genomic data produced by the CNAG every day.

Apart from myself, the group includes a postdoctoral fellow and two Ph.D. students whose I am the supervisor.

Our methods for short-read alignment (in particular the GEM mapper, S.Marco-Sola et al. Nature Methods, December 2012) are considered to be among the fastest and most accurate tools in their class, and are being adopted by an increasing number of labs and consortia.

The group is actively involved in many bleeding-edge biological projects, typically focusing on either de-novo genome assembly, or on the analysis of cell expression regulation by RNA-sequencing.

Finally, I am visiting fellow at the Pirbright Institute (formerly Institute for Animal Health), UK, one of the top institutions in the world in animal virology.

Participation in projects/consortia

2011-present. The GEUVADIS (Genetic European Variation in Disease) consortium
2010-present. The Iberian Lynx sequencing consortium
2009-2012. The ENCODE (Encyclopedia of DNA Elements) consortium
2009-2011. The Tomato sequencing consortium.

Publications

20 publications so far (18 peer-reviewed papers [including 4 Nature, 3 Nature Methods and 1 Nature Biotechnology], plus 2 book chapters).

They accumulate ~1900 citations to-date (h-index 11). For more details see http://scholar.google.com/citations?user=juUtwYAAAAAJ&hl=en.

Additional relevant information

I act as peer-reviewer for several journals and conferences (in particular, I have been nominated leading reviewer'' by Bioinformatics.)

I am often invited speaker at conferences and schools. I am one of the teachers of the Master in Bioinformatics of the University of Murcia.

Previous jobs

2008-2010. Post-doctoral position at the Center for Genomic Regulation (CRG), Barcelona, in the Bioinformatics and Genomics group.
2006-2008. Post-doctoral position at the Center for Genomic Regulation (CRG), Barcelona, in the Systems Biology group.
2006. Invited scientist at GSI/Darmstadt and Max-Planck Institut für Kernphysik/Heidelberg.
2003-2005. Post-doctoral position at the Humboldt Universität, Berlin, in the Computational Physics group.

Dr. Marques-Bonet is Principal Investigator of the group "Comparative Genomics" as a part of the Universitat Pompeu Fabra. He started his own lab in 2010 after a **Marie Curie fellowship** in Seattle, University of Washington. In 2010 he obtained the competitive **ERC Starting Grant 2010** to establish his own group centered on the analysis **genetic diversity in great apes**. In 2011, he was also granted a Spanish Grant (MICINN (BFU2011-28549), to study the effects of reorganizations in the phenotype and was selected as an **ICREA research investigator** at the Universitat Pompeu Fabra (UPF). In 2013, he was selected for the EMBO young Investigator award. He has been part of **many Genome Consortia (Neanderthal, Denisovan, Bonobo, Gorilla, Orangutan, Gibbon, Marmoset. Lynx, Cats, E-shark)**, in the section of duplications and structural variation in most of them. His group is now formed by 4 PhD students, 1 postdoc and 2 technicians whose work is focused on characterizing genetic diversity in primates, in methods for estimating selection on duplicated sequences, the evolution of epigenetics in humans or the impact of CNVs in phenotypic traits. With a total of more than 45 peer-reviewed publications, this year (2013) he has published 11 papers, 4 of them with **senior authorship** in **Nature, Plos Genetics, Genome Biology and BMC genomics.**

SELECTED PUBLICATIONS

1. Javier Prado-Martinez*, Peter H. Sudmant*, Great ape genome consortium, Evan E. Eichler+, **Tomas Marques-Bonet**+. (2013). "Great ape genetic diversity and population history". Nature, 2013.
2. Irene Hernando-Herraez, Javier Prado-Martinez, Paras Garg, Marcos Fernandez-Callejo, Holger Heyn, Christina Hvilsom, Arcadi Navarro, Manel Esteller, Andrew J. Sharp, **Tomas Marques-Bonet** "Dynamics of DNA Methylation in Recent Human and Great Apes Evolution". PLOS Genetics, 2013.
3. Belen Lorente-Galdos, Jon Bleyhl, Gabriel Santpere, Laura Vives, Oscar Ramirez, Jessica Hernandez, Roger Anglada, Greg M. Copper, Arcadi Navarro, Evan Eichler, **Tomas Marques-Bonet** "Fast exon evolution in duplicated regions in hominids". Genome Biology, 2013.
4. Javier Prado-Martinez, Irene Hernando-Herraez, Belen Lorente-Galdos, Marc Dabad, Oscar Ramirez, ..., Marta Gut, Jaume Bertranpetit, Ivo G. Gut, Teresa Abello, Ismael Mingarro, Evan E. Eichler, Carles Lalueza-Fox, Arcadi Navarro and **Tomas Marques-Bonet**. "The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild. " BMC Genomics, 2013.
5. Fereydoun Hormozdiari, Miriam Konkel, Javier Prado-Martinez, Giorgia Chiatante, Irene Hernando Herraez, Jerilyn Walker, Ben Nelson, Can Alkan, Peter H. Sudmant, John Huddleston, Claudia R. Catacchio, Arthur Ko, Maika Malig, Carl Baker, **Tomas Marques-Bonet**, Mario Ventura, Mark Batzer, and Evan E. Eichler. "Rates and Patterns of Great Ape Retrotransposition." PNAS 2013, Aug 13;110(33):13457-62
6. Neandertal Consortium. "A draft sequence and preliminary analysis of the Neandertal genome" Science, 2010.
7. Gorilla Genome Consortium. Insights into hominid evolution from the gorilla genome sequence" Nature, 2012.
8. Orangutan Genome Consortium. "Comparative and demographic analysis of orang-utan genomes" Nature, 2011 (Cover)
9. **Tomas Marques-Bonet**, Jeffrey M. Kidd, Mario Ventura, Tina A. Graves, Ze Cheng, LaDeanna W. Hillier, Zhaoshi Jiang, Carl Baker, Ray Malfavon-Borja, Lucinda A. Fulton, Can Alkan, Gozde Aksay, Priscillia Siswara, Lin Chen, Maria Francesca Cardone, Arcadi Navarro, Elaine R. Mardis, Richard K. Wilson, Evan E. Eichler. "A Burst of Segmental Duplications in the African Great Ape Genome Ancestor" Nature, 2009.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Functional Mutations	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators **(Name no more than 2; append 1 page CV for each)**	
Ivo Glynne Gut, Centro Nacional de Análsis Genómico (CNAG), Barcelona, Spain **José Ignacio Martín-Subero, University of Barcelona (UB), Barcelona, Spain**	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Background and preliminary data	
The distinction of driver and passenger mutations in tumor genomes is frequently made on grounds of function. Function in a biochemical sense involves the production of a gene product that is part of a biochemical pathway or network. Typically genomic changes that result in fusion genes, stop codons (gained or lost) or amino acid changes are viewed as functional drivers of cancer. Commonly, little emphasis is being placed on how these types of genetic variants came to be. Given that only 1-2% of the genome is coding, it seems highly unlikely that all of the commonly considered functional driver mutations are primary genomic somatic events. On the other hand the 98-99% non-exomic genome accumulates somatic mutation at about the same rate. Consequently, 99 out of 100 somatic mutations are not in the exome and do not lead to alterations of gene products. Recent studies indicate that non-exomic mutations and chromosomal breakpoints take place in genomic regions with a particular chromatin organization (Stephens et al. Cell. 2011;144(1):27-40; Fudenberg et al. Nature Biotech. 2011;29(12):1109-13;Schuster-Böckler & Lehner. Nature. 2012;488(7412):504-7 and Baca et al. Cell. 2013;153(3):666-77). Also, our own group has identified that epimutations (DNA methylation changes) in chronic lymphocytic leukemia take place mostly in particular functional domains such as enhancers (Kulis M et al. Nat Genet. 2012;44(11):1236-42). These data lead us to postulate that genetic alterations in the non-coding part of the genome may in part affect regulatory regions with a relevant role in carcinogenesis. Projects such as the ENCODE and IHEC are improving the annotation of the non-exomic part of the genome into functional chromatin states in different cell types, providing thus a rich resource to link genetic alterations in cancer and epigenomic architecture. Interesting is also that several reported recurrent exomic mutations appear to be non-random and are not the result of repairing base excised bases. Examples of this are a recurrent 2-base deletion in the NOTCH1 gene identified in CLL (Puente et al. Nature. 2011;475:101-5) and later described in other leukemias and a 2-base deletion in the GATA3 gene in luminal breast cancer (Ciriello et al. Breast Cancer Res Treat. 2013;141(3):409-20).	
Timelines & resources dedicated to project	

This proposal involves informatic, bioinformatic and statistical analysis of variant calls coming from the PanCancer analysis. For resources, we will draw on analytical resources available in our teams at the CNAG and UB. In a first stage (2 months from when mutation calls become available) we will identify across all tumor-types recurrent somatic motifs (e.g. multibase deletions), mutations in key functional classes of genes that impact DNA modifications (DNA repair genes, genes involved in epigenetic processes (DNMTs, TETs, HDAC, etc), mutations in classes of functionally important non-coding elements such as promoters, enhancers, transcription factor binding sites). Complete mutational profiles from these samples falling into these classes will be extracted and analyzed statistically (3 months).

Research proposal

The objective of this proposal is to gain insight into the genesis of de novo mutations using an integrative approach. What (epi)genomic factors are required for particular mutational profiles to appear and mechanisms to be activated?

We propose to carry out a computational search for somatic genomic driver mutations, both in coding and non-coding parts of the genome. Somatic genomic driver mutations would be variants that lead to further changes in the genome, which in turn could be disease driving mutations. Themselves they would be less likely to impact on one of the well-described cancer pathways. To achieve this we follow several lines. We will stratify the entire PanCancer somatic variant calls according to recurrent somatic mutations that are:
1) in genes that have DNA modifying properties (e.g. DNA repair genes, genes),
2) in genes that are involved in epigenetic processes (DNMTs, TETs, HDAC, etc.),
3) in non-genic functional elements (promoters, enhancers, transcription factor binding sites, etc.),
4) unusual mutations that would require sophisticated mechanisms to create them and are non-random (e.g. recurrent multi-base deletions, that occur in one particular position, without similar observations in the vicinity).

Points 1)-3) are forward directed questions – what is the downstream effect if there is a mutation in a particular gene or regulatory element? What are the base changes, structural changes, nature of cleavages (single or double strand) that are observed? In 4) we reverse the question, and ask if there is a specific mutation within a gene, or mechanistic element that carries a change relative to the reference that might be driving effects such as specific base excisions.

Legacy plans

The findings will be written up for publication to go as a companion paper with the flagship PanCancer analysis paper. The established methodology will be made available to apply to later, larger analyses.

IVO GUT is **Director of the Centro Nacional de Análisis Genómico (CNAG)** in Barcelona, one of the largest European genome sequencing operations, which he established in 2010. He received his **PhD in Physical Chemistry** from the University of Basel in 1990. His post-doctoral work was at Harvard Medical School and the Imperial Cancer Research Foundation of London, he led a group in the Department for Vertebrate Genomics at Max-Planck-Institute for Molecular Genetics in Berlin.
For the 11 years prior to CNAG he was at the Centre National de Génotypage (CNG) – CEA as Associate Director and in charge of Technology Development. He initiated and was coordinator of the 16 MEuro EU-funded project READNA on nucleic acid technology development. His research interests are genomics, high-throughput nucleic acid analysis methods, proteomics, omics technologies, automation, bioinformatics, disease gene identification, cancer, and agrogenomics. He has more than 20 years experience and has authored more than 170 research papers, 11 reviews and 12 book chapters, cited over 15000 times. He is inventor of 25 patents or patent applications, founder of 4 biotechs, and serves on numerous international advisory boards. Significant achievements are the development of pioneering methods for DNA analysis by mass spectrometry (genotyping, sequencing, DNA methylation analysis and haplotyping), bisulphite Pyrosequencing, TAMSIM a Imaging Mass Spectrometry method, execution of many of the major GWAS and ICGC studies.

SELECTED PUBLICATIONS

1. Prado-Martinez J, Sudmantet P H, Kidd J M, … **Gut I G**, Eichler E E, Marques-Bonet T. (2013). Great ape genetic diversity and population history. Nature 499(7459): p. 471-5.
2. Kulis M., Heath S, Bibikova M, … **Gut I**, C. Lopez-Otin, E. Campo and J. I. Martin-Subero (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44(11): 1236-1242.
3. Narni-Mancinelli, Jaeger EBN, Bernat C, … **Gut I G**, E. Vivier and S. Ugolini (2012). Tuning of natural killer cell reactivity by NKp46 and Helios calibrates T cell responses. Science 335(6066): 344-348.
4. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, Eick D, **Gut I**, Ferrier P, Andrau JC. (2011). . Nat Struct Mol Biol. 17;18(8):956-63.
5. Puente XS, Pinyol M, Quesada V, … **Gut I**, López-Guillermo A, Estivill X, Montserrat E, López-Otín C, Campo E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 5;475(7354):101-5.
6. Moffatt MF, **Gut I G**, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. (2010). A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 363(13):1211-21.
7. Lambert JC, Heath S, … **Gut I**, Van Broeckhoven C, Alpérovitch A, Lathrop M, Amouyel P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 41(10):1094-9.
8. Bishop DT, Demenais F, … **Gut I**, …., Lathrop GM, Barrett JH, Bishop JA. (2009). Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 41(8):920-5.
9. Hung RJ, McKay JD, … **Gut I**, … Brennan P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452: 633-637.
10. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, **Gut I**, Lathrop M and Collins R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 359, 789-799.
11. Moffatt MF, Kabesch M, Liang L, … **Gut I G**, Lathrop GM, Cookson WO. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 448: 470-473.

Dr. José Ignacio Martin-Subero is **Principal Investigator of the epigenomics group** of the Department of Pathology of the University of Barcelona. He started his own lab in 2010 with a **Ramon y Cajal contract** from the Spanish Government. Since then, he has received funding from several organizations including the European Commission, the European Hematology Association, the Spanish Government and the Fundació La Marató de TV3. He is the leader of the epigenomics work package of the **ICGC-associated Spanish CLL Genome Project**, being the first results of his research published last year in Nature Genetics (Kulis et al., 2012). He is also one of the principal investigators of the EU-funded **BLUEPRINT Consortium**, which is part of the **International Human Epigenome Consortium**. His group is currently made out of 2 PhD students, 2 postdocs and 1 technician whose work is focused on characterizing the epigenome of normal and neoplastic lymphoid cells. He has published or co-coauthored a total 120 peer-reviewed publications, being 4 of them recently published as **senior author** in **Nature Genetics, Leukemia, Biochim Biophys Acta and PLoS ONE.**

SELECTED PUBLICATIONS

1. Beà S, Valdés-Mas R, Navarro A, Salaverria I, Martín-Garcia D, Jares P, Giné E, Pinyol M, Royo C, Nadeu F, Conde L, Juan M, Clot G, Vizán P, Di Croce L, Puente DA, López-Guerra M, Moros A, Roue G, Aymerich M, Villamor N, Colomo L, Martínez A, Valera A, **Martín-Subero JI**, Amador V, Hernández L, Rozman M, Enjuanes A, Forcada P, Muntañola A, Hartmann EM, Calasanz MJ, Rosenwald A, Ott G, Hernández-Rivas JM, Klapper W, Siebert R, Wiestner A, Wilson WH, Colomer D, López-Guillermo A, López-Otín C, Puente XS, Campo E. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. PNAS 2013 Nov 5;110(45):18250-5
2. Kulis M, Queirós AC, Beekman R, **Martín-Subero JI**. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. Biochim Biophys Acta 2013 Nov;1829(11):1161-74.
3. **Martín-Subero JI**, López-Otín C, Campo E. Genetic and epigenetic basis of chronic lymphocytic leukemia. Curr Opin Hematol 2013 Jul;20(4):362-8.
4. Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, Clot G, Martínez-Trillos A, Castellano G, Brun-Heath I, Pinyol M, Barberán-Soler S, Papasaikas P, Jares P, Beà S, Rico D, Ecker S, Rubio M, Royo R, Ho V, Klotzle B, Hernández L, Conde L, López-Guerra M, Colomer D, Villamor N, Aymerich M, Rozman M, Bayes M, Gut M, Gelpí JL, Orozco M, Fan JB, Quesada V, Puente XS, Pisano DG, Valencia A, López-Guillermo A, Gut I, López-Otín C, Campo E, **Martín-Subero JI**. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 2012 Nov;44(11):1236-42.
5. Vegliante MC, Royo C, Palomero J, Salaverria I, Balint B, Martín-Guerrero I, Agirre X, Lujambio A, Richter J, Xargay-Torrent S, Bea S, Hernandez L, Enjuanes A, Calasanz MJ, Rosenwald A, Ott G, Roman-Gomez J, Prosper F, Esteller M, Jares P, Siebert R, Campo E, **Martín-Subero JI**+, Amador V+. Epigenetic activation of SOX11 in lymphoid neoplasms by histone modifications. PLoS One 2011;6(6):e21382.
6. Ammerpohl O, Haake A, Pellissery S, Giefing M, Richter J, Balint B, Kulis M, Le J, Bibikova M, Drexler HG, Seifert M, Shaknovic R, Korn B, Küppers R, **Martín-Subero JI**+, Siebert R+. Array-based DNA methylation analysis in classical Hodgkin lymphoma reveals new insights into the mechanisms underlying silencing of B cell-specific genes. Leukemia 2012 Jan;26(1):185-8.
7. **Martin-Subero JI**, et al. A comprehensive microarray-based DNA methylation study of 367 hematological neoplasms. PLoS One 2009 Sep 11;4(9):e6986. doi
8. **Martín-Subero JI**, et al. New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. Blood 2009 Mar 12;113(11):2488-97.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Population cancer genomics

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Simon C Heath, Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain

Sergi Beltran, Centro Nacional de Análisis Genómico, Barcelona, Spain

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
	Tomas Marques-Bonet, Universitat Pompeu Fabra, Barcelona, Spain **Arcadi Navarro, Universitat Pompeu Fabra, Barcelona, Spain**

Background and preliminary data

The genomes of cancer cells are the product of intense selective pressures for proliferative ability acting at the level of the populations. The pan-cancer study provides the unique opportunity to study many independent realizations of this selection process in different cancers against comparable genetic backgrounds by using standard approaches that have been available for a long time to the population genetics community. For instance, by analyzing changes in allele and haplotype frequencies or novel mutational sites and motifs across different cancer genomes from a population genetics perspective, it will be possible to identify new footprints of selection common across different samples, both within and between cancer types.

In this proposal, we would like to investigate three aspects in particular: (a) to test whether somatic mutations in different genomic regions tend to occur early or late in oncogenesis by making use of the distribution of alternate allele supporting reads of somatic variants as a proxy for cellularity and use ratios of this information to differentiate selective sweeps; (b) to look for evidences of different mutational and selective pressures in somatic (cancer genomes) vs. germ line mutations (normal population. e.g. 1KG, HapMap) by using a combination of the distribution of allele variants and typical summary statistics in populations genomics (Fst, Tajima´s D, etc..) and of methods based on haplotype properties (EHH, iHS, etc...). (c) look for particular genomic regions with particular allele co-occurrences in somatic mutations (either positive or negative) indicating synergistic or antagonistic relationships;

Timelines & resources dedicated to project

Our analyses require access to the genotype calls for all sites (not only the somatic variable sites) in both the normal and tumor samples, with information on the number of reads supporting the normal and variant allele and indications of whether the site passes the filters to be considered as a germline or somatic variant. When possible, progression cancer genotypes will be ideal, with information about cellularity and time before and after the treatment.

3-4 months should be requested for the analysis, and we estimate we can have a full report for the group within 6 months.

Research proposal

In this proposal, we would like to investigate different mutational aspects across different cancers by comparing somatic and germline variants using population genetic tools. Specifically, we would like to use the alternate allelic ratio (the proportion of reads supporting the variant allele) in somatic variants as a proxy of cellularity and selection sweeps or explore standard statistics to pinpoint specific genomic regions that might be mutational hotspots in cancer. Finally, we will look at both, over-represented and depleted co-occurrence of alleles in somatic mutations that might be proxy for specific epistatic interactions.

a) *Use of allele read frequency information to infer selective sweeps and the chronology of somatic variants during oncogenesis.* Tumor samples can contain a mixture of cells with different sub-clones of the tumor as well as normal cells. The proportion of reads supporting a somatic variant (RSSV) provides information on whether a particular mutation occurred early or late in the oncogenesis process, as an early arising mutation will be seen in all or almost all sub-clones and so should have a high variant allele ratio, and conversely a late arising mutation would be expected to have a lower variant allele ratio; this could help to differentiate between driver and passenger mutations. The estimate of the time that a mutation occurred will be biased by the presence of normal cells in the sample, so this will be accounted for in the analysis. For individual somatic variations the estimate will be very noisy, particularly at low coverage, but by combining information across all cancer samples it should be possible to see whether somatic mutations occurring in particular genomic regions tend to occur early or late, which will provide important information on the role of these mutations in oncogenesis. On top of that, we can make use of ratios of RSSV across sites (when possible) to differentiate “soft sweeps” after relapse of treatment than spurious somatic sites. The reason being that a selective sweep should homogenize the proportion of RSSV across sites, whereas the standard cell progression will leave a high heterogeneous rate of RSSV.

b) *Comparison of the selective and mutational processes in germline vs. somatic mutations.* Study of the genomic location, distribution and footprints of selection using both, statistics based on the allele frequency spectrum of sequence variants and statistics based on haplotype distributions, to compare the distribution of normal genomic variation (from the germline variants in the pan-cancer study as well as information from public sources such as the 1000 Genomes Project) with that seen in the somatic variants from all samples. This will allow the identification of genomic regions that are under different differential mutational (we will apply a machine learning approach to detect hyper-represented motifs) or selective pressures in both kind of samples, potentially indicating novel regions and features that might be relevant in many different cancer types.

c) *Correlation in the presence/absence of somatic variants between genomic regions*. We will look for evidence that somatic mutations in two regions co-occur or, conversely, that somatic mutations are rarely seen simultaneously in several regions in the same sample. Evidence of high negative or positive correlation could indicate that certain combinations of mutations are favored by epistasis or selection. To perform this analysis

we will require genotype information on all sites, whether or not a somatic variant has been called in a particular sample.

Legacy plans

The findings together with the developed methodology will be written up for publication and submitted as a companion paper to the Pan-Cancer flagship paper.

Dr. Simon Heath joined the Rockefeller University as an Assistant Professor in the Laboratory of Statistical Genetics in 1997, and then moved to Sloan Kettering Cancer Center as an Assitant Member in the Department of Human Genetics in 1999. He then moved to the Centre National de Genotypage (CNG) in Evry, France in 2002, where he was the head of the Statistical Genetics group, and became the head of the Laboratory of Statistics and Informatics at the CNG in 2008. In 2010 he moved to the Centro Nacional de Análisis Genómico (CNAG) in Barcelona where he has been the Group Leader of the Bioinformatics Development group and Team Leader of the Statistical Genomics team since the setting up of the CNAG. His research interests have included linkage analysis for complex genetic traits in large isolated populations, efficient methods for genome wide association and, more recently, computational and statistical methods for the efficient and robust analysis of next generation sequencing data. He has participated in many European and international genomics consortia, with current memberships including ICGC, IHEC and BLUEPRINT. His group consists of 1 staff scientist, 2 team leaders, 6 post-doctoral researchers, 2 PhD students and a bioinformatics technician. He has >130 peer-reviewed articles and 2 book chapters that have been cited >15,000 times and has an H-index of 54.

SELECTED PUBLICATIONS

1. Balbás-Martínez C, …, **Heath S**, Valencia A, Losada A, Gut I, Malats N, Real F (2013) Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. Nat. Genet. 45:1464-69.
2. Kulis, M, **Heath S**, …, Martin-Subero I (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat. Genet. 44(11):1236-42.
3. Heyn H,…, **Heath S,** Valencia A, Gut I, Wang J, Esteller M (2012) Distinct DNA methylomes of newborns and centenarians. PNAS 109(26).
4. Quesada V, …, **Heath S**, Gut M, Gut I, Estivill X, López-Guillermo A, Puente X, Campo E, López-Otín C (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat. Genet. 44(1):47-52.
5. Puente XS, Pinyol M, Quesada V, …, **Heath S**, …., Campo E. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 5;475(7354):101-5.
6. Purdue MP, Johansson M, …, **Heath S**, …, Lathrop M, Brennan P. (2011) Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. Nat Genet. 43(1):60-5.
7. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, **Heath S**, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. (2010) A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 363(13):1211-21.
8. Lambert JC, **Heath S**, …, Lathrop M, Amouyel P. (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 41(10):1094-9.
9. **Heath S**, Gut I, Brennan P, …, Cardon L, Lathrop, M. (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet. 16(12):1413-29.

Dr. Sergi Beltran is the Bioinformatics Analysis Group Leader at the Centro Nacional de Análisis Genómico (CNAG) since 2012. The group is devoted to the implementation and operation of sequencing data analysis and management pipelines, including identification of germline and somatic mutations. Currently, the group is formed by 3 postdocs, 1 manager and 5 technicians divided in two operational teams. Before joining the CNAG, Sergi worked at the Transcriptomics Platform of the Barcelona Science Park in 2005 and from 2006 he was a Postdoc at the Bioinformatics and Genomics Research Group at the Center for Genomic Regulation (CRG). From 2008 to 2011 he setup and managed the Bioinformatics Unit at the Scientific and Technological Centers of the University of Barcelona. Sergi has published a total of 15 peer-reviewed publications.

SELECTED PUBLICATONS

1. Lappalainen, Tuuli; Sammeth, Michael; Friedländer, Marc R et al. including **Beltran, Sergi** *"Transcriptome and genome sequencing uncovers functional variation in humans."* Nature 2013
2. Balbás-Martínez, Cristina; Sagrera, Ana; Carrillo-de-Santa-Pau, Enrique; Earl, Julie; Márquez, Mirari; Vazquez, Miguel; Lapi, Eleonora; Castro-Giner, Francesc; **Beltran, Sergi**; Bayés, Mònica et al. "Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy." Nature Genetics 2013
3. Estarás, Conchi; Fueyo, Raquel; Akizu, Naiara; **Beltrán, Sergi**; Martínez-Balbás, Marian A; "RNA polymerase II progression through H3K27me3-enriched gene bodies requires JMJD3 histone demethylase." Molecular Biology of the Cell 2013
4. Palstra, Arjan P; **Beltran, Sergi**; Burgerhout, Erik; Brittijn, Sebastiaan A; Magnoni, Leonardo J; Henkel, Christiaan V; Jansen, Hans J; van den Thillart, Guido EEJM; Spaink, Herman P; Planas, Josep V; "Deep RNA Sequencing of the Skeletal Muscle Transcriptome in Swimming Fish." PloS One 2013
5. Estarás, Conchi; Akizu, Naiara; García, Alejandra; **Beltrán, Sergi**; de la Cruz, Xavier; Martínez-Balbás, Marian A; "Genome-wide analysis reveals that Smad3 and JMJD3 HDM co-activate the neural developmental program." Development 2012
6. Pereiro, Patricia; Balseiro, Pablo; Romero, Alejandro; Dios, Sonia; Forn-Cuni, Gabriel; Fuste, Berta; Planas, Josep V; **Beltran, Sergi**; Novoa, Beatriz; Figueras, Antonio; "High-throughput sequence analysis of turbot (Scophthalmus maximus) transcriptome using 454-pyrosequencing for the discovery of antiviral immune genes." PLoS One 2012
7. Yúfera, Manuel; Halm, Silke; **Beltran, Sergi**; Fusté, Berta; Planas, Josep V; Martínez-Rodríguez, Gonzalo; "Transcriptomic characterization of the larval stage in gilthead seabream (Sparus aurata) by 454 pyrosequencing," Marine Biotechnology 2012
8. Blanco, Enrique; Pignatelli, Miguel; **Beltran, Sergi**; Punset, Adrià; Pérez-Lluch, Silvia; Serras, Florenci; Guigó, Roderic; Corominas, Montserrat. "Conserved chromosomal clustering of genes governed by chromatin regulators in Drosophila." Genome Biology 2008
9. **Beltran, Sergi**; Angulo, Mireia; Pignatelli, Miguel; Serras, Florenci; Corominas, Montserrat. "Functional dissection of the ash2 and ash1 transcriptomes provides insights into the transcriptional basis of wing phenotypes and reveals conserved protein interactions." Genome Biology 2007
10. **Beltran, Sergi**; Blanco, Enrique; Serras, Florenci; Pérez-Villamil, Beatriz; Guigó, Roderic; Artavanis-Tsakonas, Spyros; Corominas, Montserrat. *"Transcriptional network controlled by the trithorax-group gene ash2 in Drosophila melanogaster."* **Proceedings of the National Academy of Sciences 2003**

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Structural analysis of identified protein variants in pan-cancer exomes.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Marc A. Marti-Renom, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain
Ivo Gut, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Francisco Martínez-Jiménez (CNAG)	

Background and preliminary data

Over the last decade there has been an increase in application of classical comparative structure prediction methods with information theory classifiers to identify the likely effect of variants in coding protein genes in cancer. We used this structure-based analysis of missense mutations in BRCA1 gene in breast cancer to predict likely mutations that were associated to malignancy [1]. Our initial work was based on a decision-tree that required extensive manual intervention. This was later addressed by a continuation of our work using a supervised learning algorithm [2]. Such methods have now become routine in many prediction applications where the goal is to assess the likely effect of a SNP in the structure of the coding protein [3].

Here we propose to systematically apply structure prediction methods to identify the likely drivers of cancer based on somatic variant calls of the pan-cancer study.

[1] Mirkovic, N., Marti-Renom, M. A., Weber, B. L., Sali, A., & Monteiro, A. N. A. (2004). Structure-based assessment of missense mutations in human BRCA1: implications for breast and ovarian cancer predisposition. Cancer research, 64(11), 3790–3797. doi:10.1158/0008-5472.CAN-03-3009

[2] Karchin, R., Monteiro, A. N. A., Tavtigian, S. V., Carvalho, M. A., & Sali, A. (2007). Functional impact of missense variants in BRCA1 predicted by supervised learning. PLoS Computational Biology, 3(2), e26. doi:10.1371/journal.pcbi.0030026

[3] Carter, H., & Karchin, R. (2013). Predicting the Functional Consequences of Somatic Missense Mutations Found in Tumors. Gene Function Analysis, 1101(Chapter 8), 135–159. doi:10.1007/978-1-62703-721-1_8

Timelines & resources dedicated to project

After getting access to the exome somatic variant data of tumor/non-tumor pairs of the pan-cancer study, we will model 3D structures of coding proteins. Based on estimations of the numbers of amino acid altering somatic mutations, the modeling effort will take approximately 2-3 months of computational time using about 100 CPUs from CNAG's cluster. Prior to receiving the calls, Francisco Martínez-Jiménez (a PhD at the Structural Genomics Team of the CNAG) will implement and automate the modeling pipeline. In parallel, he will also develop the computational Support Vector Machine (SVM) tool that will determine whether a given mutation is associated or not to malignancy in a given tumor type. This second step will not require large amounts of computational time and is likely to be done within few days of calculation in our cluster using about 100 CPUs. Finally, the collected mutation data will be placed in a database and mined to identify likely targets with pan-cancer associations.

Research proposal

We will use the data from the pan-cancer tumor and normal exome pairs, together with clinical data, to build 3D models of all mutated (with called somatic variants) proteins. This computational analysis will consist of four main steps:

(i) Comparative modeling of as many protein as possible amino acid mutated exomes in relation to wild-type proteins

(ii) Structural analysis of all mutants by calculating from the 3D models the following features, which will train our SVM for predicting likely malignant variants.

a. Buriedness. Mutation of a more exposed residue is less likely to change the structure and therefore its function.
b. Functional Site. A mutation in a functional site is more likely to abolish function.
c. Residue and Neighborhood Rigidity. A mutation of a less rigid residue at a buried position is less likely to change the structure and therefore its function.
d. Volume Change. A large change in the volume of the amino acid residue type was considered destabilizing to the structure and therefore its function, especially when buried at a rigid position.
e. Charge Change. Mutations corresponding to changes in charge at buried positions are more likely to change the structure and therefore its function.
f. Polarity Change. Mutations corresponding to changes in polarity at buried positions are more likely to change the structure and therefore its function.
g. Mutation Likelihood. Unlikely mutations are more likely to change the structure and function.
h. Phylogenetic entropy. Mutations in evolutionary conserved positions are likely to change the structure and function.
i. Helix/Turn Breakers. Mutations in breaking helices or turns are likely to change the structure and function.

(iii) The trained SVN will be then used to predict likelihood of a mutation to be associated with a negative effect. This will constitute a structurally rich database of variants in tumor versus normal cells.

(iv) The database will be mined to identify likely drivers of disease that act across several tumor types. The clinical data will also be mined to identify early onset during the disease of such variants.

Legacy plans

We will code a ready-to-use Python library for all the steps in our proposal so that the pipeline will take as input a list of proteins and their associated variants in a given tumor and return a list of likely drivers of the disease. We will also generate a database with the structures. Results will be written up and submitted as a companion to the Pan-Cancer flagship paper.

Marc A. Marti-Renom, ICREA Research Professor http://marciuslab.org
Group Leader
Structural Genomics Team. Genome Biology Group. tel +34 934 020 542
National Center for Genomic Analysis (CNAG) fax +34 934 037 279
c/ Baldiri Reixac, 4. PCB - Tower I, 2nd floor mmarti@pcb.ub.cat
08028 Barcelona, Spain

Group Leader
Structural Genomics Group. tel +34 933 160 100
Gene Regulation, Stem Cells and Cancer Program. fax +34 933 160 099
Centre for Genomic Regulation (CRG) marc.marti-renom@crg.eu
c/ Dr. Aiguader, 88
08003 Barcelona, Spain

Since June 2006, I have led my own research group (http://marciuslab.org) first at the CIPF (Valencia, Spain) and later (January 2012) at both the National Center for Genomic Analysis (CNAG, http://cnag.cat) and the Centre for Genomic Regulation (CRG, http://crg.cat) where I have a joint appointment. Since October 2013, I am ICREA research professor. The mission of our group is to develop and use experimental and computational approaches for characterizing the molecular regulation of cells by studying the structure of macromolecules and their complexes. In particular, we focus on regulatory molecules such as proteins, RNA and chromatin. Given the limited knowledge on how these molecules fold and function, there is a great opportunity to spearhead beyond the state-of-the-art in these fields. Our research has resulted so far in over 70 peer-reviewed articles, 14 book chapters or invited reviews, and over 90 oral presentations in national and international venues. Currently, I am Associate Editor of PLoS Computational Biology. Since 2011, I coordinate two international teams funded by the EU (Era-Net Pathogenomics Grant) and the HFSP (Research Grants Award).

SELECTED PUBLICATIONS

1. F. Martínez, G. Papadatos, I. Yang, I.M. Wallace, V. Kumar, U. Pieper, A. Sali, J.R. Brown, J.P. Overington, and **M.A. Marti-Renom***. "*Target identification for active and open access compounds against Tuberculosis*" PLoS Computational Biology (2013) **9(10)**:e1003253.
2. C.U. Köser, J.M. Bryant, J. Becq, M.E. Török, M.J. Ellington, **M.A. Marti-Renom**, A.J. Carmichael, J. Parkhill, G.P. Smith & S.J. Peacock. "*Whole-genome sequencing for rapid identification, antimicrobial susceptibility testing, and typing of extensively-drug resistant Mycobacterium tuberculosis*" New England Journal of Medicine (2013) **369(3)**:290-292
3. J. Dekker, **M.A. Marti-Renom** and L. Mirny, "*Exploring the three-dimensional organization of genomes*". Nature Review Genetics (2013) **14(6)**:390–403
4. M.A. Umbarger, E. Toro, M.A. Wright, G.J. Porreca, D. Baù, S-H. Hong, M.J. Fero, J. Zhu, **M.A. Marti-Renom***, J.H. McAdams, L. Shapiro, J. Dekker and G.M. Church "*The three-dimensional architecture of a bacterial genome*". Molecular Cell (2011) **44**:252–264.
5. D. Baù, A. Sanyal, B.R. Lajoie, E. Capriotti, M. Byron, J.B. Lawrence, J. Dekker, and **M.A. Marti-Renom*** "*The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules*". Nature Structural and Molecular Biology (2011) **18**:107-115.

Up-to-date information at http://marciuslab.org/marti-renom

IVO GUT is **Director of the Centro Nacional de Análisis Genómico (CNAG)** in Barcelona, one of the largest European genome sequencing operations, which he established in 2010. He received his **PhD in Physical Chemistry** from the University of Basel in 1990. His post-doctoral work was at Harvard Medical School and the Imperial Cancer Research Foundation of London, he led a group in the Department for Vertebrate Genomics at Max-Planck-Institute for Molecular Genetics in Berlin.
For the 11 years prior to CNAG he was at the Centre National de Génotypage (CNG) – CEA as Associate Director and in charge of Technology Development. He initiated and was coordinator of the 16 MEuro EU-funded project READNA on nucleic acid technology development. His research interests are genomics, high-throughput nucleic acid analysis methods, proteomics, omics technologies, automation, bioinformatics, disease gene identification, cancer, and agrogenomics. He has more than 20 years experience and has authored more than 170 research papers, 11 reviews and 12 book chapters, cited over 15000 times. He is inventor of 25 patents or patent applications, founder of 4 biotechs, and serves on numerous international advisory boards. Significant achievements are the development of pioneering methods for DNA analysis by mass spectrometry (genotyping, sequencing, DNA methylation analysis and haplotyping), bisulphite Pyrosequencing, TAMSIM a Imaging Mass Spectrometry method, execution of many of the major GWAS and ICGC studies.

SELECTED PUBLICATIONS

1. Prado-Martinez J, Sudmantet P H, Kidd J M, … **Gut I G**, Eichler E E, Marques-Bonet T. (2013). Great ape genetic diversity and population history. Nature 499(7459): p. 471-5.
2. Kulis M., Heath S, Bibikova M, … **Gut I**, C. Lopez-Otin, E. Campo and J. I. Martin-Subero (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44(11): 1236-1242.
3. Narni-Mancinelli, Jaeger EBN, Bernat C, … **Gut I G**, E. Vivier and S. Ugolini (2012). Tuning of natural killer cell reactivity by NKp46 and Helios calibrates T cell responses. Science 335(6066): 344-348.
4. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, Eick D, **Gut I**, Ferrier P, Andrau JC. (2011). . Nat Struct Mol Biol. 17;18(8):956-63.
5. Puente XS, Pinyol M, Quesada V, … **Gut I**, López-Guillermo A, Estivill X, Montserrat E, López-Otín C, Campo E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 5;475(7354):101-5.
6. Moffatt MF, **Gut I G**, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. (2010). A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 363(13):1211-21.
7. Lambert JC, Heath S, … **Gut I**, Van Broeckhoven C, Alpérovitch A, Lathrop M, Amouyel P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 41(10):1094-9.
8. Bishop DT, Demenais F, … **Gut I**, …., Lathrop GM, Barrett JH, Bishop JA. (2009). Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 41(8):920-5.
9. Hung RJ, McKay JD, … **Gut I**, … Brennan P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452: 633-637.
10. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, **Gut I**, Lathrop M and Collins R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 359, 789-799.
11. Moffatt MF, Kabesch M, Liang L, … **Gut I G**, Lathrop GM, Cookson WO. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 448: 470-473.

Francisco Martinez Jiménez
PhD Student
Structural Genomics Team. Genome Biology Group.
National Center for Genomic Analysis (CNAG)
c/ Baldiri Reixac, 4. PCB - Tower I, 2nd floor
08028 Barcelona, Spain

http://marciuslab.org
tel +34 934 020 542
fax +34 934 037 279
fmartinez@pcb.ub.cat

Structural Genomics Group.
Gene Regulation, Stem Cells and Cancer Program.
Centre for Genomic Regulation (CRG)
c/ Dr. Aiguader, 88
08003 Barcelona, Spain

tel +34 933 160 100
fax +34 933 160 099
Francisco.Martinez@crg.eu

From 2006 until 2011 I studied a degree in Computer Science in the Universidad Complutense de Madrid (Madrid, Spain) where I did my dissertation project "*Generic architecture for Web services based on telemedicine and Google Health. Practical application in diabetes*" in collaboration with the Hospital de Toledo. In 2011, due to my interest in bioinformatics and telemedicine, I started a master degree in Bioinformatics in the Universidad Complutense de Madrid (2011- 2012). The master final thesis entitled "*Ligand-Target prediction by comparative docking*" was carried out in CNAG/CRG (Barcelona, Spain) from May until July 2012 supervised by Dr. Alfonso Valencia and Dr. Marc A. Marti-Renom. In September 2012, I started a PhD working in computational drug discovery in the Structural Genomics team led by Dr. Marc A. Marti-Renom. During the last year and a half, I have published two peer-reviewed articles and I have been working as an Assistant Teacher in the Bioinformatics Master of the Universidad de Valencia.

SELECTED PUBLICATIONS

1. **F. Martínez**, G. Papadatos, I. Yang, I.M. Wallace, V. Kumar, U. Pieper, A. Sali, J.R. Brown, J.P. Overington, and **M.A. Marti-Renom.** "*Target identification for active and open access compounds against Tuberculosis*" PLoS Computational Biology (2013) **9(10)**:e1003253.
2. López-Pelegrín, M., Cerdà-Costa, N., **Martínez-Jiménez, F.**, Cintas, A., Canals, A., Peinado, J.R., **Marti-Renom, M.A.**, López-Otín, C., Arolas, J.L. and Gomis-Rüth, F.X. *"A novel family of soluble minimal scaffolds provides structural insight into the catalytic domains of integral-membrane metallopeptidases."*. **JBC** (2013) **288** 21279–21294.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

GEM-based mapping pan-cancer pipeline

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Paolo Ribeca, Centro Nacional de Analisis Genomico, Barcelona, Spain
Ivo Gut, Centro Nacional de Analisis Genomico, Barcelona, Spain

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

Most alignment pipelines used within the ICGC are based on heuristic non-deterministic mappers like BWA. In order to increase their performance they adopt a series of plausibility criteria, and discard on the basis of rule-of-thumb considerations part of the huge search space that needs to be explored. As a result, given any sequencing read they are not guaranteed to return all the “good” alignments that exist for it within the search parameters specified by the user (and this behavior might sometimes be non-deterministic, i.e. the program might return different results depending on the run). The fact that there might be good alignments skipped for each query implies that such programs are not able, for instance, to determine whether a sequencing read is unique or not. They will also produce unreliable results when a read is repetitive and matches many locations in the genome, which can often occur in the case of cancer data. Finally, they usually employ hard-wired parameters (for instance, they usually return some set of “best” alignments which is arbitrarily selected by the program) that cannot be tuned by the user, making them a poor choice for high-precision studies.

The GEM mapper (S.Marco-Sola et al., Nature Methods, December 2012) overcomes most of those problems, providing full deterministic searches, full configurability and, on the top of that, excellent speed. Even when reporting all existing matches the GEM mapper is several times faster than BWA or Bowtie2 run in heuristic modes.

Timelines & resources dedicated to project

The project will require an initial phase, where the specific requirements of the ICGC PanCancer pipeline with respect to alignment are thoroughly discussed. This preliminary step would probably take a couple of months.

After this first phase, the actual implementation would be entirely carried out within the Algorithm Development unit of the CNAG. This is the team that originally designed and implemented the GEM mapper, as well as several pipelines for the analysis of high-throughput sequencing data based on it. Seen that the proposed project is very similar to what has already been done in the past for other international consortia, and the proponents have a consolidated experience at building data-analysis workflows, an alignment pipeline suitable to the ICGC PanCancer analysis could probably be made available in 1-2 months after the end of the preliminary definition stage.

Research proposal

The proposal focuses on designing and implementing a robust alignment pipeline based on the GEM mapper (S.Marco-Sola et al., Nature Methods, December 2012), to be used within the ICGC PanCancer analysis workflows. This will result in a more precise (deterministic, exhaustive) alignment, and in a consistent speedup.

The proposal focuses on designing and implementing a robust alignment pipeline based on the GEM mapper (S.Marco-Sola et al., Nature Methods, December 2012), to be used within the ICGC PanCancer analysis workflows. This will result in a more precise (deterministic, exhaustive) alignment, and in a consistent speedup.

Particular care will be taken to (1) determine and implement in the pipeline alignment parameters suitable to typical biological problems/protocols encountered while analyzing ICGC PanCancer data (2) determine and implement in the pipeline a read-scoring scheme suitable for downstream biological analysis (3) determine and implement in the pipeline a subset of the SAM/BAM format suitable to output all the information needed to downstream tools in a standard form.

The implementation will use in a straightforward way already available building blocks, mainly the GEMTools library (a high-throughput library for pipelining and post-processing the output of GEM mappers, https://github.com/gemtools/gemtools). The working plan above has already proven successful with similar large-scale collaborations where GEM-based pipelines have been used, for instance the GEUVADIS consortium (Lappalainen et al., Nature, September 26, 2013).

Legacy plans

An extensive documentation of alignment modes, quality scores and output format implemented in the pipeline will be written to be available to end-users.

In addition, a description of the method suitable to be embedded in publications making use of the pipeline will be provided.

Paolo Ribeca – Short CV

Research interests
A physicist by education, my research interests focus on the application of various disciplines of computer science, physics and mathematics to high-performance scientific computing in genomics and biology.
Since the inception of high-throughput sequencing techniques, I specialized on algorithms for short-read processing. I am the main architect of GEM, a software package for the analysis of short sequence reads produced by modern high-throughput sequencers (http://gemlibrary.sourceforge.net).

Current position
I currently lead the Algorithm Development unit at the Centro Nacional de Análisis Genómico (CNAG), the Spanish National sequencing center. My group provides bleeding-edge algorithmic methods, to make possible the precise and timely analysis of the ~1Tb of genomic data produced by the CNAG every day.
Apart from myself, the group includes a postdoctoral fellow and two Ph.D. students whose I am the supervisor.
Our methods for short-read alignment (in particular the GEM mapper, S.Marco-Sola et al. Nature Methods, December 2012) are considered to be among the fastest and most accurate tools in their class, and are being adopted by an increasing number of labs and consortia.
The group is actively involved in many bleeding-edge biological projects, typically focusing on either de-novo genome assembly, or on the analysis of cell expression regulation by RNA-sequencing.
Finally, I am visiting fellow at the Pirbright Institute (formerly Institute for Animal Health), UK, one of the top institutions in the world in animal virology.

Participation in projects/consortia
2011-present. The GEUVADIS (Genetic European Variation in Disease) consortium
2010-present. The Iberian Lynx sequencing consortium
2009-2012. The ENCODE (Encyclopedia of DNA Elements) consortium
2009-2011. The Tomato sequencing consortium.

Publications
20 publications so far (18 peer-reviewed papers [including 4 Nature, 3 Nature Methods and 1 Nature Biotechnology], plus 2 book chapters).
They accumulate ~1900 citations to-date (h-index 11). For more details see http://scholar.google.com/citations?user=juUtwYAAAAAJ&hl=en.

Additional relevant information
I act as peer-reviewer for several journals and conferences (in particular, I have been nominated leading reviewer'' by Bioinformatics.)
I am often invited speaker at conferences and schools. I am one of the teachers of the Master in Bioinformatics of the University of Murcia.

Previous jobs
2008-2010. Post-doctoral position at the Center for Genomic Regulation (CRG), Barcelona, in the Bioinformatics and Genomics group.
2006-2008. Post-doctoral position at the Center for Genomic Regulation (CRG), Barcelona, in the Systems Biology group.
2006. Invited scientist at GSI/Darmstadt and Max-Planck Institut für Kernphysik/Heidelberg.
2003-2005. Post-doctoral position at the Humboldt Universität, Berlin, in the Computational Physics group.

IVO GUT is **Director of the Centro Nacional de Análisis Genómico (CNAG)** in Barcelona, one of the largest European genome sequencing operations, which he established in 2010. He received his **PhD in Physical Chemistry** from the University of Basel in 1990. His post-doctoral work was at Harvard Medical School and the Imperial Cancer Research Foundation of London, he led a group in the Department for Vertebrate Genomics at Max-Planck-Institute for Molecular Genetics in Berlin.
For the 11 years prior to CNAG he was at the Centre National de Génotypage (CNG) – CEA as Associate Director and in charge of Technology Development. He initiated and was coordinator of the 16 MEuro EU-funded project READNA on nucleic acid technology development. His research interests are genomics, high-throughput nucleic acid analysis methods, proteomics, omics technologies, automation, bioinformatics, disease gene identification, cancer, and agrogenomics. He has more than 20 years experience and has authored more than 170 research papers, 11 reviews and 12 book chapters, cited over 15000 times. He is inventor of 25 patents or patent applications, founder of 4 biotechs, and serves on numerous international advisory boards. Significant achievements are the development of pioneering methods for DNA analysis by mass spectrometry (genotyping, sequencing, DNA methylation analysis and haplotyping), bisulphite Pyrosequencing, TAMSIM a Imaging Mass Spectrometry method, execution of many of the major GWAS and ICGC studies.

SELECTED PUBLICATIONS

1. Prado-Martinez J, Sudmantet P H, Kidd J M, … **Gut I G**, Eichler E E, Marques-Bonet T. (2013). Great ape genetic diversity and population history. Nature 499(7459): p. 471-5.
2. Kulis M., Heath S, Bibikova M, … **Gut I**, C. Lopez-Otin, E. Campo and J. I. Martin-Subero (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44(11): 1236-1242.
3. Narni-Mancinelli, Jaeger EBN, Bernat C, … **Gut I G**, E. Vivier and S. Ugolini (2012). Tuning of natural killer cell reactivity by NKp46 and Helios calibrates T cell responses. Science 335(6066): 344-348.
4. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, Eick D, **Gut I**, Ferrier P, Andrau JC. (2011). . Nat Struct Mol Biol. 17;18(8):956-63.
5. Puente XS, Pinyol M, Quesada V, … **Gut I**, López-Guillermo A, Estivill X, Montserrat E, López-Otín C, Campo E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 5;475(7354):101-5.
6. Moffatt MF, **Gut I G**, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. (2010). A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 363(13):1211-21.
7. Lambert JC, Heath S, … **Gut I**, Van Broeckhoven C, Alpérovitch A, Lathrop M, Amouyel P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 41(10):1094-9.
8. Bishop DT, Demenais F, … **Gut I**, …., Lathrop GM, Barrett JH, Bishop JA. (2009). Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 41(8):920-5.
9. Hung RJ, McKay JD, … **Gut I**, … Brennan P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452: 633-637.
10. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, **Gut I**, Lathrop M and Collins R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 359, 789-799.
11. Moffatt MF, Kabesch M, Liang L, … **Gut I G**, Lathrop GM, Cookson WO. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 448: 470-473.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Chromosomal environment and mutational processes in human cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia. Spanish National Cancer Research Centre (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Federico Abascal (CNIO)	

Background and preliminary data

Sequencing of thousands of cancer genomes is providing an unprecedented accumulation of data on somatic mutation in cancer. Up to 21 mutational signatures have been identified in different types of tumors (Alexandrov et al, Nature 2013), each showing distinct levels of indel frequencies, CpG mutation, overall mutation rate, or displaying the footprint of transcription-coupled repair and/or APOBEC enzymatic activity. This kind of analysis aim to identify the underlying mutational mechanisms in some cancers, bettering our understanding of the distinct mutational processes acting in the cell. In addition, these analyses allow the identification of "over-mutated" genes, which can be considered as tumour-driver genes and potential drug targets. For instance, in (Kandoth et al, 2013 Nature) up to 127 genes were identified based on their frequency of mutation. **However, as shown in (Lawrence et al, Nature 2013) background mutation frequencies vary largely across the genome and must be taken into account to call out over-mutated genes.**

Genome comparisons between species and individuals of the same species have revealed that **mutation patterns vary and co-vary across the genome.** In (Kuruppumullage et al, PNAS 2013), the human genome was segmented according to distinct patterns of neutral evolution, resulting in **regions with different rates of insertions, deletions and substitutions.**

Genomic context influences the patterns of mutation (Abascal et al *in preparation*), and clear correlations between mutation and: GC-content (Smith et al, Genome Research 2002), replication timing (Stamatoyannopoulos, Nat Genetics 2009), gene density (Duret, PLoS Biology 2013) and recombination (Lercher and Hurst TIG, 2002) have been reported. Most of these **"context" features are related each other and, in turn, are related to chromatin and chromosome structure**. For instance, lamina associated regions of the genome tend to replicate lately, be gene poor and display low GC content. In contrast, subtelomeric regions tend to have high GC content, are highly recombinogenic and diverge faster (Brown et al, Curr Biol 2010). Interestingly, **despite many chromosomal rearrangements, the higher order organization of human and mouse genomes are very similar in terms of replication timing, Hi-C inter-locus interactions and lamina association** (Chambers et al, PLoS Comp Biol, 2013). The few structurally divergent regions are associated to higher levels of divergence, large GC-content shifts and tend to be found at subtelomeric regions. **Genomic context might also determine the evolutionary fate of gene duplicates (Abascal et al, MBE, 2013)**.

Regarding somatic mutation **in cancer, it has been observed that regions with extreme levels of mutation (*kataegis*) usually occur at chromosomal breakpoints (Alexandrov et al, Naure 2013), and that mutation and replication timing co-vary (Woo and li, Nature Communications, 2012).** In (Lawrence et al, 2013 Nature) across genome mutation heterogeneity is taken into account to estimate over mutated genes, resulting in more trustable lists of candidate driver genes, and **finding that the level of gene expression and the replication timing explain a large fraction of the observed heterogeneity.**

Timelines & resources dedicated to project

We will apply HMM-based segmentation techniques to PanCancer complete genomes to identify regions with distinctive patterns of mutation at different resolutions.

We will estimate mutation rates, types of nucleotide substitution, frequency of insertions and deletions, and several trends including transition/transversion, homozigosity/heterozygosity, and strand-specific biases using as input the variant and structural calls provided by the PanCancer consortium.

Segmentations will be applied for individual samples or groups of samples (e.g. by tumor type or by type of tissue).

For germline mutational signatures, we will estimate levels of variation at the population level taking data from the 1000 Human Genomes Project, and, at the species level, based on human-chimpanzee genome comparisons.

Resulting genome segmentations will allow different comparison settings: 1) across different regions of the genome; 2) between different samples of the same type of cancer; 3) between different cancer types; and 4) under several germline-somatic comparative settings.

One major factor influencing mutation is structural variation (SV). **We propose to conduct segmentations based on raw data as well as on SV-free regions of the genome (as identified in essential research lines of the project), to assess the impact of SV on the identified signatures, and to better understand what has been referred to as *kataegis*.**

Research proposal

Despite the clear influence of genomic context on mutational patterns, our understanding is far from complete, especially in respect to the patterns of somatic mutation in cancer. We propose here to characterize mutational patterns at genome and chromosome levels, and to compare the germline and somatic mutational patterns. Evidence suggests that higher GC-content and divergence rate at subtelomeric regions are related to higher recombination rates in the germline. **We would like to explore to what extent and under which circumstances germline mutational patterns (either at population or species level) can be extrapolated to somatic mutation. This analysis will uncover differences and similarities that will better our understanding of somatic mutation in cancer.**

To characterize different mutational signatures across the genome we propose to use genome segmentation techniques in combination with variation information on tumors and normal (and related-species) samples, to **estimate mutation rates, types of nucleotide substitution, frequency of insertions and deletions, and several trends including transition/transversion, homozigosity/heterozygosity, and strand-specific biases** (the latter has been explained in terms of transcription-coupled repair).

The goal of this proposal is to increase our understanding of the mutational processes acting in cancer, and to improve our capacity to identify candidate positively selected genes, i.e. genes that have mutated more than their background. The proposal fits in the general PanCancer topics of: *integration of genome and transcriptome, integration of genome and epigenome* and *interface between germline and somatic genetics.*

Legacy plans

The methods proposed here are based on publicly available as free software or R packages. Most of the calculations are already performed as part of the collaboration of the group in cancer genome projects. As such we maintain this expertise, and the associated tools, as a essential internal resource.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Federico Abascal, PhD

Present work (since October 2011)

Staff Scientist. Structural Computational Biology Group.
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.
Phone: (+34) 917328000; email: fabascal@cnio.es

Positions held:

- **July 2009-September 2010.** Postdoctoral researcher at National Natural History Museum, with Prof. Rafael Zardoya. Department of Biodiversity. Madrid, Spain.

- **September 2006-June 2009**. I3P postdoctoral researcher at National Biotechnology Centre, with Prof. Carazo and Dr. Pascual-Montano. Biocomputing Unit. Madrid, Spain.

- **March 2004–March 2006.** Postdoctoral researcher at Universidad de Vigo/National Natural History Museum, with Profs. David Posada and Rafael Zardoya. Vigo, Spain.

PhD degree:

Ph.D. in Molecular Biology, *Universidad Autónoma de Madrid* (Spain), November 18th 2003. Dissertation title: "Genome Analysis. Methods for automatic prediction and annotation of protein function". Qualification: Excellent *Cum Laude*. Supervisor: Prof. Alfonso Valencia.

Specialization:

Main field: sequence analysis; genome analysis; evolutionary biology.
Other fields: gene expression analysis; epigenetics; data-mining.
Current research interest: genome analysis; transposable elements; GC content and gene evolution; piRNAs and other mechanisms of genome defense.

Overview of Achievements:

Author of 16 peer-reviewed articles (14 first-authorships), 2 book chapters and a book for the popularization of science: PLoS Biology, Nuc Ac Res, Bioinformatics, Mol Biol Evol, Proteins, PNAS, BMC Genomics, and others. Postdoctoral I3P fellowship. Regular teaching activities in Bioinformatics. Regular reviewer for: Bioinformatics, Mol Biol Evol, Genome Biology and Evolution, Systematic Biology, BMC Bioinformatics, BMC Genomics, BMC Evolutionary Biology, Gene, JME…

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Pan-cancer Pharmacogenomics

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia, Spanish National Cancer Research Institute (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Fatima Al-Shahrour, (CNIO)	

Background and preliminary data

Previous studies have suggested that in ~40-60% of the cases in many solid tumors the genomic information points at least one alteration that might influence therapeutic decision-making (Garraway et al Cell 2013). Still, up to now it has not been **systematically evaluates the impact of cancer genomics on therapeutic strategies**.

A number of studies have associated selected functional alterations with: a) clinical features of prognosis value (Kandoth et al Nature 2013) or b) with a limited number of drug targetable pathways (Ciriello et al. Nat Genet 2013). Although almost all samples present alterations in at least one of those "targetable" pathways and therefore, can be interpreted in terms of suggested therapeutic candidates for most cases, this approach is not comprehensive enough to select the most appropriate treatment regimen for each patient.

Here, **we propose to identify and clinically annotate a wide range of actionable genomic alterations with their associated targeted therapy for PanCancer donor**.

We have already developed a computational methodology that analyzes individual **cancer genomes** and **prioritizes individualized therapy** instead of general therapeutically actionable alterations for tumor types. This method has already applied to our CNIO internal personalize medicine initiative. For the more than 20 cases we have processed the **success rate of this strategy is of 20%** when applied to a PDX experimental models of individual tumors (manuscript in preparation).

Additionally, we have preliminary investigated this approach with available TCGA data integrating the analysis of individual cancer genomes data to identify putatively actionable tumor-specific genomic alterations. **Our integrative analysis includes as a new feature the oncogenic expression signatures associated to drug response**.

Here, we propose to apply this methodology to the PanCancer datasets to answer the following questions: **a) Identification of patients, and groups of patients, that could have been treated based on genomic data, b) prediction of drug synergistic effects and identification of new druggable pathways and c) comparison of standard tumor/tissue/key genes based classifications with the new horizontal case classification based on mutation patterns.**

Timelines & resources dedicated to project

Our proposal is based on the complete primary analysis of all tumor-normal matched samples, since we will data level 3 and clinical information. Most of the ICGC projects include donor clinical information such as stage, relapse, vital and disease status but the access to the donor's therapy information (name of therapy, duration and the clinical effect) is limited to the contributors. We might require this information for the retrospective analysis.

We have already applied our methodology to some of the TCGA data sets, so the pipeline will be available at the time PanCancer samples are available and the primary analysis completed.

Research proposal

We propose to contribute to the global PanCancer efforts with an essential project that **evaluates the impact of cancer genomics on clinical decision**. Cancer individualized treatment guidelines will be derived from the extensive study of drug therapies and genomic alterations, following our experience as part of the CNIO personalized cancer initiative.

Unfortunately, there are only few molecular predictors of efficacy for cancer drugs approved by the FDA. Recent genomics studies on cancer cell lines (CCLE and CTRP from Broad Institute and GDSC from Sanger) have proposed new gene-drug response associations providing new opportunities to expand the list of pharmacogenomics biomarkers. These studies are an important source of information (even considering the recent criticisms).

We have developed a **new computational methodology** that uses the existing information from CCLE, CTRP and GDSC resources to extrapolate **drug response using gene expression signatures**., as well as a catalog of the drugs associated to each case. This method calculates a global activation score associated to drug sensitivity for a set of oncogenic expression signatures. For this project, we'll use the samples with **WGS tumor-normal paired** integrating event types such as **mutations, copy number changes and epigenetic silencing with gene expression and methylation**.

Our pipeline would include:

a) To identify which **genes are genetically altered** (mutated, amplified or deleted, over or under-expressed and silenced) in each sample separately

b) To identify **gain or loss-of function mutations** taking into account their gene expression and methylated status and copy number variation.

c) To integrate the **oncogenic signatures activity** associated to the drug

d) To generate and refine a **comprehensive drug target (direct or indirect) catalogue** using existing public pharmacological databases and biomedical text-mining approaches

e) To develop a methodology that **prioritizes a list of druggable targets** based on patient's genomic analysis results from previous steps. If patients' clinical information available includes previous treatment, we could retrospectively correlate these mutations with the drug administered.

Additionally, this individualized approach could allow us:

- To discover new genomic biomarkers of drug response and targets for cancer drug development.
- To establish a comprehensive roadmap for selecting rational, multidrugs combination for anticancer therapy.
- To bring new driver genes to light for those patients who don't carry any known driver mutation.

This proposal fits in the PanCancer broad thems of : *Integration of genome and transcriptome, Integration of genome and epigenome, Pathway analysis, Genomic rearrangement architecture, Mutation signatures and Clinical correlations.*

Legacy plans

With this project we expect to deliver:

- A curated database of cancer pharmacogenomic data using text-mining approaches. There are some similar databases such as DrugBank and PharmGKB but don't integrate clinical candidates, experimental probes and FDA approved drugs with cancer genomic resources.

- A computational tool that evaluates and prioritizes the most appropriate therapy based on genomic profiling.

All tools, databases and methods are openly available and are will be ready to be distributed as web services or virtual machines, in the terms required by the PanCancer initiative.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Fatima Al-Shahrour, PhD

Present work (since February 2012)

Head of Translational Bioinformatics Unit. Clinical Research Program.
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.
Phone: (+34) 917328000 (ext 2409) email: falshahrour@cnio.es

Positions held:

- **October 2008–December 2011.** Computational Biologist- Postdoctoral Fellow at the group of Dr. MD. Benjamin L. Ebert. Cancer program, Broad Institute of MIT and Harvard and associated to the Department of Hematology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, US.
- **November 2007–August 2008.** Visiting Scientist at Dr. Jill P. Mesirov lab. Computational Biology and Bioinformatics program, Broad Institute of MIT and Harvard, Cambridge (MA), US.
- **March 2005–December 2008.** Research Scientist at the Functional Genomics Unit, Bioinformatics Department, under Dr. Joaquin Dopazo. *Centro de Investigacion Principe Felipe* (CIPF). Valencia, Spain.
- **February 2002–February 2005.** Bioinformatician at Dr. Joaquin Dopazo's Bioinformatics Unit. Spanish National Cancer Research Centre (CNIO). Madrid, Spain.

PhD degree:

Ph.D. in Molecular Biology, *Universidad Autónoma de Madrid* (Spain), November 14th 2006. Dissertation title: "Study and Development of Methods for the Functional Analysis of Genome- scale experiments". Qualification: Excellent *Cum Laude*. Supervisor: Dr. Joaquin Dopazo.

Specialization

- **Main field:** Computational Cancer Genomics. Functional Genomics.
- **Other fields:** RNAi screenings, Transcriptomics.
- **Current research interest:** Pharmacogenomics. Translational medicine.

Overview of Achievements

Author of 46 peer-reviewed articles and 6 book chapters, with a total of 2438 citations and an h-index of 22 (ISI web of knowledge, 22nd Nov. 2013): Cancer Cell, Nat Chem Biol, Nat Med, Nature, Genome Res, Blood, Plos Comp Biol. Principal investigator of Marie-Curie. Career Integration Grant. Associated Editor for BMC Bioinformatics. Regular reviewer for: Bioinformatics, PloS One, BMC Bioinformatics, BMC Systems Biology, BMC Genomics and Genomics. Co-organizer of The seventh international conference for the Critical Assessment of Microarray Data Analysis (CAMDA 2007). ISCB, ACCR, EACR Member.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Detection of somatic mutations in tumor samples disrupting the network of co-evolving molecular constraints

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia. Spanish National Cancer Research Centre (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
David de Juan (CNIO)	

Background and preliminary data

Our main aim is to contribute to obtaining additional information on the landscape of driver mutations. We focused this proposal on the detection of driver mutations in protein coding regions that are undetectable by current methodologies.

While current approaches rely on descriptions of evolutionary constraints based on detection of position patterns of conservation (Kumar *et al.*, 2009; Schwarz *et al.*, 2010; Reva *et al.*, 2011), **we intend to detect somatic mutations that are disruptive of networks of co-evolving molecular constraints.** This novel strategy has the potential to efficiently evaluate the impact of missense mutations in seemingly variable positions of the genome but which accessible states are strongly conditioned by epistatic effects.

Recent methodological and conceptual advances in the field of protein co-evolution have led to the realization of the surprising performance of s**tatistical models based on co-evolutionary data in describing intra- and inter-protein structural constraints resulting from physical molecular contacts** (Weigt *et al.*, 2009; Hopf *et al.*, 2012) and **our own contribution (Juan *et al.*, PNAS 2008).** The crucial step-forward of these methods is their ability to disentangle direct interdependences from a large amount of indirect spurious co-variations (Weigt *et al.*, 2009).

While these models have been very successfully applied to the problem of protein structure prediction (Hopf *et al.*, 2012), they can be additionally used to **provide a detailed and accurate representation of the network of evolutionary constraints operating on a gene**. For a review on the current status of the co-evolution in field of structural bioinformatics see our review (Juan et al, Nat Rev Genet 2013).

1000 Genomes Project Consortium *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061–1073.
Hopf,T.A. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149, 1607–1621.
Juan,D. *et al.* (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. U.S.A.*, 105, 934–939.
Juan, D. et al. (2013) Emerging methods in protein co-evolution. Nat. Rev. Genet., 14, 249–261.
Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4, 1073–1081.
Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, 39, e118.
Schwarz,J.M. *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, 7, 575–576.
Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, 106, 67–72.

Timelines & resources dedicated to project

The input information will be protein sequences deduced from the cancer genomes and the collection of mutations mapped to them.

The output will be the annotation of point mutations with their potential to disrupt the network of co-evolving residues in protein cores. This information will be interpreted in terms of epistatic interactions.

Research proposal

We propose to use a set of pre-calculated highly informative co-evolution based models for evaluating the impact of cancer somatic mutations in protein coding regions. In brief, we will provide an estimation of the evolutionary viability of the mutated protein by properly taking into account the network of statistical dependences among different positions.

This proposal fits in the general framework of PanCancer is relation to: *Mutation signatures*, *Interface between germline and somatic genetics* and *Landscape of driver mutations*

We plan to incorporate information on population genomic variants (1000 Genomes Project Consortium *et al.*, 2010) as well as from the control samples (PanCancer normal tissues) to our co-evolution based models. These extended statistical models based on the combination of evolutionary and population-based information will constitute a proper framework for the analysis of the sets of somatic mutations reported by the project.

So, **we aim to point to somatic mutations with the capacity of disrupting intra-molecular contacts essential for proper protein functioning.**

As a whole, **we intend to provide an original viewpoint for the analysis and understanding of the role of the disruption of the networks of molecular interdependences within genes along the different cohorts of cancer mutational landscapes.** Additional insights are also expectable from the analysis of our results in the light of the temporal evolution of cancer genomes, where a dynamic picture of disruption-compensation phenomena might also be observable.

Legacy plans

The software for the calculation of co-evolutionary networks has been published and it is publicly available. The execution is relatively simple and results can be pre-calculated prior to the mapping of mutations.

The more complex step is the preparation of the multiple sequence alignments that has also been formalized and can be prepared before hand.

The method(s) can be easily integrated in standard pipelines. Our intention is to maintain them as open and public.

In a second phase we will study the convenience of including the mutations in the calculation of the co-evolving positions but this is largely unexplored territory.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.

Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.

EMBO member since 2006.

Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

David Juan, MSc

Present Work (Since January 2007):

Bioinformatician at Prof. Alfonso Valencia's Structural Computational Biology Group.
Structural Biology and Biocomputing Programme.
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.
Phone: (+34) 917328000 (ext 3019) email: dadejuan@cnio.es

Positions Held:

- **(October 2000-December 2006)** Bioinformatician at Protein Design Group under Alfonso Valencia. Spanish National Biotechnology Center. Madrid, Spain.

MSc degree:

MSc in molecular Biology, *Universidad Autónoma de Madrid* (Spain), June 2003. Supervisor: Alfonso Valencia.

Specialization:

- **Main Field:** Protein co-evolution
- **Other fields:** Genome evolution, protein-protein interactions prediction, prediction of protein functional sites.
- **Current research interest:** Cancer genome evolution and epigenetics

Overview of Achievements

Author of 22 peer-reviewed articles, 2 reviews and 5 book chapters with a total of 444 citations and an h-index of 10 (ISI web of knowledge, 30th Dec. 2013): PNAS, Nat Immunol, Nat Rev Genet, Mol Biol Evol, FEBS letters, Nucleic Acids Res, Bioinformatics, J Mol Biol, Biol Open, PLOS ONE, BMC Bioinformatics, Proteins. Reviewer for: Bioinformatics, PLOS Comp Biol.

Title of abstract	
Comprehensive Analysis of the PATHOGENICITY of the Structural Variants, Rearrangements and Trans-splicing Events in Pan-Cancer samples	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Alfonso Valencia. Spanish National Cancer Research Centre (CNIO)	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Milana Frenkel-Morgenstern (CNIO)	

Background and preliminary data

Studying Fusion Transcripts in Normal and Cancerous Cells

The aberrant fusion of two distinct chromosomes by a genomic translocation event has been linked to several types of malignancy. The BCR/ABL fusion gene typical of Chronic Myelogenous Leukemia (CML) is a classic example, whereby the ABL1 protein is rendered constitutively active following fusion to the BCR locus, resulting in carcinogenic intracellular signalling. **Our goal is to identify transcript fusion events (either at the DNA or RNA level) and determine whether they can be used in cancer diagnostics, to identify the type of disease and perhaps even to direct the choice of therapy.** Our research strategy is to **catalogue the chimeric transcripts specifically expressed in different types of cancers, and to study the potential functions of their corresponding fusion proteins (Frenkel-Morgenstern, et al, 2012; Frenkel-Morgenstern, et al, 2012).** We previously generated a database of chimeric transcripts and RNA-sequencing data (The ChiTaRS database **http://chitars.bioinfo.cnio.es/** ; Frenkel-Morgenstern et al. 2013), containing more than 10,000 chimeric RNAs from humans, 233 of which have been confirmed by RNA-seq reads in 16 tissues (Human Body Map data), including ~1,000 cancer –associated chromosomal breakpoints that according to our analyses are probably pathogenic (Frenkel-Morgenstern et al. 2013). The database includes information on the expression and tissue specificity of these chimeras, as confirmed by RNA-seq data, and it includes mass-spectrometry confirmation for the presence of the translated fusion protein. Moreover, the database has advanced features to analyze "junction consistency" between two genomic regions, for example, how often the fusion events between the same two genomic regions result in a translatable gene product. Finally, all the **ChiTaRS entries are incorporated in the UniProt Knowledgebase (UniProtKB)** database as a part of the annotation of proteins in human, mouse and fruit fly.

The aim of mapping the **cancer fusion transcripts is to determine the likely pathogenicity of each fusion event in different types of cancers.** The expression level of fusion transcripts for each cell state (normal, diseased) will be found by high-throughput RNA sequencing. As the result of this study, an annotation (in more than 50 types of cancers), likely pathogenicity and the expression level of fusion transcript will be reported for all the different types of cancers in the PAN-cancer samples.

How the PAN-cancer data set will enable these analyses

1. The pathogenicity of the structural variants (rearrangements) at the description level using the description of the known breakpoints entries in the ChiTaRS database (a list of more than 1000 breakpoints in more than 50 types of cancers).

2. Prediction of the pathogenicity of the novel rearrangement using "junction consistency" analysis for the two parental genes.

3. The evaluation of the expression level and tissue specificity of the rearrangement and trans-splicing events at the RNA level by our method of mapping the chimeric junction sites by RNA-seq reads.

Previous Chimeric Transcript Analyses on Smaller Data Sets.

To assess the expression and validate the authenticity of candidate chimeric transcripts, we previously generated RNA-seq datasets. For human candidate chimeras, we utilized the Human Body Map 2.0 data generated on the HiSeq 2000 by Illumina in 2010. This dataset comprises 1,097 million (M) paired-end reads of 75 nucleotides derived from the sequencing of RNA from 16 different tissues.

- Frenkel-Morgenstern, et al., (2013). "ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data." Nucleic Acids Res 41(Database issue): D142-151.
- Frenkel-Morgenstern et al., (2012). "Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts." Genome Res 22(7): 1231-1242.
- Frenkel-Morgenstern Valencia (2012). "Novel domain combinations in proteins encoded by chimeric transcripts." Bioinformatics 28(12): i67-i74.

Timelines & resources dedicated to project

We will use the results on 1) the structural variants of the core analysis of all the samples, 2) unmapped RNA-seq reads for the chimeras expression and tissue-specificity analysis of all the cancer and normal samples.

From our ChiTaRS database, we will use:

1. A collection of 1000 breakpoints collected and manually verified from the TICdb, dbCrid, Mitellman, and Genetics Atlas databases
2. 10000 trans-splicing events, more than 1500 with the consistent junction sites
3. Cases confirmed by mass-spectrometry analysis in breast, prostate and ovarian cancers, confirmed by RNA-seq reads of 16 tissues of Human Body Map datasets

Research proposal

This proposal fits in the "*Genomic rearrangement architecture" general PanCancer aims.*

1. **The pathogenicity description analysis of the structural variants**
   Input: The structural variants of the core analysis of PAN-cancer
   Output: The annotation of every variant as found in different types of cancers, its pathogenicity

2. **The expression level of the fusions at the RNA level**
   Input: Unmapped RNA-seq reads from the core analysis of PAN-cancer
   Output: The expression level and tissue specificity of all the breakpoints and trans-splicing cases in our ChiTaRS database.

Our *Integration method for mapping of chimeric transcripts by RNA-seq reads* (***IntegChiP***) method utilizes GEM or Bowtie to align reads which cannot be mapped to the corresponding genome reference (unmapped reads). The dataset of chimeras (in the FASTA format) from the ChiTaRS database will be used as a template for the coverage by unmapped reads. The chimera expression was confirmed if at least two reads can be align to the putative chimeric junction site at least 10nt at each side of a junction with two mismatches. **The total number of RNA-seq read will be used to estimate the expression level of chimeras**. RNA-seqMap produces a tabular text file describing each chimera with the information on the genes incorporated in the chimera, junction site, number of mapped reads, the expression level using RPKM and the tissue specificity of the chimera.

3. **The prediction of pathogenicity of the novel fusions**
   Input: The structural variants of the core analysis of PAN-cancer
   Output: The annotation of every novel structural variant as found in different types of cancers, its pathogenicity using our junction consistence method. The specific features which will be predicted.
   Functional consequences of fusions of non-coding RNAs. We can analyze the trans-splicing cases I have in the ChiTaRS database of non-coding RNAs and their consequences
   Integration of genome and transcriptome. Comparative analysis of translocations and trans-splicing cases for every particular cancer.
   Pathway analysis. Pathways analysis for fusions by the context of missing protein domains using the junction consistence analysis.

Legacy plans

IntegChiP method will available for users
ChiTaRS database is already available for users
Both systems will be maintained for the next four years as part of a large European collaboration grant.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Milana Frenkel-Morgenstern, PhD

Present work (since February 2011)
Staff Scientist
In the lab of Alfonso Valencia (Structural Biology and BioComputing Programme)
Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
Phone: (+34) 917328000 (ext 3012)
email: mmorgenstern@cnio.es

Positions and Employment

1998-2002	Tutor - Mathematics and Computer Sciences Department, Bar-Ilan University, Israel.
2001- 2005	Lecturer – Popular Series in Science, Davidson Institute for Scientific Education, Rehovot, Israel.
2001-2012	Scientific Advisor - Davidson Institute for Scientific Education, Israel.
2001-2006	PhD Student - Weizmann Institute of Science, Rehovot, Israel.
2005-2012	Teaching Assistant - Feinberg Graduate School, Weizmann Institute.
2006-2009	Post-doctoral Fellow - Department of Molecular Cell Biology and Department of Structural Biology, Weizmann Institute of Science, Israel
2009-2011	Research Assistant - Spanish National Cancer Research Centre (CNIO), Madrid.
2011-Present	Staff Scientist - Spanish National Cancer Research Centre (CNIO), Spain.

Professional Memberships

2003-Present	Member of the International Society of Computational Biology (ISCB).
2007-2008	**Elected Secretary** of the Student Council at the International Society of Computational Biology (ISCBSC).
2008-Present	**Founder/Organizer** - The Annual Art & Science Exhibition @ ISMB
2012-Present	Member of the European Association for Cancer Research (EACR)
2012-Present	Member of FEBS

Awards, Honours and Fellowships

1999-2001	**Summa Cum Laude MA** in the Mathematics, Computer Sciences, Bar-Ilan University, Israel.
2006-2009	**Interdisciplinary Post-doctoral Fellowship** - The Horowitz Foundation for Studies of Complexity, Weizmann Institute of Science, Israel.
May, 2010	**Best Poster Award** - IBS 2010, Haifa, Israel.
July, 2010	**Outstanding Poster Award** - ISMB 2010, Boston, USA.
2009-2011	**Caja Navarra International Post-doctoral Fellowship** – Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

Specialization:
Cancer Genomics, Functional Genomics, Bioinformatics.

Overview of Achievements
Author of 13 peer-reviewed articles and one book with a total of 436 citations, published in Genome Research, Molecular Systems Biology, Trends in Genetics, Nucleic Acids Research, Bioinformatics, BMC Bioinformatics and others. Recipient of the Miguel Servet grant for Staff Scientists. Directly invited to review manuscripts for: Bioinformatics, PloSONE, PloS Computational Biology, BMC Genetics, BMC Bioinformatics, Molecular Oncology

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Evolutionary history of somatic mutations (including non-coding ones)

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia, Spanish National Cancer Research Centre (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
	Javier Herrero, The Genome Analysis Centre (TGAC)

Background and preliminary data

Distinguishing between driver and passenger variants is of critical importance for the understanding of cancer-related processes. In coding regions, one can predict the effect of the variant on the resulting protein and assess its severity. However, **analyzing non-coding somatic variants is a key challenge for the PanCancer initiative.**

We propose an evolutionary analysis for all the short somatic variants that will help to better predict their importance. We are especially interested in three evolutionary aspects. All these analyses **rely on the Enredo-Pecan-Ortheus (EPO) multiple alignments available in Ensembl.**

Firstly, we will use estimates of evolutionary conservation for each position in the genome. From the mammalian EPO alignments, GERP provide estimates of the conservation of each position in the genome. Driver mutations are more likely to affect highly conserved positions.

Secondly, we will infer ancestral alleles for all the short somatic variants. Ortheus is a phylogenetically aware re-constructor of ancestral sequences. These have been successfully used to polarize SNPs in the pilot data of the 1000 Genomes Project. We have recently extended the approach for the analysis of 1-bp indels. This requires re-aligning both alleles to produce reliable predictions. About 0.5% of somatic mutations and more than 1% of the indels revert a human-specific character to its ancestral state. These are unlikely to be drivers.

Lastly, we will annotate short-tandem repeats (STRs), homopolymers repeats (HRs) and other fast-evolving regions in the genome and confirm their evolutionary rate in the primate EPO alignments. These are especially relevant for short-indels as about half of the germline indels in human populations happen in these regions. Moreover, preliminary results show that ca. 70% of somatic indels are found in fast-evolving regions.

Timelines & resources dedicated to project

Mammalian GERP scores are already available in Ensembl. Also, the ancestry of any possible 1-bp variant (mutation or indel) in the human genome has been predicted and is accessible via a plugin for the Variant Effect Predictor (VEP). We plan to **use these predictions for all 1-bp somatic variants found in the PanCancer samples. We will adapt the approach for longer somatic indels.**

In addition, we will also **annotate fast-evolving regions in the human genome where germline indels are more likely to be found.** This will be used to analyze the relative incidence of somatic indels on these regions.

We will **use plugins of the Variant Effect Predictor** [McLaren et al, 2011] to annotate all the variants with this information.

This proposal fits in the PanCancer general subjects of "*Functional consequences of non-coding mutation* and *Temporal evolution of cancer genomes*"

Research proposal

We will **annotate all somatic variants in the PanCancer samples with three types of evolutionary information.** We will use **mammalian GERP scores** – which represent the level of evolutionary constraint – and study the distribution of these scores in different samples and cancer types.

We have already inferred ancestral alleles for any possible 1-bp variant in the human genome [Beal et al, in prep.]. These predictions cover both base mutations and indels. As indels affecting homopolymer repeats require special treatment, we **consider the whole homopolymer as reference and alternate alleles** (one being shorter than the other). Most 1-bp germline indels affect poly-A runs. We will **compare germline and somatic indels by looking at differences in terms of the incidence in homopolymers, segregating the results by the length of the repeat.**

For indels longer than 1-bp, pre-calculated genome-wide predictions are not available. We will use a similar approach to the one used for the 1-bp ancestral allele predictions. Namely, **we will re-align both alleles to predict the ancestral sequence with Ortheus and call the ancestral state accordingly.**

Some features in the genome tend to evolve more rapidly than the rest of the genome. For instance, it is very common to find indels in short-tandem repeats [Montgomery et al, 2013]. We will annotate these features in the genome to contrast the incidence of germline and somatic mutations on these regions.

As the prevalence of somatic variants varies greatly across cancer types [Alexandrov et al., 2013], we will **study these questions across the different types of cancer available in the PanCancer samples. We will also look at the differences among samples within each cancer type.**

This project will provide 3 types of annotations, especially relevant for the prioritization of non-coding variants.

Firstly, we will provide an estimate of the evolutionary conservation for each variant.

Secondly, we will produce ancestral allele predictions for all the short somatic mutations in the PanCancer samples.

Lastly, we will study the incidence of short variants in fast-evolving regions. We foresee the integration of these results with other similar efforts to provide clues on the severity of a given somatic variant. In addition, we will produce virtual machine with all the software required to repeat the analysis on new samples.

- McLaren et al. "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor". Bioinformatics 26 (16): 2069-2070 (June 18, 2010). doi:10.1093/bioinformatics/btq330

- Beal, et al. "Inference of the ancestral allele for any 1-bp variant in the human genome." In preparation.

- Montgomery, et al. "The Origin, Evolution and Functional Impact of Short Insertion-Deletion Variants Identified in 179 Human Genomes." Genome Research 23: 749-761 (March 11, 2013). doi:10.1101/gr.148718.112.

- Alexandrov, et al. "Signatures of mutational processes in human cancer." Nature 500: 415–421 (22 August 2013) doi:10.1038/nature12477

Legacy plans

To run the predictions on the PanCancer sample, all the software will be installed in a virtual machine. This will include the VEP, the Conservation and AncestralAlleles plugins and the Pecan and Ortheus software.

While we have no ownership of some of the pre-existing software (i.e. Pecan and Ortheus), the code used to run the analyses will be freely available to the whole research community. We anticipate distributing the source code and instructions via GitHub.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Javier Herrero

I developed my PhD research project at the CNIO (Spain). My project focused on functional genomics, more specifically on the analysis of co-expressed genes.

I joined the Ensembl team in 2004 to work in comparative genomics. I extended the system to support the storage and display of whole genome sequence alignments. To build these alignments, I developed Enredo to form the EPO pipeline (Enredo-Pecan-Ortheus) in collaboration with Benedict Paten who wrote the Pecan and Ortheus algorithms. The pipeline also runs GERP, a software that estimates an evolutionary conservation score of each basepair in the alignment and calls conserved elements from these.

In parallel, I have worked on the human ENCODE project where I have used an evolutionary perspective to analyse the different results of the study. We have looked at the conservation of different signatures in the human population and across mammalian species. While there is a good correlation for many features, others are well conserved in human population but not necessarily in all mammals.

I am currently involved with the mouse ENCODE project, helping in the comparison of the functional data between the human and mouse genomes. We have built and compared several tools for mapping ChIP-seq peaks and DNAse data across species.

I also participate in the 1000 genomes project, providing estimates of ancestral alleles on the human genome and collaborating with the Functional Interpretation Group. We are in charge of the annotation of coding and non-coding polymorphisms in the human genome. For instance, we have classified the SNPs by their predicted effect in the genome (disrupting a splice site, non-synonymous SNP, etc) and related these classes to the evolutionary conservation of these positions.

More recently, we have started to look at the ancestry of 1-bp indels in the human genome. Starting from the original primate EPO alignments, we re-align both alleles and use the predictions from Ortheus in both cases to confirm the predicted ancestral state. The difficulty of this analysis is that indels tend to appear more frequently in less stable areas of the genome where building a reliable alignment is challenging.

Scientific Career

Jun 2013 – present	**Comparative Genomics Project Leader (TGAC, UK)**
Sep 2011 – May 2013	**Ensembl Coordinator and Compara Project Leader (EMBL-EBI, UK)**
Oct 2007 – Aug 2011	**Ensembl Compara Project Leader (EMBL-EBI, UK)**
Jun 2004 – Sep 2007	**Ensembl Developer (EMBL-EBI, UK)**
Jan 1999 – Jan 2004	**PhD in Molecular Biology (Bioinformatics) by the UAM, Spain**
Jan 1998 – Dec 1998	**Graduate student in the Biophysics group (UCM, Spain)**
Sep 1992 – Jun 1997	**Degree in Chemistry (specialised in Biochemistry) by the UCM, Spain**

Publications Summary

- ➢ *First/Last author: 10 articles*
- ✓ *Senior author: 11 articles*
- ○ *Other contributions: 26 articles + 1 review*

H-index: 35 (Google Scholar; Dec 2013)

Full list: http://scholar.google.co.uk/citations?user=5aDQF4EAAAAJ&hl=en&oi=ao

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Analysis and classification of mutations in protein kinases. A family specific approach.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia. Spanish National Cancer Research Centre (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Tirso Pons (CNIO)	**Jose Maria G. Izarzugaza (CBS-DTU, Denmark)**

Background and preliminary data

Over the last 5 years we have analyzed in detail the specific characteristics of mutations in protein kinases, a protein superfamily that plays a central role in cancer. We have analyzed their structural and evolutionary characteristics, as deduced from the analysis of large multiple sequence alignments (Izarzugaza et al., 2009, 2011).

In parallel, we developed systems to complement the information about mutations available in databases with the one extracted from the literature with a text mining approach (see Krallinger et al,. 2009, Izarzugaza et al., 2012a).

Base on this information we **developed a predictor (KinMut) that prioritizes pathogenic mutations in protein kinases** by combining all these information. **KinMut outperforms the usually applied general predictors.** The **differential elements build in the kinase specific predictor are** two: statistical information about the incidence of mutations in different families of kinases and the use of methods to predict the **differential conservation at the protein family level, instead of using general predictions of conservation for the full protein supefamily** (for a description of the importance of the family specific differential information see Juan et al. 2013). The technical description of the KinMut predictor, details of the benchmarking and comparison with other predictors can be found in Izarzugaza et al., (2012b).

Finally, we have developed a **complete system (wKinMut) that integrates the predictions in an array of information** about the corresponding sequence, structure and functional information (Izarzugaza et al., 2013).

wKinMut maps the mutation in the 3D structures of corresponding proteins (or their most approximate structural model) and put them in the context of protein complexes and interactions. It also facilitates to the user the basic information about the function of the kinases and the specific mutations, if they are already known. The information is extracted databases or directly mined from the literature. The system provides information about the relation between the mutations and the protein functional and potential drug binding sites. If the mutations are already known the information about tumor of origin and histology is provided.

We have compared the **KinMut and wKinMut systems with the** 100 colon adenocarcinomas recently published. The systems have also been applied **to the cancer genome projects in which the group participates, including the CLL-ICGC project,** as well as in the recently published bladder cancer analysis apper (Balbás-Martínez et al., 2013).

- Balbás-Martínez et al., (2013) Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. Nat Genet. 45(12):1464-9.
- de Juan D et al., (2013) Emerging methods in protein co-evolution. Nat Rev Genet. 14(4):249-61
- Izarzugaza et al., (2009) Cancer-associated mutations are preferentially distributed in protein kinase functional sites. Proteins. 77(4):892-903.
- Izarzugaza et al., (2011) Characterization of pathogenic germline mutations in human protein kinases. BMC Bioinformatics. 2011;12 Suppl 4:S1
- Izarzugaza et al., (2012a) Interpretation of the consequences of mutations in protein kinases: combined use of bioinformatics and text mining. Front Physiol. 3:323
- Izarzugaza et al., (2012b) Prioritization of pathogenic mutations in the protein kinase superfamily. BMC Genomics. 13 Suppl 4:S3.
- Izarzugaza et al., (2013) wKinMut: An integrated tool for the analysis and interpretation of mutations in human protein kinases. BMC Bioinformatics. 14(1):345.
- Izarzugaza JM, Hopcroft LE, Baresic A, Orengo CA, Martin AC, Valencia A. BMC Bioinformatics. 12 Suppl 4:S1.
- Krallinger et al., (2009) Extraction of human kinase mutations from literature, databases and genotyping studies. BMC Bioinformatics. 10 Suppl 8:S1

Timelines & resources dedicated to project
The input in this case will be WGS or complete exomes with the associated point mutations in protein kinases. The output will be a detailed description of the potential pathogenicity of the mutations together with selected information about the associated evolutionary, structural and functional information, together with the links to the underlying information in databases and literature.
Research proposal
In this proposal we will address a central topics of the Pan-Cancer analysis i) *Landscape of driver mutations.* **KinMut and wKinMut** provides a comprehensive analysis of mutations in protein kinases. The systems are specfic for this protein superfamily, since they use sequence and functional information specific of these proteins. The system has been used and tuned during the application to various cancer genome analysis projects. **KinMut and wKinMut** are ppublicly available. Both systems are designed in a way in which they can be integrated in other analysis plathforms. The proposal in the case of pan-cancer is to label each mutation an estimation of their possible role in pathogenicity (prediction of the consequences of point mutations) as well as the associated functional, sequence derived and structural information. If the mutations have been already reported the associated information is reported.
Legacy plans
The **KinMut** software, **wKinMut** web server are fully functional and openly available for their external use. In the context of PanCancer webservices or virtual machines encapsulating the services offered by **KinMut** and **wKinMut** will be developed. Both systems are maintained as part of the services offered by the CNIO core bioinformatics platform.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Tirso Pons, PhD

Present work (since July 2011)

Staff Scientist, Structural Computational Biology Group, Structural Biology and Biocomputing Programme.
Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
Phone: +34 91 7328000 E-mail: tpons@cnio.es
Web: http://www.researcherid.com/rid/A-6377-2011

Previous positions

- 2006-2011: Associate Researcher at Center for Protein Research (CEP), University of Havana, Cuba.
- 2004-2011: Adjunct Professor in Bioinformatics. Department of Biochemistry, Faculty of Biology, University of Havana, Cuba.
- 1993-2005: Research Scientist at Physical Chemistry Division, Center for Genetic Engineering & Biotechnology (CIGB), Havana, Cuba.
- October 1995 - January 1996/April – June 1998: Visiting Scientist at Dr. Alfonso Valencia´s lab, National Center of Biotechnology (CNB-CSIC), Autonomous University of Madrid, Spain.
- June – November 1998: Visiting Scientist at Dr. Gert Vriend´s lab, Biocomputing Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.

PhD Degree

Ph.D. in Biology, University of Havana (Cuba), January 21th 2003. Dissertation title: "Sequence analysis, structure and functional residues prediction for glycosyl hydrolases families 32, 49, and 68" Qualification: *Summa cum laude*. Supervisors: Dr. Alfonso Valencia and Dr. Joaquín Díaz-Brito.

Specialization

- Main field: Sequence analysis, comparative protein modeling, functional residues prediction.
- Other fields: Structure-function relationship, protein-protein interactions.
- Current research interest: Understanding the structural impact of somatic mutations in cancer.

Overview of Achievements

Author of 48 peer-reviewed articles, with a total of 796 citations and an h-index of 12. Member of the editorial board of the Open Access journals: *Bioinformatics and Biology Insights* (Libertas Academica Press) and *Journal of Proteome Science and Computational Biology* (Herbert Publications Ltd). Scientific reviewer for: *PLoS Computational Biology*, *Bioinformatics*, *FEBS Letters*, *Journal of Molecular Evolution*, *Journal of Molecular Modeling*, *International Journal of Biological Macromolecules*, *Biotechnology Progress*, *Microbiology and Molecular Biology Reviews*. EMBL Alumni association member.

José M. G. Izarzugaza, PhD

Current Employment (Since August, 2012)
Assistant Professor at the Integrative Systems Biology Group (Prof. Søren Brunak)
Guest member of the Functional Human Variation Group (Dr. Ramneek Gupta)
Center for Biological Sequence Analysis (CBS)
Department of Systems Biology, Technical University of Denmark (DTU)
Phone: (+45) 528 17 521, e-mail: josemgizarzugaza@gmail.com

Employment history

2006 – 2012	**Spanish National Cancer Research Centre (CNIO)** Structural Computational Biology Group (Prof. Alfonso Valencia)
2008	**University College London (UCL)** Biomolecular Structure and Modelling Unit (Prof. Christine A. Orengo)
2005 – 2006	**Spanish National Centre for Biotechnology (CNB-CSIC)** Protein Design Group (Prof. Alfonso Valencia)
2004 – 2005	**Noray Bioinformatics** Genomics and Proteomics Group
2002 – 2003	**Agilent Technologies** Life Sciences Business Unit

PhD in Molecular Biology
Universidad Autónoma de Madrid (Spain). November 2011. Mutations in the Protein Kinase Superfamily. Supervisor: Prof. Alfonso Valencia. Qualification: Excellent, honours "*cum laude*" and "*doctor europeus*".

Publications:
13 peer-reviewed articles (1st author:10), 2 book chapters, h-index: 8 (Dec 2013)

- **Izarzugaza JMG,** Vazquez M, Pozo A, Valencia A. wKinMut: An integrated tool for the analysis and interpretation of mutations in human protein kinases. BMC Bioinformatics 2013
- Birkbak NJ, Kochupurakkal B, **Izarzugaza JMG**, Eklund AC, Li Y, Liu J, Szallasi Z, Matulonis UA, Richardson AL, Iglehart JD, Wang ZC. Tumor mutation burden forecasts outcome in patients with ovarian cancer. Plos One 2013,8(11)
- Johansen MB, **Izarzugaza JMG**, Brunak S, Petersen TN, Gupta R. Prediction of disease causing non-synonymous SNPs by artificial neural network predictor NetDiseaseSNP. Plos One 2013,8(7)
- **Izarzugaza JMG**, Krallinger M, Valencia A. Interpretation of the consequences of mutations in protein kinases: combined use of bioinformatics and text mining. Front Physiol. 2012;3:323
- **Izarzugaza JMG**, Pozo A, Vazquez M, Valencia A. Prioritization of pathogenic mutations in the protein kinase superfamily. BMC Genomics (2012) 13(Suppl 4):S3.
- **Izarzugaza JMG**, Baresic A, McMillan LEM, Orengo CA, Martin ACR, Valencia A. Characterization of pathogenic germline mutations inhuman Protein Kinases. BMC Bioinformatics 2011, 12:336.
- **Izarzugaza JMG**, Baresic A, McMillan LEM, Yeats C, Clegg AB, Orengo CA, Martin ACR, Valencia A. An integrated approach to the interpretation of Single Amino Acid Polymorphisms within the framework of CATH and Gene3D. BMC Bioinformatics 2009, 10(Suppl 8):I1.
- **Izarzugaza JMG**, Krallinger M, Rodriguez-Penagos C, Valencia A. Extraction of human kinase mutations from literature, databases and genotyping studies. BMC Bioinformatics 2009, 10(Suppl 8):II.
- Ezkurdia I, Graña O, **Izarzugaza JM**, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 2009 Jul 24.
- **Izarzugaza JM**, Redfern OC, Orengo CA, Valencia A. Cancer associated mutations are preferentially distributed in protein kinase functional sites. Proteins 2009 Jun 19.
- Baudot A, Real F, **Izarzugaza J**, Valencia A. From cancer genomes to cancer models: bridging the gaps. EMBO Rep (2009) pp.
- Pazos F, Juan D, **Izarzugaza JM**, Leon E, Valencia A. Prediction of protein interaction based on similarity of phylogenetic trees. Methods Mol Biol (2008) vol. 484 pp. 523-35.
- **Izarzugaza JM**, Juan D, Pons C, Pazos F, Valencia A. Enhancing the prediction of protein pairing between interacting families using orthology information. BMC Bioinformatics. 2008 9:35.
- **Izarzugaza JM**, Graña O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7 Proteins 2007 Aug 1.
- **Izarzugaza JM**, Juan D, Pons C, Ranea JA, Valencia A, Pazos F. TSEMA: interactive prediction of protein pairings between interacting families. Nucleic Acids Res. 2006 Jul 1;34 (Web Servers issue):W315-9.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

APPRIS – selection of principal splice isoforms and constitutive exons

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia. Spanish National Cancer Research Centre CNIO

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Michael Tress, CNIO	

Background and preliminary data

The cellular role of **alternative protein isoforms** is a topic of growing interest, both in normal cells and in cancer research. The functional annotation of alternative isoforms presents a serious challenge, not least because of the sheer quantity of genomic data that is being generated.

The **APPRIS database (appris.bioinfo.cnio.es)** addresses the challenge of annotating alternative protein isoforms with functional information. APPRIS is made up of separate modules, which a**nnotate alternative isoforms with information relating to protein structure, function, or localization, and with levels of cross-species conservation.** APPRIS provides a visual representation for each gene that allows the identification of functional changes brought about by splicing.

As part of the annotation process, **APPRIS also identifies a single principal isoform for each gene**. This is an important technical development that will improve the reliability of large-scale research projects. However, we believe that it also reflects the reality of gene organization at the cellular level. Recent results from our group suggests that most genes have a single clearly definable principal isoform and alternative protein isoforms that either have much shorter half-lives, or are expressed less frequently or in limited tissues. The results from **multiple large-scale proteomics analysis are in agreement with APPRIS – the APPRIS principal isoform is the isoform detected most frequently in proteomics experiments in 97% of genes.**

With **APPRIS we have annotated over 19,000 human genes with structural, functional and conservation information selected a unique principal isoform for 85.5% of the genes in by Ensembl.** The annotation of principal variant also allows us to determine the set of constitutive exons.

APPRIS annotations are **particularly important for the analysis of cancer associated mutations, since it provides a direct way of highlighting variants that map into principal isoforms and therefore likely to affect function**: The commonly in cancer genome papers of mapping mutations into the larger isoform is not wrong in scientific terms and creates an artificial increase in the number of mutations in coding regions and it favors the selection of long genes as potentially cancer related. The systematic implementation of appris in the intitial steps of the analysis will prevent this systematic bias in the analysis of cancer variants.

The **APPRIS database is stable (we are now on our fourth annotation of the human genome database)** and is being implemented as part of the Ensembl human genome annotation, which guarantees its update with every new Ensembl release. The database has been extended to rat, mouse, Danio and Iberian lynx and the UniProt database plans to use APPRIS to select a display isoforms for all model organisms. The set of constitutive exons provided by APPRIS is also in use in hospitals to aid in the detection of genes and mutations involved in genetic disorders.

- Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, Valencia A, Tress ML. (2013) APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 41(Database issue):D110-7.

- Harrow et al., (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22(9):1760-74

Timelines & resources dedicated to project

APPRIS provides information that is essential for the correct interpretation of point mutations in coding regions. As such the only input needed for the determination of principal isoforms are the complete genome sequences.

For the comparative study of the expression of isoforms and constitutive exons the information on normal and cancer samples will be used.

Finally, in the cases in which **expression at the RNA level will be determined based on RNAseq experiments we will use APPRIS results to analyze differential expression of principal isoforms and the possible biological role of other expressed isoforms with altered protein characteristics.**

Research proposal

In a first phase **APPRIS will use as input PanCancer complete genome and identify the principal isoforms and sets of constitutive and alternative exons for each gene.**

Knowing the set of constitutive exons will aid in **assessing the likely effect of mutations on any gene product** because mutations will can be divided into those that affect constitutive exons and those that affect alternative exons.

The information provided by APPRIS will make possible the **analysis of variation in composition of principal isoforms associated to each gen / sample / cancer type.**

In addition **APPRIS provides protein-centric annotations for all protein isoforms.** These include the mapping of protein structure, Pfam functional domains, trans-membrane helices and signal peptides to the isoforms, along with functional residues predicted by the functional residue predictor *firestar*. Firestar provides a way of mapping known mutations to ligand binding sites and catalytic residues; mutations that affect important functional residues are likely to be the most deleterious mutations (along with those that undermine the 3D structure of the protein).

In a **second phase for those cases in which RNAseq information is available APPRIS will provide basic information to analyze the level of expression of principal isoforms** (the ones likely to be functional since they contain all the key "protein like" features) and other isoforms representing alterations of the main functions associated each particular gene.

Legacy plans

APPRIS updates have been regularly published (NAR 2010, 2013**). The system functions stably as part of the GENECODE / ENCODE project (ENCODE GENECODE papers) and is used to produce Ensembl human genome annotation, which guarantees its update with every new Ensembl release.**

For the pan-cancer project APPRIS will function as a virtual machine. Therefore it should be easy to integrate it in any cloud-based analysis pipeline.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Michael Tress, PhD

Present work (since May 2006):

Staff Scientist. Structural Computational Biology Group.
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.
Phone: (+34) 917328000; email: mtress@cnio.es

Positions held:

- February 2002-April 2006. Postdoctoral research fellow at the Centro Nacional de Biotechnologia (CNB), Madrid, Spain.

PhD degree:

Ph.D. in Biology, University of Warwick (United Kingdom), January 2002. Supervisor: Prof. David Jones.

Specialization

- **Main fields:** protein structure and function prediction; proteomics; genome annotation;
- **Other fields:** sequence analysis; evolutionary biology; protein interactions.
- **Current research interest:** annotation of the human reference genome; alternative splicing; large-scale function prediction.

Overview of Achievements:

Author of 39 peer-reviewed articles (12 first-authorships, 11 last-authorships), 4 book chapters: Nature, Nucleic Acid Res, Bioinformatics, Mol Biol Evol, Proteins, PNAS, Genome Res, Genome Biol, Curr Protoc Protein Sci., PloS Biol and others.

Regular teaching activities in Bioinformatics. Regular reviewer for: Bioinformatics, Mol Biol Evol, Nucleic Acids Res, Proteins.

Co-organizer of the Critical Assessment of Techniques for Structure Prediction (CASP7.5) meeting. Official CASP assessor, CASP8.

Co-PI with Alfonso Valencia in the GENCODE Consortium, which is annotating the Human Reference genome as part of the ENCODE Project.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

FireDB and firestar, mapping of mutations to functional residues

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia. Spanish National Cancer Research Centre (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Michael Tress (CNIO)	

Background and preliminary data

The rapid growth of the primary sequence and structure databases has generated a huge number of proteins without experimental functional data. Experimental approaches for the characterization of function are expensive and difficult to automate and this has meant that researchers have turned increasingly to computational methods.

Much of this information, such as the amino acid residues implicated in molecular interactions or catalysis, can be found at the residue level, although it is often difficult to access. **FireDB (firedb.bioinfo.cnio.es), an inventory of small molecule ligand binding sites** derived from the protein structures deposited in the Protein Data Bank was developed to make use of the data in the structural and sequence databases.

All ligands and binding sites in FireDB have been annotated with information about their biological relevance by a combination of exhaustive literature searches and automatic analyses of all-against-all comparisons of binding sites. As a result **FireDB contains the largest available set of biologically relevant ligands, automatic cross-referencing to publicly available biological, chemical and pharmacological compound databases for over 95% of the ligands in FireDB and tags over 30,000 individual sites as biologically important**.

In addition to providing a vast quantity of biological and pharmacological information on its own, the functional information in FireDB also provides the raw data for the functional residue predictor ***firestar*** **to make predictions for biologically relevant ligand binding residues and catalytic residues.**

firestar, **an expert system for predicting functional residues in protein sequences, makes predictions by homology-based transfer of the biologically important information stored in FireDB.** Firestar uses measures of local sequence conservation to predict ligand binding residues and catalytic sites from the small molecule ligand binding residues organized in the FireDB database.

Firestar **is able to produce high quality results in a high throughput mode by using sequences as the only input.** Firestar was able to make reliable predictions for 40% of Pfam functional families. *firestar* has been benchmarked in the CASP8, CASP9 and CASP10 ligand binding prediction targets. The server was able to detect ligand binding residues for 94% of the sites and *firestar* outperformed all officially participating groups over the three experiments.

- Lopez et al., firestar--advances in the prediction of functionally important residues. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W235-41
- Lopez et al., firestar--prediction of functionally important residues using structural templates and alignment reliability. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W573-7
- Lopez et al., FireDB: a compendium of biological and pharmacologically relevant ligands. Nucleic Acids Res. 2007 Jan;35(Database issue):D219-23
- Maietta et al., FireDB--a database of functionally important residues from proteins of known structure. Nucleic Acids Res. 2013 Nov 15. [Epub ahead of print]
- Quesada et al,.Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2011 Dec 11;44(1):47-52

Timelines & resources dedicated to project
The *firestar* input are the predicted coding genes (protein sequences) and the position of the mutations. The **output is divided in two parts:** **a) the likelihood of the mutations to be affecting a known binding site for biological cofactors and** **b) the likelihood of the mutations to be affecting a known drug binding sites.** In both cases the information regarding the characteristics of the binding sites and ligand (or drug), the relation with similar binding sites in other proteins and the links to the relevant literature are provided.
Research proposal
Firestar provides a direct way of mapping mutations to ligand and drug binding sites. The mapping can be carried out for binding sites directly derived from experimental (x-ray) data or extrapolated to proteins with a reliable level of similarity. The system has been systematically used and updates and releases regularly produced. FIrestar has been publicly tested in the context of the CASP protein structure prediction challenge. In the 2013 competition it was the best performing server for the prediction of binding sites. **The proposal in the case of pan-cancer is to label each mutation in coding regions with: a) no information, b) affecting a binding site, c) predicted to affect a binding site, d) predicted to be affect a drug binding sites.** The information provided by **Firestar can be used as input by other methods for the prediction of the consequences of point mutations.** Additionally the identification of potential ligand binding or catalytic residues can provide important clues for the design of targeted biochemical experiments, and can be a vital part of drug design and virtual screening.
Legacy plans
Firestar updates **have been regularly published** (NAR 2010, 2013). The system **is stable and form parts of the CNIO cancer genome analysis pipeline**. It has also been used for the **analysis of the CLL-ICGC project** (Nat Genet 2011). For the pan-cancer project we can make the Firestar webserver **accessible in the form a virtual machine** that can be integrated any cloud-based analysis pipeline.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.

Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.

EMBO member since 2006.

Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Michael Tress, PhD

Present work (since May 2006):

Staff Scientist. Structural Computational Biology Group.
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.
Phone: (+34) 917328000; email: mtress@cnio.es

Positions held:

- February 2002-April 2006. Postdoctoral research fellow at the Centro Nacional de Biotechnologia (CNB), Madrid, Spain.

PhD degree:

Ph.D. in Biology, University of Warwick (United Kingdom), January 2002. Supervisor: Prof. David Jones.

Specialization

- **Main fields:** protein structure and function prediction; proteomics; genome annotation;
- **Other fields:** sequence analysis; evolutionary biology; protein interactions.
- **Current research interest:** annotation of the human reference genome; alternative splicing; large-scale function prediction.

Overview of Achievements:

Author of 39 peer-reviewed articles (12 first-authorships, 11 last-authorships), 4 book chapters: Nature, Nucleic Acid Res, Bioinformatics, Mol Biol Evol, Proteins, PNAS, Genome Res, Genome Biol, Curr Protoc Protein Sci., PloS Biol and others.

Regular teaching activities in Bioinformatics. Regular reviewer for: Bioinformatics, Mol Biol Evol, Nucleic Acids Res, Proteins.

Co-organizer of the Critical Assessment of Techniques for Structure Prediction (CASP7.5) meeting. Official CASP assessor, CASP8.

Co-PI with Alfonso Valencia in the GENCODE Consortium, which is annotating the Human Reference genome as part of the ENCODE Project.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The Rbbt framework and ICGCScout. Workflow enactment for the PanCancer projects, its infrastructure and functionalities

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alfonso Valencia, Spanish National Cancer Research Institute (CNIO)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Miguel Vazquez and Victor de la Torre (CNIO)	

Background and preliminary data

Large scale cancer analysis, i.e. **PanCancer requires the application of a variety of computational methods to datasets of molecular features and cohorts of patients.** Particularly complex are the requirements of the exploratory phases of the cancer projects.

The Pan-Cancer workflows should be: (1) reproducible for every cancer cohort, (2) configurable and easy to parametrize, (3) executable on-demand to support exploratory analysis, (4) the execution of each task in the workflow should be reusable, and (5) the complete analytical workflow should be easily to enact by third parties, including the installation and configuration of the complete infrastructure.

We have developed the **Rbbt framework** that implements these principles. The **Rbbt framework provides a workflow management system and associated web services, including a comprehensive collection of secondary and tertiary analysis functionalities**. In addition a exploratory environment, **ICGCScout (http://se.bioinfo.cnio.es) provides that incorporates the public ICGC/TCGA data and a representative selection of most useful workflows that can be executed on-demand.** An integrated installation package greatly facilitates its use. It includes the required data gathering, databases installation, and software configuration.

The Rbbt framework does not only **deal with the enactment of workflows, it also provides resource management and processing, as well as a full provenance of the results from the raw data sources**. Provenance is not just annotated as meta-data but implemented in executable code that allows a complete reproduction of the results.

PanCancer related projects will involve downloading gigabytes of genomic data together with processes results and will require the creation of dozens of fast access noSQL databases. Rbbt incorporates locking mechanisms for concurrent processes, achieving a high level of parallelism (of up to 31 CPUs at 100%). The creation and provisioning of virtual machines is reproducible even for complex workflows without further configuration requirements. Rbbt modular and decentralized design allows different groups to develop and maintain workflows, and even start REST web servers that provide remote access to their functionalities. Rbbt decouples the logic of the workflows from the technical details of how the system is deployed, and makes it especially suitable for a cloud environment.

A basic description of the system can be found in: (*Vazquez, de la Torre, Valencia A (2012) Chapter 14: Cancer genome analysis. PLoS Comput Biol.8(12):e1002824 and Gonzalez-Perez … Vazquez … Valencia et al., (2013) Computational approaches to identify functional genetic variants in cancer genomes. International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group. Nat Methods. 10(8):723-9*)

Installations of the system are already in use in Netherlands Cancer Institute and the Spanish CNAG-BSC centres. Rbbt has been instrumental in the CNIO analysis of the CLL-ICGC exomes (Quesada … Vazquez… Valencia et al., Nat Genet 2011), as well as the recent bladder cancer analysis (Balbás-Martínez … Vazquez … Valencia et al., Nat Genet 2013), and ovarian cancer (in preparation).

Timelines & resources dedicated to project
rbbt is explicitly dedicated cancer genome analysis. The current version is stable and able to (easily) include new functionalities developed by third parties. A full documentation on the use of the framework and user training materials are available. The experience acquired during the collaboration with groups in our institution and others has greatly helped to improve and debug the system. The system was developed as part of the CLL-ICGC project and it is now maintained as a core CNIO resource since it is used for the analysis of a number of institutional cancer projects.
Research proposal
We propose to use of the Rbbt framework for the PanCancer analysis. Rbbt is mature and stable, integrates a range of functionalities organizing them in use-case oriented workflows. **Rbbt and ICGCScout address the problem of system maintenance and PanCancer analysis,** addressing in particular the (*Integration of genome and transcriptome, Pathway analysis, Mutation signatures, Landscape of driver mutations and Clinical correlations).* Rbbt operates over **two molecular types: genetic variants and gene expression, and the results are put in the context of the main pathway/function databases**, including GO, NCI Nature curated pathways, Reactome, Biocarta, Pfam, Matador, COSMIC, and others. **Rbbt include the following functionalities**: mutation consequence inference, mutated isoform damage predictions (including MutationAssessor, SIFT and other popular methods and our own kinase-specific predictor), mutation frequency assessment, spectral mutation analysis, gene expression analysis, functional enrichment analysis (over-representation, rank-based, and custom statistics base on mutation frequencies), network visualization and manipulation, survival analysis, and structural analysis of protein mutations (in tertiary structures and protein complexes). **A selection of production ready workflows can be accessed at https://github.com/Rbbt-Workflows.** **We propose to implement a number of use-cases, including upon our own methods and those required by other groups in the consortium. The workflows will be provided in a way that they can be systematically applied to PanCancer.**
Legacy plans
The **rbbt framework** and the **ICGCScout application are maintained as part of the core CNIO systems** and used in our internal projects, including our analysis of individual cancer cases. Legacy versions are easy to install and maintain by third parties. The CNIO have sufficient computational capacity to carry out the necessary adaptation and development and if necessary to compute the analyses required.

Alfonso Valencia

Current Positions

- Vice-Director of Basic Research and Director of the Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)
- Director of Spanish Bioinformatics Institute (INB-ISCIII)
- Executive Editor of Bioinformatics (OUP) since 2006.
- President elect of the International Society for Computational Biology (ISCB) 2013-.

Previous Positions

- Ph.D. Biochemistry and Molecular Biology, U. Autonoma Madrid, 1988.
- EMBO Post-doctoral fellow, EMBL- Heidelberg, Chris Sander´s lab. 1989-1994.
- Group Leader, National Centre for Biotechnology Spanish (Research Professor CSIC 2004)

Selected Committees and professional activities

EMBL Scientific Advisory Committee since 2006-2012. Biozentrum U. Basel SAB 2006-. Swiss Institute for Bioinformatics SAB 2008 -. KU Lueven Center for Human Genetics 2012-. Bioinformatics Unit Curie Institue since 2011-. Intepro database SAB since 2008-2013. Coordinator of the Evaluation Committee of the Spanish Network of High-Performance Computing (2006-2011). Member of the jury of the Elsevier Grand Challenge, 2008-2009. Assessor of the CASP protein structure competition in the 9 and 10th editions. Evaluation Panel of the ERC Advance Grant schema 2008, 2010, 2012, 2014. Spanish Committee for Grant Evaluation (ANEP) 2009-2012. EMBO postdoc Fellowship committee 2009-2012. Member of various EC evaluation panels. DFG "cluster of excellence initiative (2007 and 2011. EPSFR grant committee, as ad hoc member. Founder member of the Science and Art "e-biolab" initiative. Ad-hoc reviewer for the main scientific journals and Bioinformatics/computational Biology conference/conference committees.
Founder and organizer of the BioCreative Challenges (meetings in 2004, 2007, 2011 and 2013, ESF and NSF funding). Organizer of the Biolink Text mining workshop link to ISMB since 2002.
EMBO member since 2006.
Professor Honoris Causa of the Danish Technical University DTU (2010)

Current Funding

Spanish Government (2013-2017), INB infrastructure/ ELIXIR ISCIII 2014-2018), CLL / ICGC (2009-2013), RTIC FIS 2008-2014, GENCODE 2009-2012 2013-2016, e-TOX IMI (2009-2014), Open Phacts IMI (2011-/2015), ASSET EU 7thFP (2010-2015), RD-Connect / IRDiRC (2013-2018), BLUEPRINT / IHEC (2012-2016).

Scientific Accomplishments

I have published over 300 papers in biological journals (included in Medline) and computational journals (i.e. IEEE). In terms of impact my H-index is approximately 60 with more than 7000 citations evenly distributed (not due to a single very quoted papers and in a few cases in large consortiums). My most quoted papers include collaborations with experimental biologists, analysis of biological problems related with evolution. I have published papers that are considered the foundation of areas of bioinformatics, such as: co-evolution based prediction of protein contacts and prediction of protein networks, prediction of subfamily specific residues and application of text mining to molecular biology.

Miguel Vazquez Garcia

Present work (from 2010)

Post-doc at the *Structural Biology and Biocomputing Programme*, *Spanish National Cancer Research Centre* (CNIO).

Past work

- **2005—2010**: Teaching Assistant, *Universidad Complutense*, Madrid
- **2003—2005**: Several positions as freelance/consultant in the software development industry

PhD degree:

PhD in Computer Science, *Universidad Complutense de Madrid* (Spain), July 2010: "Methods and Applications for Functional Analysis in Bioinformatics." Supervised by Alberto Pascal-Montano, Pedro Carmona-Saez, and Juan Pavón-Mestras. Qualification: Excellent *Cum Laude*.

Specialization:

- **Main interest**: data analysis
- **Specialties**: cancer genome analysis, functional analysis, systems biology
- **Additional interest**: software development

Overview of achievements

Co-authored 24 peer-reviewed articles in the fields of statistical learning, bioinformatics, and cancer research, several of them featured in high impact journals: *Nature Genetics* (2), *Nature Methods* (1), *Nucleic Acids Research* (5), *Bioinformatics* (1), among others. Statistics from *ResearchGate* (RG) and *Google Scholar* (GS): citations 498 (GS) and 115 (RG), h-index 9 (GS), total impact 164 (RG).

Reviewer for *Bioinformatics* and *BMC bioinformatics*, among other journals and conferences in bioinformatics and computer science. Co-organizer of the BioCreative III and IV community evaluation challenges. Program committee member in the *RECOMB Comparative Genomics* conference, Lyon, October 2013.

In addition to the teaching as *teaching assistant* at *Universidad Complutense*, taught courses at the *Advanced Statistics and Data Mining Summerschool* (organized by *Universidad Politécnica de Madrid*; 3 years) and participated in the *Master in bioinformatics and Computational Biology* (Co-organized by CNIO; sporadically).

VÍCTOR DE LA TORRE RUSSIS

PRESENT WORK (SINCE DECEMBER 2012)

General coordinator
National Bioinformatics Institute (INB), the Spanish node of ELIXIR
C/ Melchor Fernández Almagro, 3, E-28029 Madrid
E-mail: vdelatorre@cnio.es

PREVIOUS POSITIONS

- 2008-2012: Technician. Spanish National Cancer Research Centre
- 2006-2008: Senior Programmer & Google Analytics consultant. Trimedia e-Consulting.
- 2004-2006: Researcher. Cuban National Bioinformatics Centre.

PARTICIPATION IN PROJECTS

- 2010-current. RD-connect. An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research
- 2010-current. BLUEPRINT - A BLUEPRINT of Haematopoietic Epigenomes
- 2010.current. ASSET. Analysing and Striking the Sensitivities of Embryonal Tumours.
- 2009-current. e-TOX. Integrating bioinformatics and chemoinformatics approaches for the development of expert systems allowing the in silico prediction of toxicities
- 2009-2010. EUROCANCERCOMS. Establishing an efficient network for cancer communication in Europe
- 2009-current. MICROME. The Microme Project: A Knowledge-Based Bioinformatics Framework for Microbial Pathway Genomics
- 2008-2010. TARPOL. Targeting environmental pollution with engineered microbial systems à la carte.
- 2008-2010. Emergence. A Foundation for Synthetic Biology in Europe.

PUBLICATIONS

- Vazquez, M., de la Torre, V. & Valencia, A. Chapter 14: Cancer genome analysis. PLoS Comput. Biol. 8, e1002824 (2012).
- Azzaoui, K. et al. Scientific competency questions as the basis for semantically enriched open pharmacological space development. Drug Discov. Today 00, 1–10 (2013).
- Baudot et al. Mutated genes, pathways and processes in tumours. EMBO Rep (2010) vol. 11 (10) pp. 805-10
- Tendulkar et al. FragKB: structural and literature annotation resource of conserved peptide fragments and residues. PLoS ONE (2010) vol. 5 (3) pp. e9679

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Functional Consequences of Cancer Mutations using Structurally Annotated Protein Interaction Networks

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Prof. Dr. Christian von Mering
Institute of Molecular Life Science, University of Zurich, Switzerland and Swiss Institute of Bioinformatics
Sponsor: Dr. Jan Korbel
European Molecular Biology Laboratories, Heidelberg, Germany

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Dr. Abdullah Kahraman Institute of Molecular Life Science, U. of Zurich	

Background and preliminary data

Despite decades of research, the molecular mechanisms that lead from somatic mutations to complex diseases like cancer remains illusive. The utilization of protein interaction networks emerged as a powerful tool, however with the problem that interaction maps are often incomplete and are lacking protein structural data for functional interpretation of mutations. Recently, it has been suggested that new interactions could be predicted based on structural templates from protein complexes, which could extend current interaction maps far beyond their present coverage [1]. We therefore plan to take advantage of the structural protein data in the Protein Data Bank and extend the coverage of our STRING protein interaction database. STRING is one the most comprehensive protein-protein interaction databases with over 3500 citations from the scientific community.

One of the key applications of the new version of STRING will be the functional annotation of cancer mutation data. We will use structural data for the prediction of new interactions but we will also exploit structures to map mutations to structural interfaces in an unprecedented scale and resolution. Abdullah Kahraman, a new postdoc in my lab and the junior investigator in this proposal, is an expert in computational structural biology [2] and will lead this initiative.

We believe that the large number of whole cancer genomes in the PAN-CAN project and our efforts to map the interactome proteome-wide would mutually benefit from each other and lead to a better understanding of somatic mutations and their link to the cancerous state of cells.

[1] Zhang, Q. C. *et al. Nature* 490, 556–560 (2012).
[2] Herzog, F.*, Kahraman A.* *et al. Science* 337, 1348–1352 (2012).

Timelines & resources dedicated to project

The structural annotation of the STRING database is ongoing. We hope to have a full working annotation pipeline ready by February 2014. We have just updated our computing infrastructure with a new high-performance computing server to support the new developments. The mapping of mutations and their statistical analysis on a genome scale will likely start in March 2014.

Research proposal

Cancer mutations at coding regions from the PAN-CAN project will be mapped onto interacting proteins within the STRING database. To enable the mapping procedure, STRING proteins will be annotated with protein structures, which will be either downloaded from the Protein Data Bank/PDB (experimental) or ModBase (homology models), or if not existent predicted using homology modeling (HHpred + MODELLER) or *de novo* modeling (ROSETTA). Besides structurally annotating single proteins, protein complexes from the PDB will also be exploited as structural templates for predicting new interactions in STRING. The predictions will be based on a graph matching similarity search (ProBIS) between known domain interfaces in the PDB and proteins from STRING. To evaluate the predicted interactions, a structural interaction score will be developed, which measures the similarity of the predicted interface to known interfaces, the evolutionary conservation (ConSurf) of the interface amino acids and the geometrical and physiochemical complementarity within the interface. The structural interaction score will also serve as an additional scoring channel to the STRING database and reveal new and likely unexpected interactions between proteins.

As proteins in STRING will be structurally annotated, we will be able to make statements on the propensity of a mutation to be a driver mutation given its location on a protein (protein surface, protein core, secondary structure element), its functional impact (when located at interfaces, catalytic sites, post-translational modification sites, etc.) and its effect on protein complex stability. Furthermore, given that most human proteins interact with a multitude of other proteins via different interface region, we will also distinguish mutations based on their interface location. This distinction will allow us to study the impact of mutations on sub-networks and assess pleiotropic genes in cancer.

Furthermore, Gene Ontology enrichment studies of mutation data will be performed using the current Enrichment module of the STRING database, while protein complexes and interaction modules will be predicted with STRING's existing Clustering module.

All of the above mentioned computational analysis will be performed on the entire PAN-CAN dataset, with the aim to stratify tumor samples and identify protein, protein complex and protein network signatures. We hope that our stratification efforts will uncover network dependencies in the mutational spectrum and improve our understanding on somatic mutations and their phenotypic effects.

Legacy plans

The structurally predicted interactions will be accessible in STRING via its default web-interface at www.string-db.org. PAN-CAN annotations will be available in a optional data channel which users can turn on using the STRING "payload" mechanism.

Curriculum Vitae - Prof. Dr. Christian von Mering

Professor for Bioinformatics and Systems Biology
Institute of Molecular Life Sciences
University of Zurich
Winterthurerstrasse 190, 8057 Zurich

Email: mering@imls.uzh.ch
Website: http://www.imls.uzh.ch/vmering

Education / Employment

since Aug 2012:	Full Professor of Bioinformatics, University of Zurich, Switzerland.
Aug 2006 – July 2012:	Associate Professor of Bioinformatics, University of Zurich, Switzerland.
Sep 2001 – July 2006:	Post-Doc and Staff Scientist at the EMBL, Heidelberg (Peer Bork Group).
July 1998 – Aug 2001:	Ph.D. in Molecular Biology under the supervision of Prof. Dr. Konrad Basler awarded from the University of Zurich, Switzerland.

Honors / Functions

June 2008:	Associate Academic Editor at "PLoS Computational Biology"
May 2008:	Executive Board Member of "Swiss Institute of Bioinformatics"

Funding / Research Support

2010 – 2014:	ERC Starting Investigator Grant of 1.150.000 EUR over 5 years.
2008 – 2013:	Swiss National Science Foundation Principle Investigator grant of 450.000 CHF over 3 years and 537.000 over another 3 years.
2008 – 2016:	Swiss Institute of Bioinformatics Principle Investigator grant of 530.000 CHF over 4 years and 1.140.000 over another 5 years.
2008 – 2016:	SystemsX.ch Co-Investigator grant of 1.000.000 CHF over 9 years.

Selected Publications (* joint senior authorships)

J.F.M. Rodrigues, **C. von Mering** (2013). *HPC-CLUST: Distributed hierarchical clustering for very large sets of nucleotide sequences.* Bioinformatics, *in press*.

S. Powell, …, **C. von Mering***, P. Bork* (2013). *eggNOG v4.0: nested orthology inference across 3686 organisms.* Nucl Acids Res, *in press.*

A. Franceschini, …, C. **von Mering***, L.J. Jensen* (2013). *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucl Acids Res 41:D808-15.

M. Wang, …, **C. von Mering** (2012). *PaxDb, a database of protein abundance averages across all three domains of life.* Mol Cell Proteomics 11(8):492-500.

S. Powell, …, **C. von Mering***, P. Bork* (2012). *eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges.* Nucl Acids Res 40 (Database issue):D284-D289.

D. Szklarczyk, …, **C von Mering** (2011). *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res 39(1):D561-8.

B. Bodenmiller, …, **C. von Mering**, R. Aebersold (2010). *Phosphoproteomic Analysis Reveals Interconnected System-Wide Responses to Perturbations of Kinases and Phosphatases in Yeast.* Science Signalling 21;3(153):rs4.

S. Chaffron, …, **C. von Mering** (2010). *A global network of coexisting microbes from environmental and whole-genome sequence data.* Genome Research 20(7): 947-959

M. Weiss, …, **C. von Mering** (2010). *Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome.* Proteomics 10(6):1297-1306.

S.P. Schrimpf, …, **C. von Mering***, M.O. Hengartner* (2009). *Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes.* PLoS Biology Mar 3;7(3):e48.

L.J. Jensen, …, **C. von Mering** (2009). *STRING 8 - a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Res 37(D): p. D412-D416.

C. von Mering et al. (2007). *Quantitative phylogenetic assessment of microbial communities in diverse environments.* Science **315**(5815): p. 1126-30.

C. von Mering et al. (2002), *Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions.* Nature 417 (2002), p. 399–403.

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci*** **USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

Curriculum Vitae - Dr. Abdullah Kahraman

Senior PostDoc in the research group of Prof Dr. Christian von Mering
Institute of Molecular Life Sciences
University of Zurich
Winterthurerstrasse 190, 8057 Zurich
Email: abdullah.kahraman@uzh.ch
http://www.imls.uzh.ch/research/vonmering/people/abdullah-kahraman.html

Education / Employment

since July 2013:	Senior PostDoc at IMLS, University Zurich (Christian von Mering group).
Jan 2013 – June 2013:	Research Scientist at Institute of Biochemisty, ETHZ (Paola Picotti lab)
May 2009 – Dec 2012:	Post-Doc at IMSB at the ETH Zurich, Switzerland (Ruedi Aebersold lab).
July 2005 – April 2009:	Ph.D. in Computational Biology under the supervision of Prof. Dr. Dame Janet M. Thornton awarded from the University of Cambridge, UK.
Sept 2000 – Feb 2005:	Diploma in Bioinformatics (FH) at the University of Applied Sciences Giessen, Germany.

Honors

August 2011:	PostDoc of the Month awarded by postdocsforum.com
2005-2009:	EMBL Pre-doctoral Fellowship
2005:	FTG e.V. Certificate for obtaining the 2nd best diploma degree.

Publications (* joint authorships)

Kahraman, A. et al. (2013). *Cross-Link Driven Molecular Modeling with ROSETTA*. PLOS one e73411.

Herzog, F.*, **Kahraman, A.***, et al. (2012). *Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry*. Science 337, 1348–1352.

Kahraman, A. et al. (2011). *Xwalk: Computing and Visualizing Distances in Cross-linking Experiments*. Bioinformatics 27, 2163-2164.

Leitner, A., **Kahraman, A.***, Walzthoeni, T.* et al. (2010). *Probing native protein structures by chemical cross-linking, mass spectrometry and bioinformatics*. Mol Cell Proteomics 9, 1634-1649.

Smith, L., **Kahraman, A.**, Thornton, J. M. (2010). *Heme proteins - diversity in structural characteristics, function and folding*. Proteins 78, 2349-2368.

Kahraman, A. et al. (2010). *On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins*. Proteins 78, 1120-1136.

Kahraman, A., Thornton, J. M. (2008). *Methods to Characterize the Structure of Enzyme Binding Sites*. In Computational Structural Biology - Methods and Applications (Eds. Schwede, T. & Peitsch, M. C.), Vol. 1, pp. 189-221. 1 vols. World Scientific Publishing Co.

Kahraman, A., et al. (2007). *Variation of geometrical and physicochemical properties in protein binding pockets and their ligands*. BMC Bioinformatics 8, S1.

Kahraman, A. et al. (2007). *Shape variation in protein binding pockets and their ligands*. J Mol Biol, 368, 283-301.

Morris, R. J., **Kahraman, A.**, Thornton, J. M. (2005). *Binding Pocket Shape Analysis for Protein Function Prediction*. Acta Crystallographica Section A 61, C156–157.

Morris, R. J., …, **Kahraman, A.**, Thornton, J. M. (2005). *Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons*. Bioinformatics 21, 2347-2355.

Kahraman, A. et al. (2005). *PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics*. Bioinformatics, 21, 418-20.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of the functional information of somatic non coding variants

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Ewan Birney, Member of BASIS (ICGC Breast Cancer Consortium)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Ian Dunham, Sandro Morganella	

Background and preliminary data

There has been a growing catalog and investigation of non coding sequence function, derived from chromatin immunopreciptation experiments on transcription factors and histone modifications (Chip-seq), DNaseI hypersensitivity analysis (DNaseI-seq), RNA analysis of a variety of molecular forms (RNA-seq, including long, short, non-poly-A and chromatin associated RNA). These studies occur in both cancer derived laboratory cell lines and in different aspects of normal or normal-like cells and tissues. Large projects such as ENCODE [1], Epigenome Roadmap [2,3] and International Human Epigenome Consortium (IHEC) [4] projects all contribute this information.

The integration of this information has already been shown to be a powerful resource for germline genetic association [1,5-11] and people have already shown that there is a non-random association of somatic variants with these features . However, it is less clear what is the underlying basis for this non-random association (at least some is likely to be due to mutational biases in cancers). A credible hypothesis is that a subset of somatic non-coding variants contributes in important ways to the cancer phenotype.

1 The_ENCODE_Project_Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74, doi:10.1038/nature11247 (2012).
2 Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology 28, 1045-1048, doi:10.1038/nbt1010-1045 (2010).
3 Chadwick, L. H. The NIH Roadmap Epigenomics Program data resource. Epigenomics 4, 317-324, doi:10.2217/epi.12.18 (2012).
4 Bae, J. B. Perspectives of international human epigenome consortium. Genomics & informatics 11, 7-14, doi:10.5808/GI.2013.11.1.7 (2013).
5 Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome research 22, 1790-1797, doi:10.1101/gr.137323.112 (2012).
6 Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49, doi:10.1038/nature09906 (2011).
7 Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. Cell 152, 1237-1251, doi:10.1016/j.cell.2013.02.014 (2013).
8 Maurano, M. T., Wang, H., Kutyavin, T. & Stamatoyannopoulos, J. A. Widespread site-dependent buffering of human regulatory polymorphism. PLoS genetics 8, e1002599, doi:10.1371/journal.pgen.1002599 (2012).
9 Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. Genome research 22, 1748-1759, doi:10.1101/gr.136127.111 (2012).
10 Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature genetics 45, 124-130, doi:10.1038/ng.2504 (2013).
11 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research 40, D930-934, doi:10.1093/nar/gkr917 (2012).

Timelines & resources dedicated to project

We will dedicate one PostDoc (Sandro Morganello), and partial effort from Ian Dunham and Ewan Birney in this effort. We have preliminary working routines and assessment, developed as part of the BASIS project for the investigation.We will be able to coordinate closely with any specific compute components hosted at the EBI.

Research proposal

We propose to first explore and then quantify the relationship between non-coding somatic variants, other features of cancer biology (such as tissue of origin, stage and protein-coding gene variants) and other functional annotations from ENCODE, Epigenome Roadmap and IHEC. We will use only published datasets (of which there are ample now, and more coming during the future years), mainly to ensure that there is no complication in publication.

A key component of this investigation will be to understand and account for mutational biases. We expect that there will be a heterogeneous range of mutational processes across the genome in each cancer sample, and that these changes are non-random with respect to both large scale and small scale chromatin/genome features (from DNA replication timing to fine grained nucleosome positioning). There are two direct of benefits of viewing the mutational biases through the filter of known functional annotation. Firstly that this might inform the differences in mutation and repair processes active in different cancers, and secondly that this is the key background process to model for association with other features.

Having explored this mutational bias (and we expect to participate and collaborate with other groups with stronger expertise in the mutational modeling components) we will then turn to trying to find any evidence of cancer phenotype drivers present in the data over and above this bias. This will include the association with specific transcription factor binding sites and/or gene function, utilizing the different ways to link non-coding regions in the genome to protein coding genes including 3/4/5C, ChIA-PET and cross-sample inter-site correlation.

This analysis is distinct from, but highly synergistic with, investigation of functional information directly from cancer samples,in particular mRNA sequences and methylation data. We would be delighted to collaborate with groups working in these areas, but this analysis would focus primarily on integrating somatic variation with non-coding functional information.

Legacy plans

We will adhere to all consortium codes of practice. We will participate in appropriately storing our intermediate results and statistical tests in line with pan Cancer consortium agreements; large scale pipelines might be in an more exemplar form as we did in the ENCODE virtual machine.

Curriculum Vitae, Ewan Birney

Full Name: John Frederick William Birney
Date of Birth: 12 December 1972
Nationality: UK
Email: birney@ebi.ac.uk

77 Lancaster Road
London N4 4PL

Employment:

2012-Current : Associate Director, European Bioinformatics Institute
2000-2012: Head of Nucleotide data, European Bioinformatics Institute
Current supervisor for 4 PhD students

On a variety of SAB boards (includes Riken Institute, BCGSC, Leipzig MPI, Roslin Institute, IMP, TGAC)

1996-2000: PhD at the Sanger Centre (Supervisor, Richard Durbin)

Other positions held:

- A number of consultancy contracts, both strategic and technical in the biotech and pharmaceutical industry, including funding and finance orientated roles.
- Equity Research in SBC Warburg Pharmaceutical division (summer 1995).
- Freelance journalist (Economist) (1995).
- Research Assistant at Cold Spring Harbor Laboratory and EMBL Heidelberg.

Prizes and Awards

EMBO Member, Elected 2012
Winner of the Overton Award from the International Computational Biology Society, 2005
Winner of the Benjamin Franklin Award from Bioinformatics.org/BioIT in 2005
Winner of the Royal Society's Francis Crick Lecture in 2003

Patents:

US Provisional Patent Application 61/654295, *High-capacity storage of digital information in DNA*, filed 1 June 2012 (co-applicant with Nick Goldman)

Education:

1996-1999: PhD, St John's College Cambridge. Awarded a Scholarship
1992-1996: BA Biochemistry, Balliol College Oxford. 1st Class degree. Awarded a Scholarship

Publications

181 Peer reviewed publications, 23 in Nature (5 first/last author), 9 Science (1 last author). 1 Cell (joint last author). H-index: 83. Avg Citations/Paper 331. (Google Scholar)

Ian Dunham M.A. D.Phil

Ian Dunham
Staff Scientist
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD

Tel: 01223 492636
Fax: 01223 494468
Email: dunham@ebi.ac.uk
Citizenship: UK
DoB: 29th June 1963

EDUCATION
October 1981 – June 1985 - University of Oxford, B.A. (Hons) in Biochemistry, Class II.
October 1985 – September 1988 - University of Oxford, D. Phil. Thesis title: "Molecular Mapping of the Human Major Histocompatibility Complex." Mentor: R. Duncan Campbell.

c. Postdoctoral fellowship(s)

1989-1990 - Howard Hughes Medical Institute, Washington University Medical School, Dept. of Genetics, St. Louis, MO 63110 USA. Postdoctoral research associate in genetics. Funding HHMI. Mentor: Maynard V. Olson.

1990-1991 - Division of Medical and Molecular Genetics (Paediatric Research Unit), UMDS Guys Campus, London. Postdoctoral Research Fellow, Funding UMDS. Mentor: David R. Bentley.

1991-1993 - Division of Medical and Molecular Genetics (Paediatric Research Unit), UMDS Guys Campus, London. Wellcome Trust Postdoctoral Research Fellow. Funding Wellcome Trust. Mentor: David R. Bentley.

ACADEMIC APPOINTMENTS

April 2012 – Current – Staff Scientist, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK.

1993-1995 - Group Leader, The Sanger Centre, Hinxton, UK.
1996 – 2003 - Senior Group Leader/Senior Research Fellow, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK.
2003-Sept 2007 - Senior Investigator, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK.
Oct 2007 – July 2008 - Ensembl Developer, Regulation Team, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK.

PUBLICATIONS: 132 Publications, 11 Nature (3 first/last author), 1 Science, 4 Nature Genetics. 29,711 Citations, H-index 49 (Google Scholar)

Sandro Morganella M.A. D.Phil

Sandro Morganella

Post-doctoral Fellow

European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD

Tel: 01223 20513

Email: sandro@ebi.ac.uk

Citizenship: Italy
DoB: 26th June 1980

EDUCATION

September 2002 – October 2008 - University of Sannio, M.S. in Computer Science.
April 2009 – July 2012 - University of Italy, D. Phil. Thesis title: "Downstream Analysis of Microarray Copy Number Data ." Mentor: Michele Ceccarelli.

c. Postdoctoral fellowship(s)

2012 – Outgoing - EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. Mentor: Ewan Birney.

ACADEMIC APPOINTMENTS

Sept 2005 – Dec 2005 – Research Activity in Medical Image Analysis, University of Sannio, Benevento, Italy.
Nov 2007 – Nov 2008 - Research Activity on the field of distance learning (*e*-learning), University of Sannio, Benevento, Italy.
Jul 2009 – Nov 2009 – Security and Management of RFID Technology Events, University of Sannio, Benevento, Italy.

RECENT SCIENTIFIC PUBLICATIONS

19. Morganella S., Ceccarelli M. VegaMC: an R/Bioconductor package for fast downstream analysis of large array comprative genomic hybridization datasets. *Bioinformatics* **28**(19):2512-2514 (2012). doi: 10.1093/bioinformatics/bts453. PMID: 22815357.

20. Morganella S., Pagnotta S.M., Ceccarelli M. Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics* **22**(21):2949-2956 (2011). DOI: 10.1093/bioinformatics/btr488 . PMID: 21873327.

21. Morganella S., Cerulo L., Viglietto G., Ceccarelli M. VEGA: Variational segmentation for copy number detection. *Bioinformatics* **26**(25):3020-3027 (2010). doi: 10.1093/bioinformatics/btq586 . PMID: 20959380.

22. Zoppoli P., Morganella S., Ceccarelli M. TimeDelay-ARACNE: Reverse en- gineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*: **11**:154 (2010). doi: 10.1186/1471-2105-11-54 . PMID: 20338053.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrative pan-cancer transcriptomics analysis and links to genetics

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Alvis Brazma (EMBL, CAGEKID), Jan Korbel (EMBL)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Liliana Greger (EMBL), Mar Gonzàlez-Porta (EMBL)	**Oliver Stegle (EMBL)**

Background and preliminary data

The Brazma lab has broad experience in RNAseq data analysis, in particular we have developed and are running a pipeline at EBI to reprocess all public RNAseq data for our Expression Atlas (www.ebi.ac.uk/gxa/). We are leading the transcriptomics analysis of an ICGC project on clear cell Renal Cell Carcinoma (CAGEKID; coordinated by CEPH and McGill). CAGEKID has produced genome sequencing data from normal/cancer pairs of 80 patients, and RNAseq data from 45 pairs. These data have been deposited at EGA, while the metadata will be submitted to ICGC data portal for the next release. We have also been involved in data analysis and management of GEUVADIS, a large-scale transcriptomics project (RNA sequencing of 465 lymphoblastoid cell lines; Lappalainen T et al, Nature, 2013). In both projects we have developed pipelines to identify differential expression profiles, splicing alterations and fusion genes (Gonzàlez-Porta et al., Genome Biology, 2013, Lappalainen T et al, Nature, 2013). Using those pipelines we identified novel fusion genes in ccRCC (Integrative analysis of clear cell Renal Cell Carcinoma; *in preparation*) and in lung cancer kinome data (Majewski et al., 2013). Furthermore, we focused on expression patterns of major transcripts and identified cases in which, for a given gene, there was a switch in major transcript identity across conditions (switch events). Finally, we have also implemented tools to integrate our results with those obtained from somatic mutation analyses, revealing that even though most genes are not altered in a recurrent fashion, pathways are broadly disrupted when combining information across multiple patients.

The Stegle lab has extensive experience in the genetic analysis of molecular traits. In particular, we have pioneered methodology to handle heterogeneous expression datasets where confounding factors such as hidden batch structure, genotype structure or subtle environmental variations are present (e.g. Stegle O et al, Nat. Protoc 2012). These approaches leverage high-dimensional expression data to automatically learn confounding sample structure. Applications in leading studies include HapMap (Stranger B et al, PLoS Genet 2012) and transcription analyses of 1,000 genomes individuals (Lappalainen T et al, Nature, 2013), greatly increasing the power compared to conventional analysis approaches. Ongoing work includes some pan-cancer transcription analyses in the context of TCGA (Lehmann et al, in preparation, lead by G. Rätsch MSKCC), investigating the genetic control of splicing phenotypes by germline variants and somatic mutations.

Timelines & resources dedicated to project

Our RNAseq analysis pipelines already been have been developed and tested. They are based on paired sample analysis and they are cancer type agnostic, therefore they can be applied for cross-cancer analysis with few alterations. We will be able to distribute our analysis software to all analysis centers using virtual machines and additionally use EBI computing capacity for the required analyses, for samples distributed to the EBI pan-cancer computing node. We expect to utilize 90,000 CPU time for 1,500 RNA samples and to be able to start the analysis of at least some cancer types as soon as the project starts and we hope to be able to produce results in maximum 6 months. In terms of human resources, this will be the main project for postdoc Dr. Greger. A PhD student Gonzàlez-Porta will spend 50% of her time on the project, Brazma 20% of his time. If necessary we will tap into other resources in Brazma team, in particular we will utilize expertise of Dr. Nuno Fonseca. Overall, in Brazma group this will be a research project of the highest priority. These resources will be complemented by substantial time investment of the Stegle research group. Paolo Casale, a student in the lab will work on this full time, Stegle to contribute 50% of his time to contribute many of the genetic analyses directly.

Research proposal

We focus on RNAseq data analysis, but in conjunction with the mutation calls and structural rearrangement data provided by others. In particular, we plan to interface this proposal with the plans to investigate the interface between germline and somatic genetic variation lead by Jan Korbel, which will provide important information on genetic factors that can be correlated with changes in transcription (see objective II).

First, we aim at applying our established analysis pipelines to assess the differences between normal and tumour samples and characterise lowly recurrent events, based on paired analysis that we developed for CAGEKID. Specifically we will focus on understanding:

1) expression and splicing alterations, which can be further classified as tumour-specific or shared across multiple cancer types; in particular, we will focus on the switch events in normal vs tumour samples;
2) novel cancer specific transcripts and ncRNAs;
3) high confidence sets of novel fusion genes.

Taken together, these analyses will allow us to characterize and compare transcriptomics signatures in different cancer types, which can then be integrated with available genomics, epigenomics and clinical data to reveal cancer molecular signatures on a wider dimension and provide drug targets available for the community. Individual events will be rigorously tested for either being cancer specific or carrying pan-cancer signatures across groups of cancer types.

The second objective is to use the quantified expression and splicing events for pan-cancer genetic analyses. We will leverage the full-genome sequencing data, which is unique to the ICGC initiative. Specifically, we aim to:

1. Use full genome genetic sequencing data in conjunction with our established eQTL pipelines to identify germline eQTLs with *cis* and *trans* mechanisms. Trans association scans will be carried out unbiased and focused on selected GWAS loci to mitigate the burden of multiple testing. We will integrate information from other initiatives that link germline and somatic variation (see above).
2. Leverage the detailed genetic map of germline variability to increase the sensitivity of association tests with somatic mutations, considering local and distal genetic effects on gene expression and splicing.
3. Integrate known annotations and accessibility information from ENCODE to aggregate rare somatic variants for pathway-based and group association analyses.
4. Ty together association tests in individual cancers using both meta-analysis approaches and joint analyses in one model.

Taken together, these analyses will yield the first comprehensive map of germline and somatic genetic variations that drive molecular changes in cancer genomes. While we start with proximal, likely *cis* candidate tests, the important aim is to extend association tests to distal regions, fully leveraging full sequencing information of ICGC samples. Our approaches to increase power and robustness with respect to confounding variation and heterogeneity will be instrumental for this work.

Legacy plans

The differential gene and transcript expression results will be disseminated via the EBI Expression Atlas and linked from ENSEMBL. Any genetic associations identified will be shared using the EBI variation resources. Valuable intermediate results, such as normalized and batch corrected expression datasets will be shared with other ICGC participants. The final results will be visualized on a Data browser tailored to the project as described previously in Lappalainen T et al, Nature, 2013 (http://www.ebi.ac.uk/Tools/geuvadis-das/).

CURRICULUM VITAE **Alvis Brazma**

BSc Mathematics, University of Latvia, 1982
PhD Computer Science, Moscow State University, 1987

Current position and projects

Alvis Brazma is a Senior Scientist at EMBL and a Senior Team Leader at EMBL-EBI in charge of Functional Genomics Resources. His research interests focus on integrative data analysis, RNAseq analysis methods, and human transcriptome analysis. He leads the transcriptome data analysis for the EC funded project on renal cancer CAGEKID, which is a part of ICGC.

Selected Publications

Lappalainen T et al, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013 Sep 26;501(7468):506-11

Stefflova K et al, Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*. 2013 Aug 1;154(3):530-40

Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol*. 2013 Jul 1;14(7):R70.

Majewski IJ et al, Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing. *J Pathol*. 2013 Jul;230(3):270-6.

Rung J., Brazma A. Reuse and meta-analysis of public gene expression data. *Nature Reviews Genetics*, 2013 Feb;14(2):89-99.

Rustici G, et al. ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013 Jan 1;41(D1):D987-90.

Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012 Dec 1;28(24):3169-77.

ENCODE Project Consortium, Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74

Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res*. 2012 Dec;22(12):2376-84.

Xue V, Burdett T, Lukk M, Taylor J, Brazma A, Parkinson H. MageComet--web application for harmonizing existing large-scale experiment descriptions. *Bioinformatics*. 2012 May 15;28(10):1402-3.

Caldas J, Gehlenborg N, Kettunen E, Faisal A, Rönty M, Nicholson AG, Knuutila S, Brazma A, Kaski S. Data-Driven Information Retrieval in Heterogeneous Collections of Transcriptomics Data Links SIM2s to Malignant Pleural Mesothelioma. *Bioinformatics*. 2012 Jan 15;28(2):246-53.

Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A, Parkinson H. The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res*. 2012 Jan;40(Database issue):D64-70. doi: 10.1093/nar/gkr937.

Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N, Kurnosov P, Malone J, Melnichuk O, Petryszak R, Pultsin N, Rustici G, Tikhonov A, Travillian RS, Williams E, Zorin A, Parkinson H, Brazma A. Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2012 Jan;40(Database issue):D1077-81.

Goncalves A, Tikhonov A, Brazma A, Kapushesky M. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*. 2011 Mar 15;27(6):867-9.

Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A. A global map of human gene expression. *Nature Biotechnology*. 2010 Apr;28(4):322-4.

CURRICULUM VITAE – Dr. rer. nat. Dipl.-Ing. Jan O. Korbel

Group Leader / Principal Investigator
Genome Biology Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstr. 1, Heidelberg, Germany

Secondary affiliation:
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus, Hinxton, UK
Email: korbel@embl.de

Academic Education & Qualification

Since 2013	European Research Council (ERC) Principal Investigator at EMBL Heidelberg.
Since 2008	Group Leader / Principal Investigator at EMBL Heidelberg, in the Genome Biology Unit.
2005-2007	Postdoc at Yale University, New Haven, CT, with Mark Gerstein & Michael Snyder.
2005	PhD Molecular Biology, specialization Computational Biology, awarded from Humboldt-University Berlin & EMBL Heidelberg. PhD research mentor: Peer Bork.

Leadership in International Research Consortia

Since 2013	Steering Group Member: WGS Pan-Cancer Analysis Project.
Since 2011	Steering Group Member: 1000 Genomes Project.
Since 2011	Co-chair leading the Structural Variation Analysis Group of the 1000 Genomes Project.

Other Professional Experience

2013	Session chair, Annual Conference of American Association for Cancer Research (AACR).
2013	Session chair, Biology of Genomes Meeting, Cold Spring Harbor Laboratory.
2013	Organizing committee, 2nd EMBL Conference on Cancer Genomics.
Since 2012	Advisory board member, ICGC-affiliated "Small-Cell Lung Cancer Genome Project".

Selected Recent Publications (*joint senior authorships)

Korbel JO* & Campbell PJ* (2013). Criteria for inference of chromothripsis in cancer genomes. ***Cell*** 152:1226-36.

Korbel JO & Lee C (2013). Genome assembly and haplotyping with Hi-C. ***Nat Biotechnol***, *in press* [News & Views].

Weischenfeldt J, …, **Korbel JO*** & Schlomm T* (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** 23:159-70.

Gokcumen O, …, **Korbel JO** (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. ***Proc Natl Acad Sci*** **USA** 110(39):15764-9.

Weischenfeldt J, …, **Korbel JO** (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. ***Nat Rev Genet*** 14:125-38 [Review].

Rausch T, …, **Korbel JO** (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. ***Cell*** 148:59-71.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. ***Nature*** 491:56-65.

Mills RE, …, **Korbel JO**; for the 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. ***Nature*** 470:59-65.

Stewart C, …, **Korbel JO** & Marth GT; for the 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. ***PLoS Genet*** 7:e1002236.

Schlattl A, …, **Korbel JO** (2011). Relating CNVs to transcriptome data at fine-resolution: Assessment of the effect of variant size, type, and overlap with functional regions. ***Genome Res*** 21:2004-13.

Lam HY, …, **Korbel JO*** & Gerstein MB* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. ***Nat Biotechnol*** 28:47-55.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. ***Nature*** 467:1061-73.

Kasowski M, …, **Korbel JO*** & Snyder M* (2010). Variation in transcription factor binding among humans. ***Science*** 328:232-5.

LILIANA STALEVA GREGER

Address 31 Alice Bell Close, Cambridge, CB4 1GN
Tel (home) 01223 424011/079 82899997
E-mail lgreger@ebi.ac.uk
Nationality Bulgarian (EU citizen)

EDUCATION:

2005-2009	**MSc in Bioinformatics** Manchester University, Faculty of Life Sciences by distant learning (with distinction)
1994-1998	**PhD in Molecular Biology** Institute of Molecular Biology, Department of Genetics, Bulgaria, laboratory of Prof. Pencho Venkov
1997-1998	**Visiting PhD Student** National Institute of Health, National Cancer Institute, USA laboratory of Dr. David Garfinkel
1992-1993	**MSc in Gene and Cell Engineering** Sofia University St. Kliment Ohridski, Faculty of Biology, Bulgaria Average grade from the State Examinations 5.50 (out of 6)
1986-1991	**BSc in Biotechnology** Sofia University, Faculty of Biology, Bulgaria

PROFESSIONAL EXPERIENCE:

2010 – present	**Bioinformatician, Postdoctoral Fellow**, EMBL-EBI
2008 – 2010	**Bioinformatician**, Cambridge University, Department of Pathology
2003-2008	Career break (raised children) Studied for MSc degree in Bioinformatics
1999-2003	**Post-Doctoral Fellow**, New York University, School of Medicine, Dermatology Research Laboratories, laboratory of Prof. Seth J. Orlow, New York, USA

AWARDS:

1997-1998	Wood and Whelan Research Fellowship from the International Union of Biochemistry and Molecular Biology
2008-2010	Daphne Jackson Trust fellowship for returners to science, engineering and technology after a career break

PATENT:

Orlow S, Manga P and Staleva L., U. S. Patent "Assay for Melanogenesis", WO/2003/095645, World Intelectual Property Organization

PUBLICATIONS:

Lappalainen T,.. **Greger L** et. al. Transcriptome and genome sequencing uncovers functional variation in humans. (2013). Nature 501:506-11

Majewski IJ, ..**Greger L** et al. Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing. (2013). J Pathol. 230:270-6

Cooke J, Zhang H, **Greger L**, Silva AL, Massey D, Dawson C, Metz A, Ibrahim A, Parkes M.

Mucosal genome-wide methylation changes in inflammatory bowel disease. (2012). Inflamm. Bowel Dis. 18:2128-37

Ibrahim AE, Arends MJ, Silva AL, Wyllie AH**, Greger L**, Ito Y, Vowler SL, HuangTH, Tavaré S, Murrell A, Brenton JD. Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. (2011). Gut 60:499-508

Staleva L and Orlow SJ. Ocular albinism 1 protein: Trafficking and function when expressed in Saccharomyces cerevisiae. (2006). Exp. Eye Res. 82:311-318

Pesheva M, Krastanova O**, Staleva L**, Dentcheva V, Hadzhitodorov M, and Venkov P. The Ty1 transposition assay: a new short-term test for detection of carcinogens. (2005). J. Microbiol. Methods 61:1-8

Staleva L, Hall A. and Orlow SJ. Oxidative stress activates FUS1 and RLM1 transcription in the yeast Saccharomyces cerevisiae in an oxidant-dependent manner. (2004). Mol Biol Cell. 15:5574-5582

Staleva L, Manga P, and Orlow SJ. Pink-eyed Dilution Protein Modulates Arsenic Sensitivity and Intracellular Glutathione Metabolism. (2002).Mol. Biol. Cell 13: 4206-4220

Staleva L, and Venkov P. Activation of Ty transposition by mutagens. (2001). Mutat. Res. 474 :93-103

Staleva L, Gugova R, Venkov P, Waltscheva L, and Golovinsky E. Genotoxic effect of 4-aroyl-1-(2-chloroethyl)-1nitrosohydrazinecarboxamides on *Saccharomyces cerevisiae* cells. (1998). J Cancer Res Clin Oncol. 124: 321-325

Staleva L, Waltscheva L, Golovinsky E, and Venkov P. Enhanced cell permeability increases the sensitivity of a yeast test for mutagens. (1996). Mutat Res. 370: 81-89

MAR GONZÀLEZ-PORTA

mar@ebi.ac.uk | http://www.ebi.ac.uk/~mar

EDUCATION

UNIVERSITY OF CAMBRIDGE & EMBL-EUROPEAN BIOINFORMATICS INSTITUTE, Cambridge, UK
PhD in Bioinformatics, October 2010 – 2014 (expected)

UNIVERSITY ROVIRA I VIRGILI, Tarragona, Spain
BSc in Biotechnology, September 2006 – April 2010, best academic record

RESEARCH EXPERIENCE

EMBL-EUROPEAN BIOINFORMATICS INSTITUTE, Cambridge, UK
Predoctoral Fellow, Functional Genomics Group

since 10/2010 — RNA sequencing for the study of splicing – supervised by Dr. Alvis Brazma

- Investigated transcriptome diversity and the extent to which it could lead to protein diversity.
- Explored the role of splicing in cell cycle progression and applied innovative analysis strategies based on intronic reads.
- Studied splicing alterations in renal cancer as part of an ICGC project (CAGEKID).
- Developed a tool to visualize changes in splicing across conditions (LOREM - http://tinyurl.com/kf4gkgd).
- Explored the integration of transcriptomics (RNA-seq) and proteomics (SWATH-MS) data for the validation of predicted splicing changes.

CENTER FOR GENOMIC REGULATION, Barcelona, Spain
Intern, Computational Biology of RNA Processing Group

06/2010 – 09/2010 — Investigated differences in gene expression and alternative splicing in human populations by analysing RNA-seq data.

LAUSANNE UNIVERSITY & SWISS INSTITUTE OF BIOINFORMATICS, Lausanne, Switzerland
Intern, Evolutionary Bioinformatics Group

06/2009 – 09/2009 — BSc thesis: Evolutionary insights into metazoan miRNA expression and function

UNIVERSITY ROVIRA I VIRGILI, Tarragona, Spain
Student Researcher, Biochemistry and Biotechnology Department

2008 – 2010 — Developed tools to determine the interplay between CpG islands and miRNAs in the regulation of gene expression in humans.

PUBLICATIONS

1. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., **Gonzàlez-Porta, M.**, et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), 506–511.
2. **Gonzàlez-Porta, M.**, Frankish, A., Rung, J., Harrow, J., & Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome biology, 14(7) R70.
3. Roux, J., **Gonzàlez-Porta, M.**, & Robinson-Rechavi, M. (2012). Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs. Nucleic acids research, 40(13), 5890–5900.
4. **Gonzàlez-Porta, M.**, Calvo, M., Sammeth, M., & Guigó, R. (2012). Estimation of alternative splicing variability in human populations. Genome research, 22(3), 528–538.

Personal Details

Voice/Email/WEB +44 (0) 1223 494 101 / Oliver.stegle@ebi.ac.uk / www.ebi.ac.uk/stegle

Date of Birth 08th August 1981

Citizenship Germany / EU

Education & Academic Employments

Nov 2012- ***EMBL-European Bioinformatics Institute, Hinxton, Cambridge, UK***
Research Group Leader

June 209-October 2012 ***Max Planck Institute for Developmental Biology & MPI for Intelligent Systems, Tübingen, Germany***
Postdoctoral researcher, fellowships from Volkswagen Foundation & Marie Curie Intra European FP7

2005 –2009 ***Ph.D. Physics,* University of Cambridge, UK**
Advisor: Prof. David J.C. MacKay
Fellowships from the German National Academic Foundation & the Cambridge Gates Trust.

–2005 ***Master of Advanced Studies in Mathematics (MAST),* University of Cambridge, UK**
Theoretical physics: Quantum Field theory, Quantum Information Theory, Statistical Field Theory.

Selected publications

*denotes equal contribution. Selected from more than 25 publications in total, see http://scholar.google.com/citations?user=ClSXZ4IAAAAJ .

Lappalainen, Tuuli, Sammeth Michael, et al., …, **Oliver Stegle**, …, Dermitzakis ET *Transcriptome and genome sequencing uncovers functional variation in humans.* ***Nature 2013***

Julien Gagneur*, **Oliver Stegle***, Zhu Chenchen, Jakob Petra, Tekkedil Manu M., Aiyar Raeka S., Schuon, Ann-Kathrin, Pe'er Dana, and Steinmetz Lars M. *Interactions Reveal Causal Pathways That Mediate Genetic Effects on Phenotype.* ***PLoS Genetics 2013***

Nicolo Fusi, Christoh Lippert, Karsten Borgwardt, Neil Lawrence**, Oliver Stegle.** Detecting regulatory gene-environment interactions with unmeasured environmental factors**.** ***Bioinformatics 2013***

Oliver Stegle*, Leopold Parts*, Matias Piipari, John Winn, Richard Durbin *Using Probabilistic Estimation of Expression Residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* ***Nature Protocols 2012***

Xiangchao Gan*, **Oliver Stegle***, et al., …, Richard M. Clark, Gunnar Rätsch, Richard Mott *Multiple reference genomes and transcriptomes for Arabidopsis thaliana***. Nature 2011**

Leopold Parts*, **Oliver Stegle***, John M. Winn, Richard Durbin *Joint Genetic Analysis of Gene Expression Data with Inferred Cellular Phenotypes.* ***PLoS Genetics 2011***

Oliver Stegle*, Leopold Parts*, Richard Durbin, John M. Winn *A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies.* ***PLoS Computational Biology 2010***

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

How many somatic mutations drive cancer?

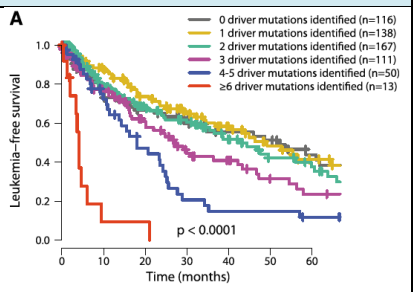
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Peter Campbell, Wellcome Trust Sanger Institute (ICGC breast, bone and chronic myeloid cancers)
Michael Stratton, Wellcome Trust Sanger Institute (ICGC breast, bone and chronic myeloid cancers)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Inigo Martincorena, Sanger Institute Kevin Dawson, Sanger Institute	

Background and preliminary data

We do not know how many somatic mutations it takes to drive a cancer. Estimates of this number in the literature are generally vague, but have typically fallen in the range 1-10. We have shown that the number of driver mutations correlates with outcome in myelodysplastic syndromes (Figure), suggesting that this is a clinically meaningful question. The collection of 2,000 whole cancer genomes enables us to address this fundamental question of cancer biology with complete data and sufficient sample size to enable robust estimation.



Recurrence, above that expected by chance, has been the fundamental principle underpinning the identification of driver mutations in cancer genomes for the last 20 years. This approach requires an appropriate model of the expected ('background' or 'neutral') mutation rate – we have developed maximum likelihood and Bayesian models of the ratio of non-synonymous to synonymous mutations (dN/dS) for coding substitutions – this work started with kinase screens (Greenman et al, Nature 2006), but has continued into the exome era. Run across the TCGA exomes, we identify many known and novel candidate cancer genes (Table for top 7 tumour suppressor genes). This approach is rather flexible – one can estimate the excess of missense mutations at a particular hotspot; the excess of nonsense mutations or splice site mutations (as shown in Table); or the excess of missense mutations in the whole gene footprint. For example, for *CASP8*, the dN/dS ratio is estimated at 3.4, suggesting that 3.4/4.4 (77%) missense mutations are present in excess. With 81 missense mutations observed, this means that ~60 are likely to be driver events.

We believe this principle, identifying a background, expected rate for a given mutational process coupled with quantifying the statistical excess of a given functional consequence, provides a general framework for estimating the number of driver mutations in a given cancer genome.

gene_name	cds_ID	n_all	n_syn	n_mis	n_non	n_splice	wMIS	wNON	pMIS	pNON	qMIS	qNON
CDKN2A	CCDS56565	102	12	52	27	11	2.164364995	26.24007821	0.009399426	0	0.317962046	0
CASP8	CCDS42798	134	10	81	41	2	3.408023508	18.10118086	1.98E-05	0	0.018233961	0
ARID1A	CCDS285	457	65	208	183	1	1.351632185	16.08414728	0.029667264	0	0.462436886	0
TP53	CCDS11118	1285	83	958	227	17	4.671559901	14.7615248	0	0	0	0
PBRM1	CCDS43099	175	20	100	54	1	1.685263607	9.815896272	0.024228486	0	0.438983626	0
ARHGAP35	CCDS46127	195	27	118	50	0	1.810880317	9.713587122	0.003156057	0	0.226050513	0
ARID2	CCDS31783	266	34	162	68	2	1.59655644	7.007725476	0.009179523	0	0.316717085	0

Timelines & resources dedicated to project

Months 1-3: Defining the ontology of functional consequences associated with somatic mutations
Months 1-12: Annotation of somatic mutations from 2,000 cancer genomes according to functional ontology
Months 1-12: Development of statistical models for background mutation rate
Months 1-15: Development of framework for estimating excess recurrence of mutations for each class
Months 9-15: Integration across 2,000 patients with analysis of distribution by tumour type, category of mutational process, etc.

International Cancer Genome Consortium

Research proposal

We aim to generate estimates of the number of driver mutations per patient across 2,000 whole cancer genomes. This will be achieved by, firstly, attributing a series of functional consequences to each somatic mutation; secondly, defining a model for the background rate of passenger mutations across functional consequences; and, thirdly, developing a statistical framework to estimate (and count) the excess mutations with observed functional consequence over that expected for the background model.

Phase 1 – An ontology of functional consequences: We recognise that driver mutations come in many different guises, with many different effects. Inference of driver status therefore requires us to identify excess recurrence across multiple scales. Borrowing from the rationale of the GO classification of gene function, we will define a hierarchical ontology of functional consequences for mutation categories. The classification structure of a V600E mutation in *BRAF*, for example, might be "T>A chr7:140453131" < "V600E *BRAF*" < "Missense kinase domain *BRAF*" < "Altered structure *BRAF*" < "Alteration RAS-RAF pathway". The classification structure for the tandem duplication that generates the *KIAA1549-BRAF* fusion gene in pilocytic astrocytoma might be {"*KIAA1549-BRAF* fusion gene" < {"*KIAA1549* fusion"; "*BRAF* fusion"} < {"Altered structure *KIAA1549*"; "Altered structure *BRAF*"} < "Alteration RAS-RAF pathway"; {"Copy number gain of *TTC26*"; "Copy number gain of *PARP12*",...}}. Note that the hierarchical structure will enable these two events to be considered together at the functional level "Altered structure *BRAF*", allowing integration across mutational categories. For non-coding mutations, we will define both heuristic bins of a fixed size and also bins according to ENCODE chromatin state segmentation.

Phase 2 – A model for neutral mutations: The success of the dN/dS ratio relies on the accuracy of the synonymous mutation rate as a proxy for the expected number of neutral mutations. We will develop models for the expected background distribution of each mutational process (eg T>A mutations or tandem duplications). We will employ a rather general crossed variance components structure that will incorporate effects related to local genomic features (eg sequence context); regional genomic features (eg gene footprint, replication timing, chromatin structure, GC content etc); and mutational process-specific features (eg size distribution of tandem duplications).

Phase 3 – A statistical framework for counting excess mutations with a given functional consequence: Each mutation will be attributed to an underlying mutational process. Given the background model of neutral mutations for each process, we will be able to estimate an expected distribution of the number of tumours exhibiting a specific functional consequence. Essentially, we will peel back from the most specific functional consequence to the most general – for example, estimating first the excess numbers of V600E mutations, then the excess *BRAF* kinase domain mutations etc. In this way, it will be possible to attribute to each mutation a probability that the recurrence rate of its various functional consequences is in statistical excess. Summing these probabilities will then lead to an estimate of the number of driver mutations per patient. A second statistical framework is a probabilistic Bayesian approach. In this approach, we would construct a Monte Carlo sampler, which would sample, for each cancer genome, the set of putative driver variants (and selection intensities), from a posterior distribution specific to the cancer genome and the individual patient. From the output of this sampler we could compute marginal posterior probabilities, including the posterior probability that each of the candidate variants is a driver, and the posterior distribution of the number of driver variants. These outputs are easy to understand and interpret. We could also obtain the marginal posterior distribution of the selection intensity experienced by each putative driver variant.

This strategy will provide a comprehensive estimate of the number of driver mutations per tumour as well as the driver strength of any particular mutation, allowing us to study the distribution of these numbers across tumour types and mutational classes, and their association with clinical outcome and other metadata.

Legacy plans

The following outputs will be generated from these analyses:

1. A pragmatic ontology of functional consequences for all classes of somatic mutation
2. A statistical framework for estimating background mutational rate across mutation processes
3. Models to estimate the numbers of driver mutations

All statistical algorithms will be released with the main publication through interwoven code, output and explanatory text in the RStudio / knitr format.

CURRICULUM VITAE

NAME
Dr Peter J Campbell

POSITION TITLE
Head of Cancer Genetics & Genomics,
Wellcome Trust Sanger Institute

EDUCATION/TRAINING

FIELD OF STUDY	INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED
Mathematics and Statistics	University of Otago, New Zealand	BSc Hons (1st Class)	1994
Medicine	University of Otago, New Zealand	MB ChB (Distinction)	1995
Haematology	Royal Australasian College of Physicians	FRACP	2003
Haematology	Royal College of Pathologists of Australasia	FRCPA	2003
Haematology	University of Cambridge	PhD	2006

SELECTED PEER-REVIEWED PUBLICATIONS

Papaemmanuil E, Rapado I, …, Greaves M **and Campbell PJ**. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. **Nature Genetics** 2013 (in press).

Alexandrov LB, Nik-Zainal S, …, **Campbell PJ** and Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013, 500(7463), 415-21.

Nik-Zainal S, Van Loo P, … Futreal PA, Stratton MR, **and Campbell PJ**. The life history of 21 breast cancers. **Cell** 2012, 149(5), 994-1007.

Nik-Zainal S, Alexandrov LB, … Futreal PA, **Campbell PJ** and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. **Cell** 2012, 149(5), 979-993.

Papaemmanuil… E, Cazzola M, Futreal PA, Stratton MR, **and Campbell PJ**. Somatic *SF3B1* mutation in myelodysplasia with ring sideroblasts**. N Engl J Med** 2011, 365(15):1384-95.

Stephens PJ, Greenman CD, …, Futreal PA **and Campbell PJ**. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. **Cell** 2011, 144(1), 27-40.

Greenman CD, Pleasance ED, …Futreal PA, Stratton MR, **and Campbell PJ.** Estimation of rearrangement phylogeny for cancer genomes. **Genome Res**. 2012, 22(2), 346-61.

Campbell PJ, Yachida S,… Iacobuzio -Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. **Nature** 2010, 467(7319), 1109-13.

Pleasance ED, Stephens PJ, … Stratton MR, Futreal PA, **and Campbell PJ**. A small cell lung cancer genome with complex signatures of tobacco exposure. **Nature** 2010, 463(7278), 184-90.

Campbell PJ, Stephens PJ, , … Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics** 2008, 40(6), 722-9.

Campbell PJ, Pleasance ED, … Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. **Proc Natl Acad Sci USA** 2008, 105(35), 13081-6.

Campbell PJ, Scott LM, … Harrison CN, Green AR. Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. **Lancet** 2005, 366(9501), 1945-1953.

CURRICULUM VITAE
Full Name: Michael Rudolf STRATTON
Present Position Director, Wellcome Trust Sanger Institute
Qualifications Institution Dates
BA Physiological Sciences Brasenose College, Oxford 1979
MB BS (Medical Degree) Guys Hospital, London 1982
PhD Institute of Cancer Research, London 1989
FRCPath London 1991
Fellowships
Fellow of the Academy of Medical Sciences FMedSci 1999
Fellow of the Royal Society FRS 2008
Member of EMBO 2010
Prizes and Awards
2013 The GHA Clowes Award, American Association of Cancer Research, USA
2013 The Louis-Jeantet Prize for Medicine, the Louis-Jeantet Foundation, Switzerland
2013 The Sergio Lombroso Award in Cancer Research, Weizmann Institute, Israel
2013 The Ernst W Bertner Award, MD Anderson Cancer Centre, USA
2013 The AACR Distinguished Lecturer in Breast Cancer Research, San Antonio Breast Cancer Symposium, USA
2012 Estella Medrano Memorial Lecture Award, Society for Melanoma Research, USA
2012 Royal Physiographical Society Medal, Lund, Sweden
2011 The Massachusetts General Hospital Award in Cancer Research, Massachusetts General Hospital, USA
2010 The C. Chester Stock Award, Memorial Sloan Kettering Cancer Centre, USA
2010 The Lila Gruber Award for Cancer Research, American Academy of Dermatology, USA
2008 The BioMedicum Helsinki Medal, Finland
2007 The AstraZeneca Award, Biochemical Society UK.
2002 The Lennox K Black Award for Excellence in Medicine, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
2001 The Tom Connors Award, British Association of Cancer Research, UK
SELECTED RECENT PUBLICATIONS
Alexandrov LB, Nik-Zainal S, ..., Campbell PJ and **Stratton MR**. Signatures of mutational processes in human cancer. ***Nature*** 2013, 500, 415-21.
Alexandrov LB, Nik-Zainal S, … Campbell PJ, **Stratton MR**. Deciphering signatures of mutational processes operative in human cancer. ***Cell Rep***. (2013), 3, 246-59
Nik-Zainal S, Alexandrov LB, ... Futreal PA, Campbell PJ and **Stratton MR**. Mutational processes molding the genomes of 21 breast cancers. ***Cell*** 2012, 149, 979-993.
Stephens PJ, Tarpey PS, … Campbell PJ, P Futreal PA and **Stratton MR**. The landscape of cancer genes and mutational processes in breast cancer. ***Nature*** 2012, 486, 400-404.
Bignell GR, Greenman C, Davies H, …Campbell PJ, Futreal PA and **Stratton MR.** Signatures of mutation and selection in the cancer genome. **Nature** 2010, 463(7283), 893-8.
Stephens PJ, McBride DJ, Lin ML, ….Campbell PJ, Futreal PA and **Stratton MR**. Complex landscapes of somatic rearrangement in human breast cancer genomes. **Nature** 2009, 462(7276), 1005-10.
Greenman C, Stephens P, … Wooster R, Futreal PA, **Stratton MR**. Patterns of somatic mutation in human cancer genomes. ***Nature*** 2007**,** 446, 153-158.

Inigo Martincorena, PhD

Wellcome Trust Sanger Institute, CB10 1SA, United Kingdom
im3@sanger.ac.uk

Current position

EMBO Postdoctoral Fellow, Wellcome Trust Sanger Institute, Cambridge, UK.
Junior Research Fellow, Queens' College, University of Cambridge.

Education

2012. PhD (Evolutionary genomics). University of Cambridge and EBI–EMBL, UK.
2007. Degree in Biology. University of Navarra, Spain. 98.7% (44 A+, 7A)
2007. Degree in Biochemistry. University of Navarra, Spain. 99.6% (22 A+, 1A)

Selected publications

Evolution and functional genomics *(variable mutation rates and selection)*
** denotes corresponding author.*

Martincorena I*, Seshasayee A, Luscombe NM*. 2012. *Evidence of non-random mutation rates suggests an evolutionary risk management strategy*. Nature. (485)95-98.

Martincorena I*, Luscombe NM. 2012. *Non-random mutation: The evolution of targeted hypermutation and hypomutation*. BioEssays. 35(2):123-30.

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. *Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.* Cell. 152(3):453-66.

Cancer *(selection and cancer gene discovery)*

Wong C, Martincorena I, Rust AG, Rashid M, Alifrangis C, Alexandrov LB, Tiffen JC, Kober C, ICGC, Green AR, Massie CE, Nangalia J, Lempidaki S, Döhner H, Döhner K, Bray SJ, McDermott U, Papaemmanuil E, Campbell PJ, Adams DJ. *Inactivating CUX1 mutations promote tumorigenesis*. Nature Genetics. *In Press.*

Murchison E, Wedge D, Alexandrov LB, Fu B, Martincorena I, Ning Z, Tubio JMC, Werner EI, Allen J, Barboza De Nardi A, Donelan EM, Marino G, Fassati A, Campbell PJ, Yang F, Burt A, Weiss RA and Stratton MR. *The genome of a transmissible cancer reveals the origin and history of an ancient lineage*. Science. *In Press.*

Nangalia J, [...], Campbell PJ, Green AR. *Somatic CALR mutations in JAK2-wildtype essential thrombocythemia and myelofibrosis.* New England Journal of Medicine. *In Press.*

Yen J, […], Futreal PA. 2013. *The genetic heterogeneity and mutational burden of engineered melanomas in zebrafish models.* Genome Biology. 14(10):R113.

Selected awards and fellowships

2013-2015. Queen's College Junior Research Fellowship, University of Cambridge.
2013-2014. EMBO Long-Term Postdoctoral Fellowship.
2010-1011. *EIOI* Fellowship. Ministry of Science and Innovation. Government of Spain.
2009-2012. Cambridge University European Trust bursary.
2008, 2011. Two *Caja Madrid Foundation* Postgraduate Scholarships.
2008. Marie Curie Fellowship for Early Stages. European Commission.
2007. National Award for Excellence in Academic Performance in Biochemistry.

Kevin J. Dawson

Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus, Hinxton,
Cambridge CB10 1SA, UK

Phone: +44 (0) 1223 495395
Email: kd7@sanger.ac.uk

Education

B.Sc. Genetics, 1987–1990.

University College London, Department of Genetics and Biometry.

Classification: 2.1 (*upper second class*).

Ph.D. Theoretical Population Genetics, 1991–1993.

University of Edinburgh. (Supervisors: Prof. N. H. Barton, F. R. S., and Prof. W. G. Hill, F. R. S.)

Thesis: *The dynamics of statistical associations between many genes.*

Employment

1994. Post-doctoral Research Associate, University of California, Davis.

1995-1996. Post-doctoral Research Associate, University of Edinburgh.

1997–1999. Post-doctoral Research Associate, CNRS, Montpellier, France.

2000–December 2010. BBSRC Band 5 Research Scientist (Statistical Geneticist).

2000–2001. Long Ashton Research Station.

2001–December 2010. Rothamsted Research (formerly Rothamsted Experimental Station).

2012–present. Staff Scientist, Cancer Genome Project. Wellcome Trust Sanger Institute.

Research interests

I am a statistical geneticist with a background in population genetics and Bayesian inference. Much of my research has been on developing Bayesian methods for inference problems in population genetics, and more recently, in cancer genomics.

Selected publications

[1] Ian J. Wilson and Kevin J. Dawson. A Markov chain Monte Carlo strategy for sampling from the joint posterior distribution of pedigrees and population parameters under a Fisher-Wright model with partial selfing. *Theoretical Population Biology*, 72:436–458, 2007.

[2] Kevin J. Dawson and Khalid Belkhir. An agglomerative hierarchical approach to visualisation in Bayesian clustering problems. *Heredity*, 103:32–45, 2009.

[3] Eric Bazin, Kevin J. Dawson, and Mark A. Beaumont. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, 185:1411–1423, 2010.

[4] Niccolo Bolli, Hervé Avet-Loiseau, David C. Wedge, Peter Van Loo, Ludmil B. Alexandrov, Inigo Martincorena, Kevin J. Dawson, Francesco Iorio, Serena Nik-Zainal, Graham R Bignell, Jonathan W. Hinton, Jose Tubio, Stuart McLaren, Sarah O'Meara, Adam P. Butler, Jon W. Teague, Laura Mudie, Yu Tzu Tai, Masood A. Shammas, Adam S. Sperling, Mariateresa Fulciniti, Paul G. Richardson, Florence Magrangeas, Stephane Minvielle, Philippe Moreau, Michel Attal, Thierry Facon, P. Andrew Futreal, Kenneth C. Anderson, Peter J. Campbell, and Nikhil C. Munshi. Heterogeneity of somatic mutations, clonal architecture and genomic evolution in multiple myeloma. *Nature Communications*, 2013. In press.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Discovering New Links Between Infectious DNA Sequences and Cancer Development.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Colin Cooper, University of East Anglia, Norwich UK, ICGC UK Prostate project

Rosalind Eeles, Institute of Cancer Research, London, UK, ICGC UK Prostate project

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Dan Brewer, University of East Anglia, Norwich, UK
Zsofia Kote-Jarai, Institute of Cancer Research, London, UK

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Mario Caccamo, Director of The Genome Analysis Centre, Norwich Research Park, Norwich, UK

Lindsay Hall, University of East Anglia, Norwich, UK

Background and preliminary data

Infectious agents are involved in the development of a variety of human cancers. The involvement in papilloma virus infection in the development of cervical cancer, of hepatitis virus in the development of liver cancer, and EBV in the development of nasopharyngeal cancer are well established. Similarly the links between *Helicobacter pylori* infection and stomach cancer and *Schistosomiasis* infection and bladder cancer have been well documented. We hypothesise that other infectious agents are directly linked to cancer development, but as yet remain undefined. Indeed, for some cancer types there are already clues that new infectious agents may be involved in cancer development. Prostate cancer incidence, for example, seems at least in part to be linked to genital infections (e.g. Cancer, 2012, 118, 4437-4442). Most recently, analysis of mutations from 7,042 cancer samples across 30 different cancer types (*Nature* 2013, 500, 415-421) identified 21 distinct mutational signatures, but the causes of only 9 these signatures are currently know. These data further confirm the view that many underlying causes of cancer remain to be identified, and thus this represents a critical area for study.

The availability of Whole Genome DNA Sequence (WGS) data from many different cancer type genomes from different geographic locations via the WGS pan-cancer platform makes this an ideal time to start probing and analysing cancer DNA sequences for infectious components. We will contribute DNA sequences from the UK ICGC Prostate Cancer project and will use these as part of the 2000 genome WGS pan-cancer dataset to determine associations between infection and many different cancer types. It should be noted that the discovery of infectious DNA sequences, associated with a particular cancer type, does not fulfill either classic or molecular Koch criteria for causation. Such associations will therefore form the starting point for additional sets of studies designed to establish a causal link. There are currently no preliminary results for this project.

Timelines & resources dedicated to project

Timelines:
Jan 2014-Feb 2014 Identification and construction of database of potentially medically relevant bacteria, viruses and parasites:
Jan 2014-June 2014 Construction of analysis pipeline
July 2014-Dec 2014 Analysis of sequence data
Jan 2015- June 2015 Design and completion of validation experiments
July 2015 onwards. Design and execution of additional studies to establish causation. Paper writing.

Resources: In house bioinformatics resources at Norwich and the Institute of Cancer Research, London. Total 1 FTE

Research proposal

The starting point for this project will be unaligned DNA sequences after human genome assembly. The approach used to interrogate these sequences is similar to those used in large scale metagenome and microbiome projects (Microbial Informatics and Experimentation 2012, 2:3). Two strategies can be employed for metagenomics samples: reference-based assembly (co-assembly) and *de novo* assembly. We will initially use the former simpler approach.

Searching for potentially medically relevant bacteria, viruses and parasites: The Norwich Research Park (NRP) represents an ideal location for compiling a list of relevant DNA sequences. The main strength of the NRP for this project is the high concentration of world-leading scientists; including over 70 research groups focusing on microbiology, the gut and the microbiome and bioinformatics development (http://www.micron.ac.uk). This component of the project will be guided by the bioinformatics resources of The Genome Analysis Centre (TGAC) at Norwich, where around one third of its activities are currently devoted to analysis of bacterial and viral genomes. Critical steps are: (i) gathering genome sequences; (ii) aligning unmatched reads to each genome; (iii) determining whether there is significant evidence for the presence of individual infectious agents or families of agents through depth of coverage in unique regions of sequence. MEGAN is a useful tool that will be used in (iii) for processing and visualizing results derived from BLAST searches in a functional or taxonomic dendrogram. We will also utilise a number of large-scale databases that process, annotate and deposit metagenomic datasets: MG-RAST, IMG/M, and CAMERA are three such systems.

Statistical analysis: Its is important to establish difference between cancer samples and cancer types. A variety of statistical tools exist to achieve this: for example the Primer-E package allows for a range of multivariate statistical analyses, including the generation of multidimensional scaling (MDS) plots, and identification of the species that are different between two samples (SIMPER).

Interpretation and downstream analysis: The human microbiome consists of 10^{14} total cells (10 times the number of human cells) and an estimated 10,000 bacterial species/strains. Yet, excluding the main microbiome niches such as the gut, skin, genitourinary tract, the extent of bacterial translocation into human tissues is, in most cases, unknown. It is highly likely that numerous differences in the presence or abundance of bacterial and viral sequences between different cancer types will be established in our studies. Assessing their significance will represent a critical step involving confirmation of differences and associations in larger clinical series, and establishing epidemiological associations between infection and cancer development. To expedite these studies we already have prostate cancer resources in place. Links will be made to centers contributing WGS data to this project for investigations of other cancer types.

Links to other studies and WGS pan-cancer projects: We consider that a critical component of this study will be to overlay data on infections with other types of data from the WGS pan-cancer project: this will include survival and details of genome structure for example. Correlations between presence of infectious agent and contributing mutational signatures will be made to determine whether infections can explain unidentified signatures.

Legacy plans

Publically available database of viral, bacterial and parasite DNA sequences present in human cancers

Discovery of new links between infectious agents and cancer development.

Professor Colin Stephen Cooper, BSc PhD DSc FMedSci

Email: colin.cooper@uea.ac.uk Telephone +441603592246

Education

BSc	Class 1 Biochemistry	University of Warwick	1975
PhD	Biochemistry	University of Birmingham	1978
DSc		University of Warwick	1991

Positions Held

2011_date: Chair of Cancer Genetics, University of East Anglia
2010-date: Head of the UK component of the International Cancer Consortium (ICGC)_Prostate Cancer project
1989-2011: Head of the Molecular Carcinogenesis Section at the Institute.
1997-2002: Head of Laboratories (Sutton Site, Institute of Cancer Research).

Major Grant funding

2013-2016 Bob Champion Cancer Trust award of £186k to support senior bioinformatics post.
2013-2016. Andy Ripley studentship of £90,000 to support bioinformatics PhD post
2012-2014 Prostate Cancer Foundation (USA) grant of $1M to carry out chemoprevention studies in prostate cancer.
2012-2014 Movember Grant of $1M(AUS) to carry out urine biomarker studies in prostate cancer
2011-2013 Cancer Research UK grant of up to £4M to carry out the International Cancer Genome Consortium (ICGC)_Prostate Cancer project.
2001-2011. National Cancer Research Institute £5.41M grant (funded by Department of Health, MRC, Cancer Research UK) to run the South of England Prostate Cancer Collaborative.

References (selected recent)

1. Cooper CS, Eeles R, Wedge D, et al. The Life History of Multifocal Prostate Cancer: Multiple Independent Clonal Expansions in Neoplastic and Normal Prostate Tissue. Nature Medicine. [under review]
2. Olmos D, Brewer D, Clark J, ..Cooper CS et al. Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. The Lancet Oncology. 2012, 13: 1114-24
3. Santarius T, Shipley J, Brewer D, Stratton MR, and Cooper CS. A census of amplified and overexpressed human cancer genes. Nature reviews. Cancer, 2010, 10: 59-64
4. Reid AH, Attard G, ……Cooper CS. Novel, gross chromosomal alterations involving PTEN cooperate with allelic loss in prostate cancer. Mod Pathol. 2012, 25: 902-10
5. Kote-Jarai Z, Olama AA, … Cooper, CS…..Easton DF, Eeles RA; Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study Nat Genet. 2011, 43: 785-91
6. McCarthy F, Dennis N, Flohr P, Jhavar S, Parker C, Cooper CS, High-density tissue microarrays from prostate needle biopsies.J Clin Pathol. 2011, 64: 88-9
7. Attard G, de Bono JS, Clark J, Cooper CS. Studies of TMPRSS2-ERG gene fusions in diagnostic trans-rectal prostate biopsies. Clin Cancer Res. 2010, 16: 1340
8. Reid AH, Attard G, …., **Cooper CS**; Transatlantic **Prostate** Group. Molecular characterisation of ERG, ETV1 and PTEN gene loci identifies patients at low and high risk of death from prostate cancer. Br J Cancer. 2010, 102: 678-84.
9. Attard G, Cooper CS, de Bono JS. Steroid hormone receptors in prostate cancer: a hard habit to break? Cancer Cell. 2009 Dec 8;16(6):458-62.

Professor Rosalind Anne EELES

Posts Held:

Jan 2010 -	Professor of Oncogenetics & Team Leader, Oncogenetics team, Institute of Cancer Research
Dec1994 -	Honorary Consultant, Cancer Genetics & Clinical Oncology, Royal Marsden NHS Trust
Dec 1994 - Jan 2010	Team Leader in Translational Cancer Genetics, Institute of Cancer Research
Apr 2004-	Head of Clinical Cancer Genetics Unit, Royal Marsden NHS Foundation Trust

Education And Qualifications

Jun 2012	FMedSci
Jan 2010	Professor of Oncogenetics
May 2000	FRCP Elected Fellow of the Royal College of Physicians London, UK
Mar 2000	PhD Germline and Somatic Mutations in the TP53 Gene in Breast and Other Cancers (University of London)

Personal statement

Prof Eeles is an internationally recognised expert in genetic predisposition to prostate cancer for which she was elected as a Fellow of the Academy of Medical Sciences in June 2012. Her research programme links bench to bedside research and involves identification of genetic variants which predispose to prostate cancer and their clinical application in targeted screening and prostate cancer care.

Current Grants held

PG13 – 001 (PI:Eeles) PCUK/Movember £200,703 Dec 2013-Nov 2016
Identification of DNA Repair gene mutations as a predisposition to early onset and aggressive prostate cancer.

GAP1 Funding Award (PI: Eeles) Movember £236,282 Nov 2012-Oct 2014
Finding coding variants which predispose to clinically relevant prostate cancer.

C5047/A15007 (PI: Eeles) CR-UK £1,000,000 Oct 2012-Sep 2014
Translational research in individuals with genetic predisposition to cancer.

C5047/A13232 (PI: Eeles) CR-UK £133,049 Sep 2011– Aug 2014
The IMPACT Study – This Study is the identification of men with a genetic predisposition to prostate cancer by targeted screening in BRCA1/2 mutation carriers and controls.

C5047/A14835 (Joint PI: Eeles/Cooper) CR-UK £621,753 Sep 2011- Mar 2014
Full Project: (ICGC) International Cancer Genome Consortium: The Prostate Cancer Initiative: a UK-North American-French-Asian Partnership.

(PI:Eeles) The Ronald & Rita Mcaulay Foundation. US$250,710 Apr 2011 – Mar 2014
THE IMPACT STUDY: This study is the identification of men with a genetic predisposition to prostate cancer by targeted screening in BRCA1/2 mutation carriers and controls.

1U19CA148537-01 (PI: Henderson) NIH US$1,670,706 Jul 2010 – Jun 2014
ELLIPSE: Elucidating Loci Involved in Prostate Cancer Susceptibility

Recent Relevant Publications (Selected from a peer-reviewed list of 312 publications)

Eeles RA et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013;45(4):385-91.

Al Olama AA, et al & Eeles RA. A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. Hum Mol Genet. 2013 ; 22(2):408-15.

Goh CL, et al & Eeles RA. Clinical implications of family history of prostate cancer and genetic risk single nucleotide polymorphism (SNP) profiles in an active surveillance cohort. BJU Int. 2013;112(5):666-73

Kote-Jarai Z, et al & Eeles RA. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression. Hum Mol Genet. 2013; 22(12):2520-8.

Castro E, et al & Eeles R. Germline BRCA Mutations Are Associated With Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer. J Clin Oncol. 2013; 31(14):1748-57.

Leongamornlert D, et al & Eeles R, Kote-Jarai Z. Germline BRCA1 mutations increase prostate cancer risk. BJC 2012;106(10):1697-701.

Goh CL, & Eeles RA. Genetic Variants Associated with predisposition to prostate cancer and potential clinical implications. J Intern Med. 2012; 271(4): 353-65.

Dr Daniel Brewer, MSci MRes PhD

Leadership roles

2012 instated as principle (starred) co-author on ICGC Prostate UK Project publications.
2011 onwards a bioinformatics lead and deputy Chair of ICGC Bioinformatics Committee on ICGC Prostate UK Project.
2009 onwards bioinformatics lead on Prostate Cancer Map project.
2008-2011 head bioinformatician in the section of molecular carcinogenesis.
2006-2013 lead scientist for bioinformatics analysis in Cooper lab.

Employment

Sept 2013 onwards visiting worker (TGAC)
Sept 2013 onwards Senior Bioinformatics Officer (University of East Anglia)
NGS analyses and development of appropriate filtering strategies.
Data analysis of large data sets produced by microarray experiments investigating prostate cancer.
Development of diagnostic, prognostic and predictive tests.
Application of novel or new algorithms to large datasets.
Design and development of database web applications.
Jan 2006 to Aug 2013 Senior Bioinformatics Officer (Institute of Cancer Research)

Education and Qualifications

Oct 2002 to Jan 2006 PhD in Computational Biology (ICH & CoMPLEX, University College London)
Oct 2001 to Sept 2002 MRes in Biological Complexity (Distinction) (CoMPLEX, University College London)
Oct 1996 to June 2000 Physics - MSci. (First Class Honours) (Imperial College, London)

Selected Publications

I have led bioinformatics components in each of these publications:
Cooper CS, Eeles R, Wedge D, et al. The Life History of Multifocal Prostate Cancer: Multiple Independent Clonal Expansions in Neoplastic and Normal Prostate Tissue. Nature Medicine. [Submitted] (Joint senior author)
Olmos D, Brewer D, Clark J, et al. Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. The lancet oncology. 2012;2045(12):1–11. (Joint first author)
Santarius T, Shipley J, Brewer D, Stratton MR, and Cooper CS. A census of amplified and overexpressed human cancer genes. Nature reviews. Cancer, vol. 10, 2010, pp. 59-64.
Barenco M, Brewer D, Papouli E, et al. Dissection of a complex transcriptional response using genome-wide transcriptional modelling. Molecular systems biology. 2009;5:327. (Joint first author)

Dr Zsofia Kote-Jarai

Senior Staff Scientist, The Institute of Cancer Research

Posts Held:
Jan 2010 - Senior Staff Scientist, Oncogenetics Team, The Institute of Cancer Research, Sutton, UK
2005-2009- Staff Scientist, Oncogenetics Team, The Institute of Cancer Research, Sutton, UK
1998 - 2005- Postdoctoral Scientist, Oncogenetics Team, The Institute of Cancer Research, Sutton, UK
1995-1997 - Staff Scientist, Dep Molecular Biology, National Institute of Oncology, Budapest, Hungary
1993-1994 - Postdoctoral Research Fellow, Friedrich Miescher Institute, Basel, Switzerland

Education and Qualifications

1985 MSc Biology Eotvos Lorand University, Faculty of Natural Sciences, Budapest, Hungary
1987 PhD Botany and Ecology Eotvos Lorand University, Faculty of Natural Sciences, Budapest, Hungary
1995 PhD Molecular Genetics Eotvos Lorand University, Faculty of Natural Sciences, Budapest, Hungary

1989-91 Postdoctolal Fellow Ohio State University, Columbus Ohio, USA

Personal statement
I have a very broad range of experience in human molecular genetics as I have worked in four international research centres and my strategic planning and managerial skills have been essential to set up and execute the numerous projects I have been involved in. I have been working on the genetics of prostate cancer for the last ten years, leading the research in the experimental research group of Prof Rosalind Eeles' Oncogenetics team at The Institute of Cancer Research. We have published (5 papers in Nature Genetics) on genome-wide association studies in prostate cancer and currently I am managing research projects to follow up our results from this in addition to introducing new approaches to identify rare genetic variants contributing to prostate cancer risk. In my current position I have been involved and lead several national and international genetic studies in prostate cancer that identified the largest number of common genetic variants which affect risk of the disease. The focus of our research also extends to clinical application of our findings and to targeted screening of men at high risk of prostate cancer.

Recent Relevant Publications
Eeles RA et al, Kote-Jarai, Z and Douglas Easton. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013 45(4):385-91
Kote-Jarai Z et al. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression, Hum Mol Genet. 2013 22(12):2520-8.
3.Castro E,... Kote-Jarai Z, Eeles R. Germline BRCA Mutations Are Associated With Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer. J Clin Oncol. 2013 May 10;31(14):1748-57
Leongamornlert D, et al & Kote-Jarai Z. Germline BRCA1 mutations increase prostate cancer risk. Br J Cancer 2012;106(10):1697-701.
Kote-Jarai Z et al. BRCA2 is a moderate penetrance gene contributing to young onset prostate cancer: implications for genetic testing in prostate cancer patients. Br J Cancer. 2011; 105(8):1230-4.
Kote-Jarai Z, et al. Seven novel prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nature Genetics 2011;43(8):785-91.
Al Olama AA, Kote-Jarai Z et al Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet. 2009; 41(10): 1058-60
Eeles RA, Kote-JArai Z et al. Identification of seven novel prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009;41(10):1116-21
Eeles R, Kote-Jarai Z et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genetics 2008; 40(3):316-321.

Dr Mario Jose Caccamo

Director: The Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH. UK
Email: mario.caccamo@tgac.ac.uk

Employment History
2013 - present Director – The Genome Analysis Centre
2013 - present Acting Director – The Genome Analysis Centre
2009 – present Head of Bioinformatics – The Genome Analysis Centre.
2009 - present External Faculty Member – John Innes Centre.
2009 - present Honorary Senior Lecturer – University of East Anglia.
2007 - 2009 Scientific Software Engineer, European Genome-phenome Archive, European Bioinformatics Institute (EBI).
2005 - 2007 Principal Computer Programmer, Zebrafish/Pig Genome Assembly
The Wellcome Trust Sanger Institute.
2005 - 2005 Acting Project Leader, Zebrafish Genome Analysis Group
The Wellcome Trust Sanger Institute.
2003 - 2005 Senior Computer Biologist,
The Wellcome Trust Sanger Institute.

Education

2003	PhD in Computer Science	BRICS University of Aarhus, Denmark
1998	MSc in Computer Science	UNICAMP, Campinas, Brazil
1996	BSc in Computer Science	Universidad Nacional del Sur, Bahía Blanca, Argentina

Publications (selected recent publications):
"The zebrafish reference genome sequence and its relationship to the human genome." Howe K, Clark M. et al. (Consortium manuscript) Nature (2013)
"Metagenomic study of the viruses of African straw-coloured fruit bats: Detection of a chiropteran poxvirus and isolation of a novel adenovirus" Baker K. et al. Virology dx.doi.org/10.1016/j.virol.2013.03.014
"A physical, genetic and functional sequence assembly of the barley genome." Mayer KF et al. Nature 2012 doi: 10.1038/nature11543.
"Evolution of an Eurasian avian-like influenza virus in naïve and vaccinated pigs." Murcia PR, et al. PLoS Pathog. 2012;8(5):e1002730. May 2012
"De novo assembly and genotyping of variants using colored de Bruijn graphs". Iqbal Z, Caccamo M, et al. Nature Genetics doi 10.1038/ng.1028. January 2012.
"Advances in bacterial transcriptome and transposon insertion-site profiling using second-generation sequencing." Febrer M, McLay K, Caccamo M, Twomey KB, Ryan RP. Trends Biotechnology. November 2011;29(11):586-94

Research Funding (selected recently awarded grants):
Nornex: An open consortium for molecular understanding of ash dieback disease. BBSRC & DEFRA (Co-I)
Triticeae Genomics for Sustainable Agriculture (PI). (BBSRC LoLa BB/J003743/1)
Trans-national Infrastructure for Plant Genomic Science (Work Package Leader) (EU FP7 transPLANT)
Effect of Chromatin modification on meiosis: wheat, a model for polyploid crops BB/J009334/1 (Co-I)
Draft sequence of the barley genome (Co-I). (BBSRC BB/I00663X/1A)
RevGenUK, the next generation: establishing a TILLING boutique for a UK-based reverse genetics community resource (Co-I). (BBSRC BB/I026030/1)
Scientific Activities
Associated Editor – BMC Bioinformatics
Co-Chair of Bioinformatics Expert Working Group – Wheat Initiative
BBSRC Institute Strategic Programme Grant (TGAC): Bioinformatics for data-driven Science (2012-2017)–PI

Dr Lindsay J Hall

Address: Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
E-mail:Lindsay.Hall@uea.ac.uk; Tel: +44 (0)1603255167

Education and Employment

2011-current: Research Leader, Institute of Food Research, Norwich Research Park, Norwich.
2011-current Lecturer and Principle Investigator in Gastrointestinal Sciences, Norwich Medical School, University East Anglia, Norwich Research Park, Norwich.
2007 – 2011 Post Doctoral Research Fellow, Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland.
2003 – 2007: Ph.D, University of Cambridge. The Wellcome Trust Sanger Institute, Hinxton, Cambridge.
1999 – 2003: BSc (Hons) Upper Second Class, Microbiology, University of Glasgow.

Publications (LAST 5 YEARS)

Jonathan M. Williams, Carrie A. Duckworth, Alastair J. M. Watson, Mark R. Frey, Jennifer C. Miguel, Michael D. Burkitt, Robert Sutton, Kevin R. Hughes, Lindsay J. Hall, Jorge H Caamaño, Barry J. Campbell, D. M. Pritchard. A mouse model of pathological small intestinal epithelial cell apoptosis and shedding induced by systemic administration of lipopolysaccharide. Dis. Model. Mech. (2013) doi:10.1242/dmm.013284

Watson AJM and Hall LJ. Regulation of host gene expression by gut microbiota. Gastroenterology. (2013) doi:pii: S0016-5085(13)00239-4.

Hall LJ, Murphy CT, Quinlan A, Hurley G, Shanahan F, Nally K, Melgar S. Natural killer cells protect mice from DSS-induced colitis by regulating neutrophil function via the NKG2A receptor. Mucosal Immunol. (2013) 6(5):1016-26

Hall LJ, Murphy CT, Quinlan A, Hurley G, Shanahan F, Nally K, Melgar S. Natural Killer cells protect against mucosal and systemic infection with the enteric pathogen Citrobacter rodentium. Infect. Immun. (2013) 81(2):460-9

Watson AJM, Hall LJ, Hughes KR. Cell shedding – old questions answered. Gastroenterology. (2012) 143(5):1389-91.

Hall LJ/Fanning S, van Sinderen D. Bifidobacterium breve UCC2003 surface exopolysaccharide is a beneficial trait mediating commensal-host interaction through immune modulation and pathogen protection. Gut Microbes. (2012) 1;3(5).

Murphy CT, Hall LJ, Hurley G, Quinlan A, Shanahan F, Nally K, Melgar S. FTY720 Impairs Mucosal Immunity and Clearance of the Enteric Pathogen Citrobacter rodentium. Infect Immun. (2012) 80(8):2712-23

Hall LJ and Watson AJM. Role of Autophagy in NOD2-Induced Inflammation in Crohn's Disease. Gastroenterology. (2012). 142(4):1032-4.

Hall LJ/Fanning S, Cronin M, Zomer A, MacSharry J, Goulding D, O'Connell-Motherway M, Shanahan F, Nally K, Dougan G, van Sinderen D. Bifidobacterial surface-exopolysaccharide facilitates commensal-host interaction through immune modulation & pathogen protection. PNAS. (2012). 7;109(6):2108-13.

Hall LJ/ Jansen AM, , Clare S, Goulding D, Holt KE, Grant AJ, Mastroeni P, Dougan G, Kingsley RA. A Salmonella Typhimurium-Typhi genomic chimera: A murine-typhoid model to study Vi polysaccharide capsule function. PLoS Pathogens. (2011) 7(7):e1002131.

Hall LJ, Clare S, Dougan G. Probing local innate immune responses after mucosal immunisation. J Immune Based Ther Vaccines. (2010) 8:5.

Murphy CT, Moloney G, Hall LJ, Quinlan A, Faivre E, Casey P, Shanahan F, Melgar S, Nally K. Use of bioluminescence imaging to track neutrophil migration and its inhibition in experimental colitis. Clin Exp Immunol. (2010) 162(1): 188-96.

Hall LJ, Faivre E, Quinlan A, Shanahan F, Nally K, Melgar S. Induction and activation of adaptive immune populations during acute and chronic phases of a murine model of experimental colitis. Dig Dis Sci. (2010) 56(1): 79-89.

Hall LJ, Clare S, Dougan G. NK cells influence both innate and adaptive immune responses after mucosal immunization with antigen and mucosal adjuvant. J Immunol. (2010) 184(8):4327-37.

Hall LJ, Clare S, Pickard D, Clark SO, Kelly DL, El Ghany MA, Hale C, Dietrich J, Andersen P, Marsh PD, Dougan G. Characterisation of a live Salmonella vaccine stably expressing the Mycobacterium tuberculosis Ag85B-ESAT6 fusion protein. Vaccine (2009) 27(49):6894-904.

GRANTS

Wellcome Trust New Investigator Award (2013-2018): The role of the early life microbiota in colonisation resistance development. £1.3 million. Sole Investigator

UEA PhD Studentship (2013-2016): The role of adult gut microbiota in the regulation of intestinal cell shedding and epithelial integrity. UK/EU PhD registration fees and £12,000: Principal supervisor.

NMS Translational Medicine Grant (2013-2014): Analysis of the gut microbiota composition in neonates after probiotic treatment. £50,000. Principle Investigator

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The landscape of telomere lengths, interstitial telomeric repeats, and telomerase activity in multiple cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Ros Eeles, Institute of Cancer Research, ICGC UK Prostate project
Rebecca Fitzgerald, Hutchison/MRC Research Centre, ICGC UK Oesophageal project

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Andy Lynch, Cancer Research UK Cambridge Institute Zsofia Kote-Jarai, Institute of Cancer Research	

Background and preliminary data

Numerous papers have commented on a) the shortening of telomeres with age, b) the genomic instabilities that can arise from telomere shortening, and c) the alternate lengthening of telomeres, ALT, that can occur in cancers (for example, Cell 152 (3) 390–393). Elsewhere, the effect of parental age on telomeres has been documented (see for example PNAS 109 (26) 10251-6). Naturally then, there have been efforts to infer telomere length from whole genome sequencing data (BMC Genomics 11:244, Genome Biology 13:R113, https://github.com/zd1/telseq).

Telomerase is primarily responsible for maintaining telomere lengths but has been shown to play other roles in cancer (reviewed in Cell 144(5) 646-74). The key protein-coding gene contributing to the telomerase complex is *TERT*, and germline variants have been identified that are associated with cancer and affect telomerase expression (Hum Mol Genet 22(12) 2520-2528), and which are associated with telomere length and also with cancer (Hum Mol Genet 10.1093/hmg/ddt355). Further, somatic gains of the *TERT* gene (Cell 150(2) 251-263) and somatic mutations in *TERT*'s promoter regions (Nat Communications 4:2185) have been associated with cancers.

Many other genes and non-coding RNAs have been identified as having roles in the regulation and maintenance of telomeres (Nat Rev Gen 6(8) 611-622) and many of the telomere-independent functions of telomerase have been elucidated (Cell 144(5) 646-74), including a role in Wnt signaling (Nature 460 66-72).

In addition to this, other telomere-like sequences exist in the genome. These interstitial telomeric repeats are sources of chromosomal instability, and can exhibit MSI-style behaviour in cancer (Genomics 68(2) 111-117).

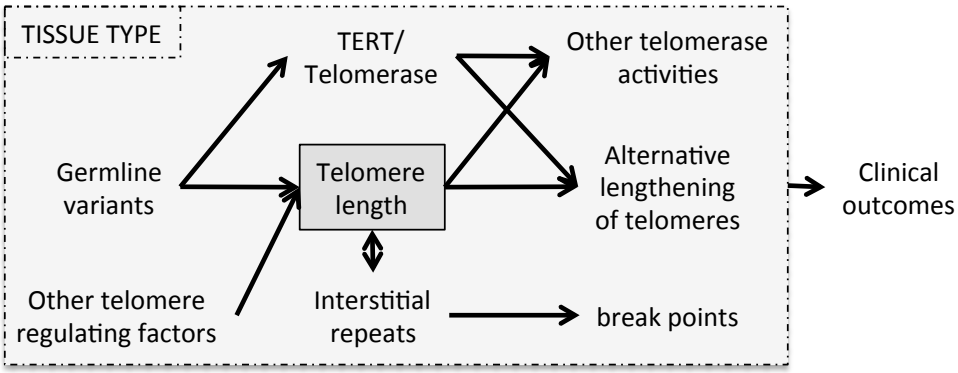
Timelines & resources dedicated to project

- One full-time computational biology PhD student (Henry Farmery, supervised by Andy Lynch and David Neal) 2013 to 2016/7
- 10% of Andy Lynch's time, 10% of Mike Smith's time, 5% of Zsofia Kote-Jarai's time.
- Input from the prostate steering group (https://prostatedatabase.org.uk/ICGC/steeringGroup.php)

Dec 2013-Aug 2014	Development of tool for estimating telomere lengths
Jan 2014-Mar 2014	Identification of gene lists and germline variants to consider
Jul 2014-Oct 2014	Retrieval of germline variants of interest.
Aug 2014-Nov 2014	Estimation of telomere lengths and interstitial telomeric repeats
Nov 2014	Access to ICGC pan-cancer somatic mutation calls and expression data
Sep 2014-Feb 2015	Analyses + Manuscript preparation

Research proposal

- At the core of our proposal is the development of a new tool to estimate telomere lengths from WGS data that builds on approaches that have gone before by taking into account ploidy (more telomere doesn't necessarily mean longer telomeres, it may just mean more telomeres), cellularity (important if there is a change in telomere length in the cancer), and interstitial telomeric repeats, ITRs, (which combined may be expected to contribute of the order of a telomere's worth of hexamer repeats even if they remain stable). This tool will be used to estimate telomere lengths for the samples with whole-genome sequencing data. It will be validated using some public resources and in-house experimentation.
- A simple comparison of telomere lengths, and changes in telomere lengths across a range of tumour-types will be the first analysis.
- In estimating the telomere lengths we shall have to identify and remove 'telomeric' reads corresponding to ITRs. In doing so we will generate a genome-wide map of ITRs. In considering the tumour-normal pairs we will generate a catalogue of the stabilities of ITRs in different cancer types.
- We have previously identified in Prostate cancer, germline variants that are associated with the expression levels of *TERT*, and variants that are associated with telomere length. We will investigate in which cancer-types these associations hold to gain greater understanding of the functional behaviour.
- We will model telomere length in terms of *TERT* expression, patient age, patient's parents' ages (if available), and known germline variants. We will look for somatic mutations and additional germline variants (focusing on genes known to be involved in telomere lengthening or regulation) that contribute to or otherwise modify the model.
- In doing so we will identify samples for which the telomere length is longer than can be explained by the known factors. These are candidates for ALT taking place. Differences in numbers between tumour types will be examined. Differential expression analyses between candidates and non-candidates will be performed. Genes that are consistent across tumour-types will be investigated.
- Where telomere length is not as great as would be indicated by telomerase activity, we will check for evidence of activity for known pathways involved in other telomerase activities and chart these across cancer types.
- Depending on what information is available on clinical outcomes, we will investigate whether any of these factors are associated with outcome.
- Finally, ITRs have been associated with chromosomal instability, and we will investigate whether the break points called in these samples do associate with ITRs



- All analyses will be across the full set of tumour types, using the observed differences in behaviour between tumour types to aid interpretation of the results.

Legacy plans

1. We shall produce a tool for estimating telomere lengths from WGS data, specifically written with the considerations of cancer samples in mind.
2. We will provide a genome-wide catalogue of interstitial telomeric repeats, their degrees of instability, and their associations with different cancer types.

Professor Rosalind Anne EELES

EDUCATION AND QUALIFICATIONS

Jun 2012 **FMedSci**
Jan 2010 Professor of Oncogenetics
May 2000 **FRCP** Elected Fellow of the Royal College of Physicians London, UK
Mar 2000 **PhD** Germline and Somatic Mutations in the *TP53* Gene in Breast and Other Cancers (University of London)

POSTS HELD:

Jan 2010 - Professor of Oncogenetics & Team Leader, Oncogenetics team, Institute of Cancer Research
Dec1994 - Honorary Consultant, Cancer Genetics & Clinical Oncology, Royal Marsden NHS Trust
Dec 1994 - Team Leader in Translational Cancer Genetics, Institute of Cancer Research
Jan 2010
Apr 2004- Head of Clinical Cancer Genetics Unit, Royal Marsden NHS Foundation Trust

Personal statement

Prof Eeles is an internationally recognised expert in genetic predisposition to prostate cancer for which she was elected as a Fellow of the Academy of Medical Sciences in June 2012. Her research programme links bench to bedside research and involves identification of genetic variants which predispose to prostate cancer and their clinical application in targeted screening and prostate cancer care.

Current Grants held

1. *PG13 – 001 (PI:Eeles)* ***PCUK/Movember*** *£200,703 Dec 2013-Nov 2016*
   Identification of DNA Repair gene mutations as a predisposition to early onset and aggressive prostate cancer.
2. *GAP1 Funding Award (PI: Eeles)* ***Movember*** *£236,282 Nov 2012-Oct 2014*
   Finding coding variants which predispose to clinically relevant prostate cancer.
3. *C5047/A15007 (PI: Eeles)* ***CR-UK*** *£1,000,000 Oct 2012-Sep 2014*
   Translational research in individuals with genetic predisposition to cancer.
4. *C5047/A13232 (PI: Eeles)* ***CR-UK*** *£133,049 Sep 2011– Aug 2014*
   The IMPACT Study – This Study is the identification of men with a genetic predisposition to prostate cancer by targeted screening in BRCA1/2 mutation carriers and controls.
5. *C5047/A14835 (Joint PI: Eeles/Cooper)* ***CR-UK*** *£621,753 Sep 2011- Mar 2014*
   Full Project: (ICGC) International Cancer Genome Consortium: The Prostate Cancer Initiative: a UK-North American-French-Asian Partnership.
6. *(PI:Eeles)* ***The Ronald & Rita Mcaulay Foundation****. US$250,710 Apr 2011 – Mar 2014*
   THE IMPACT STUDY: This study is the identification of men with a genetic predisposition to prostate cancer by targeted screening in BRCA1/2 mutation carriers and controls.
7. *1U19CA148537-01 (PI: Henderson)* ***NIH*** *US$1,670,706 Jul 2010 – Jun 2014*
   ELLIPSE: Elucidating Loci Involved in Prostate Cancer Susceptibility

Some Recent Relevant Publications (Selected from a peer-reviewed list of 312 publications)

1. **Eeles RA** et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2012; accepted.
2. Leongamornlert D, et al & **Eeles R**, Kote-Jarai Z. Germline BRCA1 mutations increase prostate cancer risk. BJC 2012;106(10):1697-701.

3. Goh CL, & **Eeles RA**. Genetic Variants Associated with predisposition to prostate cancer and potential clinical implications. J Intern Med. 2012; 271(4): 353-65.
4. Kote-Jarai Z, & **Eeles R**. *BRCA2* is a moderate penetrance gene contributing to young onset prostate cancer: implications for genetic testing in prostate cancer patients. Br J Cancer. 2011; 105(8):1230-4.
5. Kote-Jarai Z, et al & **Eeles RA**. Seven novel prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nature Genetics 2011;43(8):785-91.
6. Al Olama AA et al & **Eeles R***, Easton DF* Multiple loci on 8q24 associated with prostate cancer susceptibility. * Equal contribution. Nat Genet. 2009; 41(10): 1058-60
7. **Eeles RA**, et al & Easton DF.. Identification of seven novel prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009;41(10):1116-21
8. **Eeles R**, et al & Easton DF. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genetics 2008; 40(3):316-321.

CURRICULUM VITAE

Rebecca Clare Fitzgerald

Home address: The Wix, 147 High Street, Harston, Cambridge CB22 7QD. Tel: +44 (0)1223-870420
Business address: MRC Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre, Box 197, Cambridge Biomedical Campus, Cambridge CB2 0XZ
Tel: +44 (0)1223 763287, E-mail: rcf29@mrc-cu.cam.ac.uk
Web page: www.mrc-cu.cam.ac.uk/our_research/Rebecca_Fitzgerald/index.html

Date of birth: 30 September 1968 **GMC Registration number**: 3616372
Nationality: British **Date of full registration**: February 1993
Family Status: Married with four children **Date of entry to specialist register**: October 2002

Education and Degrees

2013 F.Med.Sci., London
2006 F.R.C.P., London
1997 M.D., University of Cambridge
1995 M.R.C.P., London
1993 M.A., University of Cambridge
1991 M.B. B.Chir., University of Cambridge
1989 B.A. (Hons), Medical Sciences Tripos, University of Cambridge
1986-1991 Girton College, Cambridge (Choral Scholar)
1976–1986 St. Brandon's School, Clevedon (Scholar, Chemistry, Physics, Mathematics, Biology)

Awards and Scholarships

2012 NIHR Research Professorship (five year term, start date to be agreed)
2011 NHS Innovation Challenge Prize for work on the cytosponge
2008 Lister Institute Research Prize
2007 Sir Francis Avery Jones Medal awarded by British Society of Gastroenterology for contribution to research in the field.
2007 Royal College of Physicians Goulstonian Lecturer
2005 National Cancer Research Institute – Research Prize for translational research
2005 UK selected candidate for 'European Rising Star in Gastroenterology' prize lecture at United European Gastroenterology Week in Copenhagen.
2004 Westminster Medal and prize for excellence in Science and Technology
2002 National Clinician Scientist Award

Present appointments

Dec 02-present **Programme Leader** (Tenure status granted June 2007)
Cancer Cell Unit, Hutchison/MRC Research Centre, Hills Road, Cambridge CB2 0XZ

Jul 01–present **Honorary Consultant Gastroenterology and Oncology**
Addenbrooke's Hospital, Cambridge
SpR appointment initially - Honorary Consultant April 2003

Oct 02–present **Director of Studies in Medicine & College Lecturer in Medical Sciences**
Trinity College, Cambridge

Oct 10-present **Adjunct Faculty of Cambridge Research Institute**

Memberships of Learned Societies and Other Organisations

2012 Member of Screening, Prevention and Early Diagnosis Advisory Group (SPED)
2012 Member of Biomarker Expert Review Panel of the Science Committee, Cancer Research UK
2012 Member of NICE Interventional Procedures Committee
2012 Member of NCRN Oesophago-Gastric Subgroup

2011 Member of MD Committee, University of Cambridge School of Clinical Medicine
2011 Scientific Board for International Cancer Genome Consortium
2010 Trustee and Founding member Heartburn Cancer Awareness Support (Registered Charity Number 1136413)
2008 Fellow of the American Gastroenterological Association (AGAF)
2007 Fellow of Institute of Learning and Teaching (member 2002, fellow 2007)
2006 Member of British Society of Gastroenterology Research and Oesophageal Committees
2002 Member of the Association of Physicians

Dr Andy **Lynch**

(www.andrewlynch.co.uk)

EMPLOYMENT HISTORY

Present

2008- Senior Research Associate, CRUK Cambridge Institute, University of Cambridge
2013- College Lecturer in Mathematics, Downing College, Cambridge

Past

2006-2008 Research Associate, Department of Oncology, University of Cambridge
2002-2006 Research Associate, Centre for Applied Medical Statistics, University of Cambridge
2001-2002 Research Associate, Centre for Process Analytics and Control Technology, Newcastle

QUALIFICATIONS AND AWARDS

2010 Chartered Scientist, The Royal Statistical Society (on behalf of the Science Council)
2006 Chartered Statistician, The Royal Statistical Society
2002 PhD, Probability and Statistics, University of Sheffield
1998 Certificate of Advanced Study in Mathematics, University of Cambridge (MMath 2011)
1997 BA, Mathematics, Downing College, University of Cambridge (MA 2001)

ACTIVE AREAS OF RESEARCH

High-Throughput Technologies – Investigating biases and fundamental properties of new technologies to develop improved data summary methods and better analysis tools *(see e.g. BMC Genomics, 2009, 10:588; Statistical Methods in Medical Research, 2009, 18:437-452; Bioinformatics, 2011, 27:713-714)*

Inference Problems in Genomics – Developing methods to estimate properties that cannot directly be measured such as FDR, library complexity, complementary mutations that mask each other etc. *(see e.g. PNAS, 2008, 105:10067-10072)*

Design of Experiments – Decisions such as platform choice, and practical designs to combat sample mix-ups or to improve resource allocation in large-scale genomics studies *(see e.g. BMC Genomics, 2010, 11:540; PLoS ONE, 2012, 7:e41815)*

Cancer Genomics Collaborations – A named 'key-collaborator' on the CRUK-funded Oesophageal ICGC grant. A collaborator on the CRUK-funded Prostate ICGC project. Previously a researcher on the METABRIC study of 2000 breast tumours. *(see e.g. Nature, 2012, 486:346–352; EMBO Journal, 2011, 30:2719-2733)*

SELECTED OTHER PROFESSIONAL ACTIVITY

2011- BMC Cancer Editorial Board member
2005-2010 Statistical Methods in Medical Research, Editorial Advisory Board member
2013 European Bioconductor Developers' Workshop organizing committee member
2011-2012 Royal Statistical Society, Bioinformatics Study Group member
2011 Isaac Newton Institute, Design of Experiments Programme co-organizer
2011 3rd Cambridge Statistics Initiative special one-day meeting co-organizer

TEACHING EXPERIENCE

See www.andrewlynch.co.uk/home/teaching for details

RESEARCH

Co-author of approximately forty peer-reviewed articles, multiple more in edited volumes, and four software packages (details at http://www.andrewlynch.co.uk/home/publications).

BIOGRAPHICAL SKETCH

Dr Zsofia Kote-Jarai – Senior Staff Scientist, The Institute of Cancer Research

EDUCATION AND QUALIFICATIONS

1985 MSc Biology Eotvos Lorand University, Faculty of Natural Sciences, Budapest, Hungary
1987 PhD Botany and Ecology Eotvos Lorand University, Faculty of Natural Sciences, Budapest, Hungary
1995 PhD Molecular Genetics Eotvos Lorand University, Faculty of Natural Sciences, Budapest, Hungary

1989-91 Postdoctolal Fellow Ohio State University, Columbus Ohio, USA

POSTS HELD:

Jan 2010 - Senior Staff Scientist, Oncogenetics Team, The Institute of Cancer Research, Sutton, UK
2005-2009- Staff Scientist, Oncogenetics Team, The Institute of Cancer Research, Sutton, UK
1998 - 2005- Postdoctoral Scientist, Oncogenetics Team, The Institute of Cancer Research, Sutton, UK
1995-1997 - Staff Scientist, Dep Molecular Biology, National Institute of Oncology, Budapest, Hungary
1993-1994 - Postdoctoral Research Fellow, Friedrich Miescher Institute, Basel, Switzerland

Personal statement

I have a very broad range of experience in human molecular genetics as I have worked in four international research centres and my strategic planning and managerial skills have been essential to set up and execute the numerous projects I have been involved in. I have been working on the genetics of prostate cancer for the last ten years, leading the research in the experimental research group of Prof Rosalind Eeles' Oncogenetics team at The Institute of Cancer Research. We have published (5 papers in Nature Genetics) on genome-wide association studies in prostate cancer and currently I am managing research projects to follow up our results from this in addition to introducing new approaches to identify rare genetic variants contributing to prostate cancer risk. In my current position I have been involved and lead several national and international genetic studies in prostate cancer that identified the largest number of common genetic variants which affect risk of the disease. The focus of our research also extends to clinical application of our findings and to targeted screening of men at high risk of prostate cancer.

Recent Relevant Publications

1. Eeles RA et al, **Kote-Jarai, Z** and Douglas Easton. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013 45(4):385-91
2. **Kote-Jarai Z** et al. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression, Hum Mol Genet. 2013 22(12):2520-8.
3. 3.Castro E,… **Kote-Jarai Z**, Eeles R. Germline BRCA Mutations Are Associated With Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer. J Clin Oncol. 2013 May 10;31(14):1748-57
4. Leongamornlert D, et al & **Kote-Jarai Z.** Germline BRCA1 mutations increase prostate cancer risk. Br J Cancer 2012;106(10):1697-701.
5. **Kote-Jarai Z** et al. *BRCA2* is a moderate penetrance gene contributing to young onset prostate cancer: implications for genetic testing in prostate cancer patients. Br J Cancer. 2011; 105(8):1230-4.
6. **Kote-Jarai Z**, et al. Seven novel prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nature Genetics 2011;43(8):785-91.
7. Al Olama AA**, Kote-Jarai Z** et al Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet. 2009; 41(10): 1058-60
8. Eeles RA, **Kote-JArai Z** et al. Identification of seven novel prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009;41(10):1116-21
9. Eeles R, **Kote-Jarai Z** et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genetics 2008; 40(3):316-321.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

High-definition reconstruction of sub-clonal composition and sub-clone specific computational analyses across cancers.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Ville Mustonen, Sanger Institute, ICGC's Bioinformatics Analyses and Mutation Consequences and Pathways working groups.

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Dr. Andrej Fischer/a new recruit and Ignacio Vazquez Garcia (PhD student)	

Background and preliminary data

Background

Adult cancers often exhibit extensive genetic heterogeneity in the form of sub-clonality that complicates their interpretation. Understanding such population variation is important - it can affect therapy success possibly because of pre-existing low frequency resistance mutations harboured by sub-clones. Computationally, attempts to characterise sub-clones is challenging as the main genomic assay, next generation sequencing, does not yield direct information on long range haplotypes when applied to mixed cell populations such as cancers. Here we propose to apply a probabilistic algorithm cloneHD (Fischer et al, *in preparation*) to perform clone reconstruction from short read data across the pan-cancer data set. The method factors in, and exploits, possible correlations across time (longitudinal data), across space (multi region and/or metastatic samples) and across genomes caused by events such as copy number changes.

Preliminary data

We have tested the performance of the approach using simulated data and by applying it to two case studies. A single breast cancer sample (1,2) and time resolved samples of chronic lymphocytic leukemia (3). The output of the method consist of number of clones detected in the sample, their population frequencies across time or space, and finally clone specific posterior genotype probabilities for locus specific copy numbers and all somatic variants detected. Figure 1a) depicts somatic SNP frequency trajectories (>4000) across five time points from CLL003 sample (see Ref (3)). Figure 1b) shows genome wide reads depths for these samples with a visible loss/gain dynamics in chromosome 8. We ran cloneHD on this data set to reconstruct sub-clone dynamics of this cancer Figure 1c). The inferred evolution closely matches the one inferred in Ref (3) via deep sequencing of the observed coding variants from WGS to a mean depth of 100000X Figure 1d). cloneHD does not detect the small clone seen only at time point a) in the targeted deep sequencing and splits the green clone from Figure 1d) to two separate clones. cloneHD clones 2 and 4 are close in their genotypes based on a clustering analysis (data not shown) and indeed forcing a three clone solution merges them together. As such the reconstruction problem is mathematically demanding and often more than a single solution explains the data almost or equally well. We have noticed that this degeneracy is greatly reduced for cancers with more than a single sample. Nevertheless, the analysis can also work for single sample tumours, e.g. , we have used cloneHD to data from Refs (1,2) with good results.

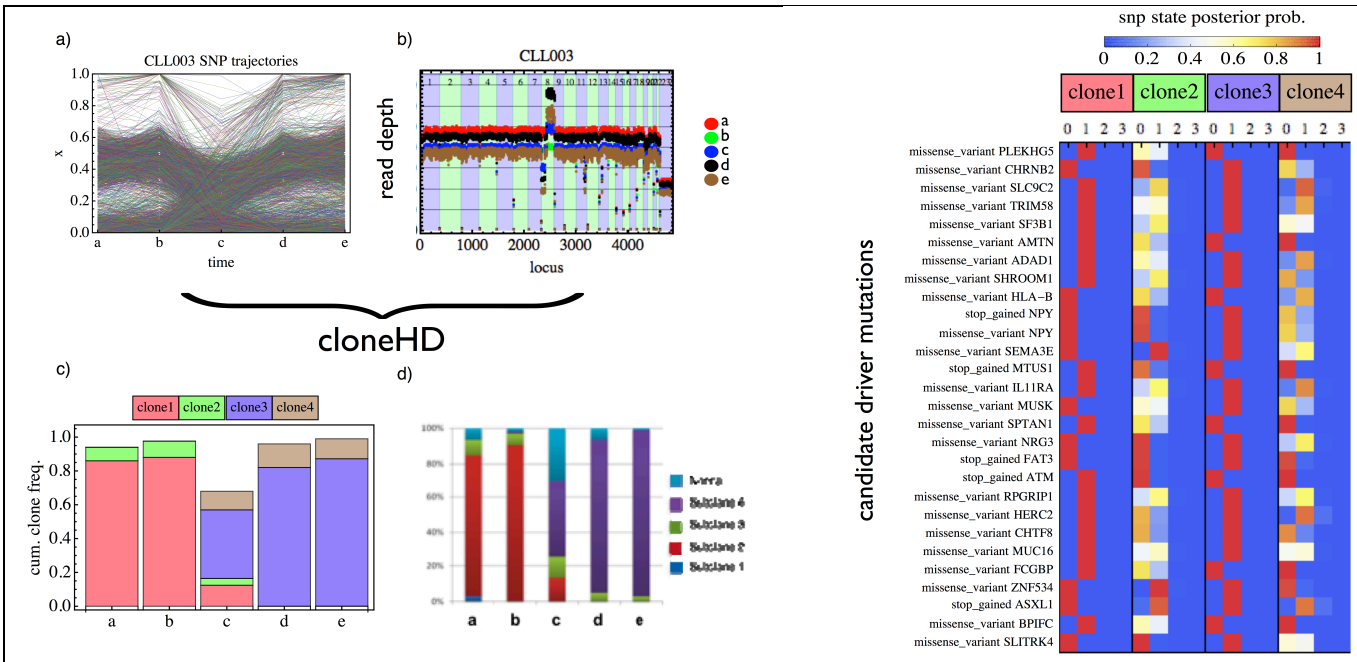


Figure 1 (left): Application of cloneHD to CLL003 data (see Ref (3) for the study reporting the data). a) somatic SNP trajectories across five time points. b) Genome-wide read depth profiles for these samples. c) Sub-clonal evolution as inferred by cloneHD (white depicts normal). d) From SI text of Ref (3) (blue depicts normal), inferred evolution using targeted deep sequencing closely matches (see text) what we get from NGS with cloneHD.

Figure 2 (right): Application of cloneHD to CLL003 data (see Ref (3)). Sub-clone specific posterior probabilities for somatic SNPs for variants that ensembl variant effect predictor lists to have a possible effect. For instance, PLEKHG5 missense mutation, is part of clone1 (one copy) with high confidence but is missing from clone3.

[1] Nik-Zainal et al. Mutational processes molding the genomes of 21 breast cancers. Cell (2012) vol. 149 (5) pp. 979-93

[2] Nik-Zainal et al. The Life History of 21 Breast Cancers. Cell (2012) vol. 149 (5) pp. 994-1007.

[3] Schuh et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. Blood (2012) vol. 120 (20) pp. 4191-6

Timelines & resources dedicated to project

Our analysis depends on good quality set of somatic SNP and B-allele calls (we need read depth and variant allele counts for all loci where calls have been made). Dr. Fischer is to focus first on the sub-clone inference but he is likely to leave during the coming spring. In that case we plan to recruit a post-doc whose main project would be cancer evolution and this pan-cancer analysis. Depending on the timelines and what kind of a candidate we manage to recruit; if needed the PI (Mustonen) will execute the analysis himself. Our student, Ignacio (second author in our cloneHD manuscript) is expected to contribute to but not drive this analysis as he has already other big commitments.

Research proposal

We propose to run cloneHD on all of the ~ 2000 whole cancer genomes focusing on cases where more than a single sample is available (e.g. pancreatic, prostate, CLL, breast, chronic myeloid and ovarian cancers). The first task will be to estimate the number of sub-clones and their fractions in these samples. Then depending on the details (sequencing depth, sub-clone sizes, and number of samples for the given cancer) we expect to have some number of cancers where we can reconstruct sub-clone specific genotypes (see preliminary work Figure 2) to be used for further computational analyses. The analyses we are thinking contain running of our mutation signature algorithm EMu (4) on the sub-clones, statistics of what kind of candidate functional mutations different clones have accumulated and also trying to understand if the inferred sub-clones can be interpreted via an evolutionary tree. Presently cloneHD does not assume an evolutionarily tree connecting the clones but we are considering to include such a constraint in the future.

(4) Fischer et al. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology (2013) vol. 14 (4) pp. R39

Legacy plans

We share and/or publish all algorithms that are used in this project. We are happy to share cloneHD with the community upon submission of our manuscript describing it (planned for Dec. 2013).

Ville Mustonen

The Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, United Kingdom
http://www.sanger.ac.uk/research/faculty/vmustonen/
vm5@sanger.ac.uk

Research Statement

Our group focuses on discovering and understanding functionally relevant genetic variation by utilising and developing computational genomics and evolutionary theory based methodology. We increasingly work with systems with direct relevance to human health, e.g., in the context of cancer genomics and evolution of drug resistance where we have ongoing, large scale, experimental evolution assays running with collaborators. Our work is collaborative and cross-disciplinary. We have a record of successful research collaborations working together with clinicians and experimentalists.

Summary of Academic Work

10/2009-present. Computational Genomics Faculty member at the Sanger Institute
Leading a research group in computational genomics and evolutionary theory.

11/2004-10/2009. Postdoctoral Fellow at University of Cologne.
Researcher in the group of Professor Michael Lässig working on quantitative biology.

Education

University of Oxford, United Kingdom
2002-2005 Doctor of Philosophy, Physics
Thesis: *Wetting, filling and interface dynamics*. Supervisor: Professor Douglas Abraham.

Helsinki University of Technology, Finland
1997-2001 Master of Science, Computational Eng., GPA: 4.4/5.0, *with distinction*

Summary of Publications

- Methods to quantify selection from time-resolved genetic data (Genetics 2011, Bioinformatics 2012, MBE 2012, PLOS Pathogens 2012).
- An evolutionarily informed scoring system for cancer mutations and methods for studying mutational signatures in the cancer genome (Genetics 2011, Genome Biology 2013).
- Identification of key observables for adaptive dynamics and predicting their behaviour under various evolutionary scenarios (PNAS 2010).
- Generalisation of the notion of fitness landscape to systems where selection itself is dynamic leading to a new concept of fitness seascape (TIGS 2009).
- Quantitative models for transcription factor binding site evolution (PNAS 2005,2008).
- A new statistical framework to analyse genomic polymorphism and substitution data (PNAS 2007).

Refereeing

- Reviewer for journals including *Science, PLOS Biology, PLOS Pathogens, Genome Research, Genetics, PLOS Comp. Biol., PLOS Genetics, Bioinformatics, Nucleic Acid Research, Molecular Biology and Evolution, Cell Reports, Molecular Ecology, Phys. Rev. Lett., Euro Phys. Lett.*

Dr. Andrej Fischer Curriculum Vitae

Contact Information

Address Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, CB10 1SA, Hinxton, United Kingdom

Phone +44-1223-49-2393

Email af7@sanger.ac.uk

Academic Positions

04/2012–present **Wellcome Trust Sanger Institute,** Postdoctoral fellow with Dr. Ville Mustonen, mathematical models and computational methods for cancer genomics and cancer therapy.

Education

02/2009–12/2011 **Doctor of Philosophy,** Institute for Theoretical Physics, University of Cologne, magna cum laude (0.7), thesis title: *Minimal models of evolution: germline fitness based scoring of cancer mutations and stochastic tunneling under strong recombination*, supervisor: Professor Alexander Altland.

10/2009–09/2010 **Research project,** Wellcome Trust Sanger Institute, project title: fitness based scoring of somatic cancer variation, host: Dr. Ville Mustonen.

10/2003–12/2008 **Diploma in Theoretical Physics,** Institute for Theoretical Physics, University of Cologne, final grade: 1.1 (highest possible: 1.0), thesis title: *Semi classical methods for discrete stochastic processes*, thesis grade: 1.0, supervisor: Professor Alexander Altland.

06/2002 **Abitur,** German final secondary school examination, Burggymnasium Essen, final grade: 1.2 (highest possible: 1.0)

Awards

04/2012–03/2013 **German Research Foundation (DFG),** Postdoctoral research fellowship.

10/2007–12/2011 **Bonn-Cologne Graduate School for Physics and Astronomy (BCGS),** Student scholarship.

02/2006–09/2008 **German National Academic Foundation,** Student scholarship.

Conferences, talks and posters

09/2013 **Bertinoro Computational Biology Meeting,** Bertinoro, Italy, invited talk title: *The Value of Monitoring in Controlling Evolving Populations.*

07/2013 **International Cancer Genome Consortium,** Mutation Consequences and Pathways Subgroup of the Bioinformatics Analyses Working Group, talk title: *Localizing Mutational Processes in the Cancer Genome.*

10/2012 **Memorial Sloan-Kettering Cancer Center,** New York City, NY, USA, Symposium on Systems Biology of Diversity in Cancer, poster title: *Probabilistic inference of mutational processes: application to 21 breast cancers.*

Ignacio Vázquez-García

CONTACT INFORMATION

Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
United Kingdom

Date of birth: 26/12/1988
Tel.: +44 (0) 7534 622 486
E-mail: ivg@sanger.ac.uk
Website: www.damtp.cam.ac.uk/people/i.vazquez-garcia/

EDUCATION

10/2011 – present **DAMTP, University of Cambridge**, Cambridge, United Kingdom
Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

Ph.D. student, Wellcome Trust PhD Programme in Mathematical Genomics and Medicine

Four-year doctoral training programme in mathematical and computational biology. The first two terms of the first year were devoted to acquiring advanced computational skills and a solid understanding in life sciences; culminating in two 8-week rotations projects.

Currently carrying out my PhD under the supervision of Dr Ville Mustonen (WTSI) and Prof. Simon Tavaré (DAMTP/CRUK CI). My research interests are in developing analytical, computational and experimental methods to unravel the laws of evolution: how mutations create variation, while genetic drift, recombination, and selection alter the variant frequencies. A main focus thus far has been on inference methods for time-dependent models of selection and adaptive genome evolution.

09/2007 – 07/2011 **Imperial College of Science, Technology & Medicine**, London, United Kingdom

MSci Physics with a Year in Europe

Degree course focused on the fundamental aspects of theoretical and experimental physics, and mathematical methods in the physical sciences.

Modules studied include: Computational Physics, General Relativity, Statistical Physics, Quantum Field Theory, Quantum Optics, Quantum Theory of Matter.

Awarded grade: First Class Honours

09/2009 – 07/2010 **Swiss Federal Institute of Technology (EPFL)**, Lausanne, Switzerland

Year abroad under the ERASMUS/SOCRATES exchange programme

Modules studied include: Lasers, Numerical Methods, Statistical Mechanics.

MSci Project at the "Laboratory of Photonics and Quantum Measurements" (LPQM).

- Dissertation Topic: *Novel Optomechanical Resonators for Ground-State Cooling*
- Adviser: Prof. Tobias J. Kippenberg

TEACHING

01/2013 – 06/2013 Supervisor for the Part II course of the Mathematics Tripos in "Statistical Physics", DAMTP, University of Cambridge, United Kingdom.

02/2010 – 06/2010 Teaching assistant for the first-year undergraduate course "Physique générale II", Department of Geoscience, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

PUBLICATIONS

- A. Fischer, I. Vázquez-García, C. Illingworth and V. Mustonen, *High-definition reconstruction of clonal composition from next-generation sequencing data* (in preparation).

CONFERENCES AND WORKSHOPS

- *Quantifying selection in evolving populations using time resolved genetic data* (contributed talk), Quantitative Laws of Genome Evolution, Como (Italy), 27 to 5 July 2013.
- *Cell-fate decisions in mammalian immune cells* (contributed talk), British Society for Immunology, Mathematical Modelling Affinity Group Meeting, Cambridge (United Kingdom), 25 to 26 June 2012.

HONOURS AND AWARDS

Ibercaja Predoctoral Research Scholarship, Fundación Ibercaja, 2011
Sir Arthur Lacland Prize in Languages, Imperial College London, 2009
Scholar of the Deutsche Akademische Austauschdienst (DAAD), 2009

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

A comprehensive characterization of the mutational processes operative in human cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Mike Stratton, Wellcome Trust Sanger Institute
Peter J. Campbell, Wellcome Trust Sanger Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Ludmil B. Alexandrov, Wellcome Trust Sanger Institute Serena Nik-Zainal, Wellcome Trust Sanger Institute	N/A

Background and preliminary data

All cancers originate from a single cell that starts to behave abnormally due to the acquired somatic mutations in its genome. These somatic mutations may be the consequence of the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagenic exposures, enzymatic modification of DNA, or defective DNA repair. Different mutational processes often generate different combinations of mutation types, termed 'mutational signatures'. Until recently, the knowledge of the mutational processes and their mutational signatures has been very limited.

Recently, we developed a novel mathematical approach that allows deciphering the signatures of mutational processes from somatic mutations derived from next generation sequencing data (Alexandrov *et al.*, Cell Reports 3 (1), 246-259). Initial application of this approach was performed on the somatic mutations derived from the whole genomes of 21 breast cancer patients revealing multiple distinct mutational signatures of substitutions (Nik-Zainal *et al.*, Cell 149 (5), 979-993). Deciphering the distinct mutational signatures operative in these breast cancer samples provided the means for timing their activity across different cancer sub-clones (Nik-Zainal, Van Loo, Wedge *et al.*, Cell 149 (5), 994-1007).

We recently expanded the mutational signature analysis performed on the 21 breast cancer genomes by examining almost 5 million somatic mutations identified in 7,042 cancer samples (507 from whole genome and 6,535 from whole exome sequences) from 30 different types of human cancer (Alexandrov et al., Nature 500 (7463), 415-421). This first global signatures analysis revealed more than 20 mutational processes operative in human cancer. Each of the cancer types had at least two mutational processes operative in it, while some (e.g., cancers of the liver and uterus) had up to 6 distinct mutational processes. Remarkably, most of the cancer samples had at least two mutational signatures active in them. We were able to propose etiology for 11 of these mutational processes.

These analyses, however, had a number of shortcomings that could be addressed in the forthcoming ICGC Pan-Cancer analysis. If this was done there could be very substantial improvements in the biological insights obtained into mutational signatures.

- They were restricted to certain classes of mutations, namely substitutions and indels, with no attention to rearrangements and copy number changes.
- In the previous studies, indel calling was highly variable in quality and therefore there could only be limited exploration of indel-based signatures.
- Most of the cancer cases previously analyzed were exomes. Power calculations and empirical observations indicate that, in general, a small number of whole genome sequences are more powerful than a large number of exome sequences in extracting substitution and indel signatures. Indeed, in some cancer types the number of substitutions and indels available from exome sequences was so limited that only a very crude assessment of mutational signatures was possible.
- Some cancer types were not included at all in the previous analyses, particularly certain childhood cancers and some cancer types in which there are known exposures but data was not available (e.g., aristolochic acid and aflatoxin induced cancers).
- The extent to which the genome landscape was introduced into signature characterization was limited to transcriptional strand bias. In principle, there could be many other features of the landscape that distinguish between signatures and hence provide further insights into their etiologies and mechanisms.
- The extent of sequence context was a single base 5' and 3' to the mutated base, but this could potentially be extended further using large numbers of whole genomes.
- Associations between mutational signatures and known etiological factors, e.g. exposures or mutated cancer genes, were made previously but could be much elaborated in a more wide ranging analysis including methylation profiles

International Cancer Genome Consortium

and other potential genomic or clinical features.

- Since the previous publications we have been working on methodological improvements in mutational signature extraction to address previous limitations. We anticipate that these will increase the discrimination and characterization of mutational signatures allowing us to progress to a definitive set.

We therefore propose to apply our previous knowledge and expertise for identifying signatures of mutational processes to the pan-cancer dataset and create a comprehensive atlas of mutational processes operative in human cancer.

Timelines & resources dedicated to project

The mutational signatures analysis will not directly use sequencing data but it will rather rely on already identified somatic variants. As such, the timeline for this project will be highly dependent on the availability of the core variant calling data. We will performed analysis both rapidly (as data becomes available) and globally (when all data is finalized):

1. Mutational signatures analysis in ~8,000 exomes:
    a. Separate analysis by cancer type (December 2013 to September 2014, as data becomes available)
    b. Global analysis in all exomes (July 2014 to September 2014)
2. Mutational signatures analysis in ~2,000 whole genomes:
    a. Separate analysis by cancer type (December 2013 to September 2014, as data becomes available)
    b. Global analysis in all genomes (July 2014 to September 2014)
3. Generating of consensus mutational signatures from genome and exome data (July 2014 to September 2014).
4. Final association analysis.

The analysis by cancer type will allow rapid results and checkpoints throughout the project. We will try to perform association analysis for each cancer type, provided there is sufficient data and the final association analysis will be performed with the consensus mutational signatures.

While mutational signature analysis is computationally intensive, we expect that the computing resources available at the Sanger Institute will be sufficient to perform this analysis. However, it may be helpful in some regards to use the cloud resources that will be available to ICGC.

Research proposal

We propose to perform extensive mutational signatures analysis on these Pan-Cancer data. We will start by using our previously developed mathematical model and framework (Alexandrov et al., Cell Reports 3 (1), 246-259) and extend it to include additional mutation types.

Previously, our analysis incorporated base substitutions and their immediate sequencing context (with and without transcriptional strand bias) as well as small insertions and deletions (indels) and *kataegis*. In principle, our framework can be applied to a wider repertoire of mutation types. We will incorporate rearrangements and copy number changes (and potentially even epigenetic changes, depending on the available data) to provide a comprehensive overview of the operative mutational processes. Our mathematical model is easily extendable to include these additional features and only minor modifications are required to our computational framework.

Mutational signatures extraction will be performed both on a cancer-by-cancer basis as well as in a global set including all samples. Analysis will be performed separately for cancer genomes and exomes. Global consensus mutational signatures will be derived by unsupervised non-negative matrix factorization of all derived mutational signatures. We will try to identify the nature of these global mutational signatures by associating them with genomic features or the presence of somatic mutations.

Previously, we performed a targeted association study between mutations in known genes and the activity of mutational processes (*e.g.*, mutation of MLH1 associated with the activity of one of the identified mutational signatures). Here, we propose to build upon this and provide a systematic analysis of the association between mutational signatures and somatic mutations in known cancer genes.

Further, we will elaborate our current methods to improve the extraction and characterization of mutational signatures. Lastly, to localize the regions of the genome where the mutational processes are active, we will use the consensus mutational signatures to estimate the activity of each process across different sequence windows. The size of the windows will be cancer type dependent as different cancer types exhibit different prevalence of somatic mutations. We will attempt to associate any clustering of somatic mutations belonging to a specific mutational process with features from ENCODE.

Legacy plans

As in all our previously mutational signatures analyses, all the required data and software will be freely available to download (e.g., see Alexandrov et al., Nature 500 (7463), 415-421) and replicate the results. Previously, we provided only a MATLAB implementation of our framework but we plan to provide an ANSI C implementation and R package to make it more useable for the wider bioinformatics community.

CURRICULUM VITAE

Full Name: Michael Rudolf STRATTON
Present Position Director, Wellcome Trust Sanger Institute

Qualifications	**Institution**	**Dates**
BA Physiological Sciences	Brasenose College, Oxford	1979
MB BS (Medical Degree)	Guys Hospital, London	1982
PhD	Institute of Cancer Research, London	1989
FRCPath	London	1991

Fellowships

Fellow of the Academy of Medical Sciences	FMedSci	1999
Fellow of the Royal Society	FRS	2008
Member of EMBO		2010

Prizes and Awards

2013 The GHA Clowes Award, American Association of Cancer Research, USA
2013 The Louis-Jeantet Prize for Medicine, the Louis-Jeantet Foundation, Switzerland
2013 The Sergio Lombroso Award in Cancer Research, Weizmann Institute, Israel
2013 The Ernst W Bertner Award, MD Anderson Cancer Centre, USA
2013 The AACR Distinguished Lecturer in Breast Cancer Research, San Antonio Breast Cancer Symposium, USA
2012 Estella Medrano Memorial Lecture Award, Society for Melanoma Research, USA
2012 Royal Physiographical Society Medal, Lund, Sweden
2011 The Massachusetts General Hospital Award in Cancer Research, Massachusetts General Hospital, USA
2010 The C. Chester Stock Award, Memorial Sloan Kettering Cancer Centre, USA
2010 The Lila Gruber Award for Cancer Research, American Academy of Dermatology, USA
2008 The BioMedicum Helsinki Medal, Finland
2007 The AstraZeneca Award, Biochemical Society UK.
2002 The Lennox K Black Award for Excellence in Medicine, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
2001 The Tom Connors Award, British Association of Cancer Research, UK

SELECTED RECENT PUBLICATIONS

Alexandrov LB, Nik-Zainal S, ..., Campbell PJ and **Stratton MR**. Signatures of mutational processes in human cancer. ***Nature*** 2013, 500, 415-21.

Alexandrov LB, Nik-Zainal S, … Campbell PJ, **Stratton MR**. Deciphering signatures of mutational processes operative in human cancer. ***Cell Rep***. (2013), 3, 246-59

Nik-Zainal S, Van Loo P, ... Futreal PA, **Stratton MR**, and Campbell PJ. The life history of 21 breast cancers. ***Cell*** 2012, 149, 994-1007.

Nik-Zainal S, Alexandrov LB, ... Futreal PA, Campbell PJ and **Stratton MR**. Mutational processes molding the genomes of 21 breast cancers. ***Cell*** 2012, 149, 979-993.

Stephens PJ, Tarpey PS, … Campbell PJ, P Futreal PA and **Stratton MR**. The landscape of cancer genes and mutational processes in breast cancer. ***Nature*** 2012, 486, 400-404.

Pleasance ED, Stephens PJ … **Stratton MR**, Futreal PA, Campbell PJ. A small-cell lung cancer genome with complex signatures of tobacco exposure. ***Nature*** 2010 463, 184-190.

Greenman C, Stephens P, … Wooster R, Futreal PA, **Stratton MR**. Patterns of somatic mutation in human cancer genomes. ***Nature*** 2007***,*** 446, 153-158.

CURRICULUM VITAE

NAME	POSITION TITLE
Dr Peter J Campbell	Head of Cancer Genetics & Genomics, Wellcome Trust Sanger Institute

EDUCATION/TRAINING

FIELD OF STUDY	INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED
Mathematics and Statistics	University of Otago, New Zealand	BSc Hons (1st Class)	1994
Medicine	University of Otago, New Zealand	MB ChB (Distinction)	1995
Haematology	Royal Australasian College of Physicians	FRACP	2003
Haematology	Royal College of Pathologists of Australasia	FRCPA	2003
Haematology	University of Cambridge	PhD	2006

SELECTED PEER-REVIEWED PUBLICATIONS

Papaemmanuil E, Rapado I, …, Greaves M **and Campbell PJ**. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. **Nature Genetics** 2013 (in press).

Alexandrov LB, Nik-Zainal S, …, **Campbell PJ** and Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013, 500(7463), 415-21.

Nik-Zainal S, Van Loo P, … Futreal PA, Stratton MR, **and Campbell PJ**. The life history of 21 breast cancers. **Cell** 2012, 149(5), 994-1007.

Nik-Zainal S, Alexandrov LB, … Futreal PA, **Campbell PJ** and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. **Cell** 2012, 149(5), 979-993.

Papaemmanuil E, Cazzola M, …Futreal PA, Stratton MR, **and Campbell PJ**. Somatic *SF3B1* mutation in myelodysplasia with ring sideroblasts**. N Engl J Med** 2011, 365(15):1384-95.

Stephens PJ, Greenman CD, …, Futreal PA **and Campbell PJ**. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. **Cell** 2011, 144(1), 27-40.

Greenman CD, Pleasance ED, …Futreal PA, Stratton MR, **and Campbell PJ.** Estimation of rearrangement phylogeny for cancer genomes. **Genome Res**. 2012, 22(2), 346-61.

Campbell PJ, Yachida S,… Iacobuzio-Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. **Nature** 2010, 467(7319), 1109-13.

Pleasance ED, Stephens PJ, … Stratton MR, Futreal PA, **and Campbell PJ**. A small cell lung cancer genome with complex signatures of tobacco exposure. **Nature** 2010, 463(7278), 184-90.

Campbell PJ, Stephens PJ, , … Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics** 2008, 40(6), 722-9.

Campbell PJ, Pleasance ED, … Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. **Proc Natl Acad Sci USA** 2008, 105(35), 13081-6.

Campbell PJ, Scott LM, … Harrison CN, Green AR. Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. **Lancet** 2005, 366(9501), 1945-1953.

CURRICULUM VITAE

Full Name: Ludmil B. Alexandrov
Present Position PhD student at Wellcome Trust Sanger Institute

Qualifications	**Institution**	**Dates**
BS Computer Science (Summa Cum Laude)	Neumont University	2005-2007
MPhil Computational Biology (Distinction)	University of Cambridge	2009-2010
PhD Molecular Biology	University of Cambridge	2010-current

Research Experience

Cancer Genome Project, Wellcome Trust Sanger Institute	2010-current
Usheva Lab, Harvard Medical School	2007-2010
Nuclear Safeguards Science & Technology, Los Alamos National Lab	2005-2006
Applied Mathematics and Plasma Physics Group, Los Alamos National Lab	2004-2005

Scholarships and Awards

2010 PhD scholarship at the Wellcome Trust Sanger Institute, University of Cambridge
2010 Director's award for distinctly high performance on the MPhil program
2009 MPhil scholarship at Department of Applied Mathematics and Theoretical Physics, University of Cambridge
2005 Grady Booch scholarship, covering 100% tuition at Neumont University
2006 Winner of CA Technologies's Iron man software coding competition
2003 Winner of the mathematical tournament of the cities organized by the Russian Academy of Sciences and Kvant magazine
2002 Bronze medal at the International Olympiad in Informational Technologies
2002 First place at the Bulgarian National Olympiad in Information Technologies

PUBLICATIONS

I have 18 published manuscripts (including ones in press) and I have presented my research at 11 conferences (including 6 oral presentations, two of which were invited talks and one was a keynote talk). According to Google Scholar my published manuscripts have been cited a total of 509 times and my h-index is 9. Selected list of publications is given below.

Alexandrov LB, Nik-Zainal S, ..., Campbell PJ and Stratton MR. Signatures of mutational processes in human cancer. ***Nature*** 2013, 500, 415-21.

Alexandrov LB, Nik-Zainal S, … Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. ***Cell Rep***. (2013), 3, 246-59

Nik-Zainal S, **Alexandrov LB**, ... Futreal PA, Campbell PJ and **Stratton MR**. Mutational processes molding the genomes of 21 breast cancers. ***Cell*** 2012, 149, 979-993.

Nik-Zainal S, Van Loo P, ... Futreal PA, Stratton MR, and Campbell PJ. The life history of 21 breast cancers. ***Cell*** 2012, 149, 994-1007.

Murchison EP, Schulz-Trieglaff OB, ... Evers DJ, Stratton MR. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. ***Cell*** 2012, 148, 780-791.

Alexandrov LB, … Usheva A. Computer Modeling Describes Gravity-Related Adaptation in Cell Cultures. ***PLOS ONE*** e8332

Serena Nik-Zainal (MA, MBBChir, MRCP, PhD) came to the UK from Malaysia on a PETRONAS Malaysia Berhad scholarship. She completed undergraduate pre-clinical studies with a 1st Class Honours and graduated from medicine at Cambridge University 2001. After general professional training as a physician, Serena specialised in Clinical Genetics, before she undertook a PhD with Mike Stratton at the Wellcome Trust Sanger Institute (WTSI) on the Wellcome Trust Clinical Research Training programme, exploring breast cancer using next-generation sequencing (NGS) technology.

During her research training, she demonstrated how detailed analyses of all mutations present in whole-genome sequenced breast cancers could reveal mutational *signatures,* imprints left by the mutagenic processes that have occurred throughout cancer development. In particular, she identified a novel phenomenon of localised hypermutation termed "kataegis". Furthermore, capitalising the digital nature of NGS technology, the principles of constructing an evolutionary tree were developed in an ultradeep-sequenced cancer and revealed striking intra-tumour heterogeneity present in breast cancers.

Serena has been awarded the Robin Winter Prize (CGS 2012), the Susan G. Komen Prize (EACR 2012), the AACR-Pezcoller Scholar-in Training Award (2013) and the Wellcome-Beit Prize (2013) for her work. Serena has since been awarded a Wellcome Trust Intermediate Clinical Research Fellowship and is currently an Honorary Consultant in Clinical Genetics. She is currently pursuing biological understanding of the mutational signatures that were identified during her PhD. Serena continues to be involved in analyses of large datasets. Serena is keen to see the responsible implementation of genomic medicine and is part of the Society and Personal Genomes Working Group at the Wellcome Trust Sanger Institute.

Selected publications:

1. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., et al. (2013) Signatures of mutational processes in human cancer, Nature 500, 415-421.
2. Alexandrov, L.B., Nik-Zainal, S., Wedge, D. C., et al. (2013) Deciphering signatures of mutational processes operative in human cancer, Cell Reports 3, 246-259.
3. Taylor, B.J.**, Nik-Zainal, S.**, Wu, Y.L., (2013) DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. Elife 2.
4. Nik-Zainal, S., Alexandrov, L. B., Wedge, et al. (2012) Mutational Processes Molding the Genomes of 21 Breast Cancers, Cell 149, 979-993.
5. **Nik-Zainal, S., Van Loo, P.**, Wedge, D. C.**, et al. (2012) The life history of 21 breast cancers, Cell 149, 994-1007.
6. Stephens, P. J., Tarpey, P. S., Davies, H., et al. (2012) The landscape of cancer genes and mutational processes in breast cancer, Nature 486, 400-404.
7. Papaemmanuil, E., Cazzola, M., Boultwood, J., et al. (2011) Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts, The New England journal of medicine 365, 1384-1395.
8. Nik-Zainal, S., Strick, R., Storer, M., et al (2011) High incidence of recurrent copy number variants in patients with isolated and syndromic Mullerian aplasia, J Med Genet 48, 197-204.
9. Nik-Zainal, S., Cotter, P. E., Willatt, L. R., Abbott, K., and O'Brien, E. W. (2011) Ring chromosome 12 with inverted microduplication of 12p13.3 involving the Von Willebrand Factor gene associated with cryptogenic stroke in a young adult male, Eur J Med Genet 54, 97-101.
10. Campbell, P. J., Yachida, S., Mudie, L. J., et al. (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer, Nature 467, 1109-1113.
11. Mefford, H. C., Sharp, A. J., Baker, C., et al. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes, The New England journal of medicine 359, 1685-1699.
** denotes shared first author

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The transcriptional consequences of somatic mutation across human cancer.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Adam Shlien PhD[1] and Peter Campbell MD, PhD[2]

(1) Hospital for Sick Children, ICGC Breast Cancer Working Group and Expression Subgroup. (2) Wellcome Trust Sanger Institute, ICGC Bone Cancer and Chronic Myeloid Disorders Groups.

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Roland Arnold, PhD[1]; Young Seok Ju, MD, PhD[2]	Michael Brudno, PhD[1]

Background and preliminary data

The disordered transcriptomes of cancer encompass direct effects of somatic mutation on transcription; co-ordinated secondary alterations in transcriptional pathways; and increased transcriptional noise. Altered transcript structure can take many forms. **There has been little systematic effort to describe, measure and quantify first-order transcriptional consequences across all classes of somatic mutation found in well-annotated cancer genomes.**

We developed a suite of algorithms to exhaustively characterise the cancer transcriptome: in so doing we aimed to wring maximum detail on the structure of cancer transcripts from RNA-sequencing data. Previous work has examined gene and mutation expression alone or has focused exclusively on one facet of transcript structure (such as fusion genes or alternative splicing) without allowing for the discovery of multiple or complex events or the involvement of the antisense strand. We implemented a seed-and-extend mapping algorithm to find reads that span different regions of the genome, and then developed a discordant pair analysis algorithm, drawing these results together with a set of methods to arrange the results into biologically meaningful categories. The primary advantage of our approach is the comprehensive detection of all possible somatic changes in the transcriptome, including events missed by other methods. These would include compound events present in *cis*, such as fusion transcripts involving alternative splice forms and exon skips with cryptic splice sites; internal exon shuffling (reusage); post-transcriptional modifications such as early polyadenylation sites; non canonical transcript junctions, or those involving lowly-expressed transcripts that are not present in reference databases.

We used these methods to understand the inter-relationship between somatic mutation and transcript structure in breast cancer (Shlien A … Stratton MR and Campbell PJ, Cell, [submitted]). **We propose to apply these methods to the analysis of ICGC and TCGA's matched genome-transcriptome pairs to discover the transcriptional consequence of somatic mutation.**

Timelines & resources dedicated to project

The proposed project can start as soon as the raw RNA-Seq data is available (FASTQ or BAM files with unaligned reads). The complete project, which will integrate both DNA and RNA data, will depend on the availability of high quality somatic variants from matched genomes. Sufficient resources and staff will be available for the successful completion of this project. The team has experience handling data of this size and complexity. In addition, our collaborator, Dr. Michael Brudno, who directs the Hospital for Sick Children's Centre for Computational Medicine will make a state-of-the-art high performance computing facility available to us (>10,000 threads, 30 TB of RAM, 220TB of local storage and 2 PB of high performance storage).

Research proposal

We propose to apply our unique analysis methods to uncover the transcriptional consequences of somatic mutation.

Objective 1: *Comprehensive characterization of the somatic transcriptome.*
1A. Quality control. There is greater variability in RNA handling and RNA-Sequencing library preparation than there is for DNA sequencing. Each transcriptome will therefore undergo rigorous quality control to ensure evenness of coverage and high complexity of the sequenced reads.
1B. Structural changes. Using methods we previously developed, we will identity exon skips, exon reusages, alternative donor and acceptors, early polyadenylation sites, and fusion gene pairs, for all samples across all cancer types. Each somatic transcriptional change will be classified as in-frame, out-of-frame, canonical, or cryptic. We will also determine normalised gene expression levels (FPKM) for all genes.

Objective 2: Systematic integration of transcriptome with genome data. We will classify all changes as having a genomic or post-genomic origin. Examples of genomic changes include highly expressed genes due to somatic mutation, or tandem duplications leading to aberrant splicing. These associations may pinpoint novel mutations. Post-genomic changes are those not preceded by a genomic change, including complex transcript abnormalities, allele-specific expression and RNA edits.
2A. Effects of point mutations and RNA edits. The expression levels of each variant base will be assessed. Putative nonsense mutations (found in the genome) will be confirmed by their absence or reduction in the transcriptome. We will look for the expression of novel isoforms of genes containing genomic variants at splice donor, acceptor or branch sites. We will delve into the relationships between coding mutations and allele-specific expression, between intronic mutations and aberrant splicing, and between mutations in untranslated exonic regions and gene expression.
2B. Direct effects of genomic rearrangements on transcriptome structure. The association between genomic structural rearrangements and the resultant transcriptional aberrations will be evaluated using custom software. Each subtype of somatic rearrangement (translocations, deletions, inversions, tandem and non-tandem duplications and complex events) will be examined to look for evidence of novel or unusual transcription, including expressed fusion genes and truncated transcripts.
2C. Integration with unpublished data. We will complement ICGC's data with our own paired tumor data (RNA and DNA sequencing) from various cancer types.
2D. Pan-cancer genome-transcriptome analyses. We will explore the state of the transcriptome across all available cancer types and subtypes. We will use a network-based approach to find genes that are exclusively rearranged in one cancer type. We hope to find a consensus transcriptomic state for each subtype that extends beyond gene expression patterns.

Legacy plans

The ICGC pan-cancer analysis represents a singular opportunity to understand the effect of the genome on the somatic transcriptome. The legacy of the proposed project will be a richly annotated transcriptomic data set, that has been interwoven with paired genomic data. The results of these analyses will be freely available to the scientific community in multiple file formats (SQLite database, R data stores, and text files). The code will be maintained in a controlled versioning system (github).

Adam Shlien, PhD

The Hospital for Sick Children, Toronto, Canada. Email: adam.shlien@sickkids.ca.
Phone: 416-813-6205

Position

Scientist, Hospital for Sick Children.

Associate Director, Translational Genetics, Hospital for Sick Children.

Education/Training

Postdoctoral Fellow, Wellcome Trust Sanger Institute.	2010-2012
PhD student, Medical Biophysics, Hospital for Sick Children	2006-2010

Awards

- H.L. Holmes Award, National Research Council Canada 2011
- Long Term Fellowship, European Molecular Biology Organization 2011
- Fellowship, Canadian Institutes of Health Research (declined) 2011
- Frederick Banting and Charles Best Canada Graduate Scholarship 2009

Most Significant Publications

- Shlien A … Malkin D. Excessive germline copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci* 2008

- Shlien A … Malkin D. A common molecular mechanism underlies two phenotypically distinct 17p13.1 microdeletion syndromes. *Am J Hum Genet.* 2010

- Shlien A … Stratton MR and Campbell PJ. The direct transcriptional consequences of somatic mutation in breast cancer. *Cell* [submitted]. 2013

- Cooke SL, Shlien A … Stratton MR, McDermott UA and Campbell PJ. Processed pseudogenes acquired somatically during cancer development. *Nature* [2nd revision]. 2013

CURRICULUM VITAE

NAME	POSITION TITLE
Dr Peter J Campbell	Head of Cancer Genetics & Genomics, Wellcome Trust Sanger Institute

EDUCATION/TRAINING

FIELD OF STUDY	INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED
Mathematics and Statistics	University of Otago, New Zealand	BSc Hons (1st Class)	1994
Medicine	University of Otago, New Zealand	MB ChB (Distinction)	1995
Haematology	Royal Australasian College of Physicians	FRACP	2003
Haematology	Royal College of Pathologists of Australasia	FRCPA	2003
Haematology	University of Cambridge	PhD	2006

SELECTED PEER-REVIEWED PUBLICATIONS

Papaemmanuil E, Rapado I, …, Greaves M **and Campbell PJ**. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. **Nature Genetics** 2013 (in press).

Alexandrov LB, Nik-Zainal S, …, **Campbell PJ** and Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013, 500(7463), 415-21.

Nik-Zainal S, Van Loo P, … Futreal PA, Stratton MR, **and Campbell PJ**. The life history of 21 breast cancers. **Cell** 2012, 149(5), 994-1007.

Nik-Zainal S, Alexandrov LB, … Futreal PA, **Campbell PJ** and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. **Cell** 2012, 149(5), 979-993.

Papaemmanuil E, Cazzola M, …Futreal PA, Stratton MR, **and Campbell PJ**. Somatic *SF3B1* mutation in myelodysplasia with ring sideroblasts**. N Engl J Med** 2011, 365(15):1384-95.

Stephens PJ, Greenman CD, …, Futreal PA **and Campbell PJ**. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. **Cell** 2011, 144(1), 27-40.

Greenman CD, Pleasance ED, …Futreal PA, Stratton MR, **and Campbell PJ.** Estimation of rearrangement phylogeny for cancer genomes. **Genome Res**. 2012, 22(2), 346-61.

Campbell PJ, Yachida S,… Iacobuzio-Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. **Nature** 2010, 467(7319), 1109-13.

Pleasance ED, Stephens PJ, … Stratton MR, Futreal PA, **and Campbell PJ**. A small cell lung cancer genome with complex signatures of tobacco exposure. **Nature** 2010, 463(7278), 184-90.

Campbell PJ, Stephens PJ, , … Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics** 2008, 40(6), 722-9.

Campbell PJ, Pleasance ED, … Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. **Proc Natl Acad Sci USA** 2008, 105(35), 13081-6.

Campbell PJ, Scott LM, … Harrison CN, Green AR. Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. **Lancet** 2005, 366(9501), 1945-1953.

Curriculum Vitae, Dr. Roland Christian Arnold

663 Huron Street, Apartment 5, Toronto, ON M5R 2R8 CANADA
Phone: (+1) 647-707-4158, Email: roland.arnold@sickkids.ca

Research goals: I believe that integrated analyses of comprehensive genomic, transcriptomic, and proteomic data-sets are necessary to fully understand cellular alterations induced by cancer. Such analyses provide the key information necessary for personalized medicine and will greatly improve diagnostic and treatment of cancer. I aim to develop profound, tailor-made bioinformatics systems which will allow a rapid integrative analysis of cancer samples. I will apply this systems to uncover novel biology of different cancer types and to implement diagnostic procedures based on high-throughput sequencing data.

Education and work experience

Since December 2013	**Postdoctoral fellow** at the Hospital for Sick Children, Toronto, Paediatric Laboratory Medicine. Supervisor: Prof. Dr. Adam Shlien
2010 - 2013	**Postdoctoral fellow** at the University of Toronto, Terrence Donnelly Centre for Cellular and Biomolecular Research, Kimlab. Supervisor: Prof. Dr. Philip Kim
2005-2010	**Ph.D. in bioinformatics**, Technische Universität München. Dissertation title: *"Development of computational models of the chlamydial interactomes and bacterial secreted proteins".* Supervisors: Prof. Dr. Hans-Werner Mewes, Dr. Thomas Rattei Finished with '*magna cum laude*'
2000-2005	**Degree in bioinformatics** (Diplom Bioinformatiker Univ.), Technische Universität München, Ludwig Maximilians Universität München, joint curriculum. Topic of final thesis: "*Integrative Analysis of Functional Modules in Bacteria",* Supervisor: Prof. Dr. Hans-Werner Mewes

Awards and fellowships

2012-2013	Post-Doctoral fellowship from the Ontario Ministry of Research and Innovation
2010-2011	Government of Canada Awards, Post-Doctoral Research Fellowship (PDRF)
2011	European Commission ERA-NET PathoGenoMics Ph.D. Award

Key publications (in total 22 publications)

Vizeacoumar FJ*, **Arnold R*** et al**:** *A Negative Genetic Interaction Map in Isogenic Cancer Cells Reveals Cancer Cell Vulnerabilities s*
Mol Syst Biol. 2013 Oct 8;9:696 PMID: 24104479

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T., *Sequence-based prediction of type III secreted proteins,*
Plos Pathogens 2009 Apr;5(4); PMID: 19390696

Arnold R*, Rattei T*, et al., *SIMAP--The similarity matrix of proteins.*
Bioinformatics. 2005 Sep 1;21 Suppl 2:ii42-ii46. PMID: 16204123

*) co-first author

Curriculum Vitae

Young Seok Ju, MD, PhD

Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK
• Mobile phone +44-7796-002947; E-mail jueenome@gmail.com, ysj@sanger.ac.uk

EDUCATION AND TRAINING

Seoul National University College of Medicine, Seoul, Republic of Korea **2007-2010**
Ph.D. in Biochemistry and Medical Genomics, Advisor: Dr. Jeong-Sun Seo, M.D., Ph.D.

Seoul National University College of Medicine, Seoul, Republic Korea **2001-2007**
M.D. (B.S in Medicine), GPA 3.82/4.30 (17^{th} out of 173)

RESEARCH EXPERIENCE

Postdoctoral Research **Apr.2013-present**
Postdoctoral fellow, Cancer Genome Project Team, Wellcome Trust Sanger Institute, Hinxton, UK
Advisor: Dr. Mike Stratton & Dr. Peter Campbell

Postdoctoral Research **2010-2013**
Researcher, Life Science Institute, Macrogen Inc., Korea (Substitution of mandatory military service)

Selected PUBLICATIONS

1. The transcriptional landscape and mutational profile of lung adenocarcinoma. Jeong-Sun Seo*, **Young Seok Ju***, Won-Chul Lee*, Jong-Yeon Shin, June Koo Lee *et al.*, *Genome Research* 22(11). (2012) (*Co-first author)
2. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. **Young Seok Ju**, Won-Chul Lee, Jong-Yeon Shin, Seungbok Lee, Thomas Bleazard *et al.*, *Genome Research* 22(3). (2012)
3. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. **Young Seok Ju**, Jong-Il Kim, Sheehyun Kim, Dongwan Hong, Hansoo Park *et al*., *Nature Genetics* 43(8). (2011)
4. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Hansoo Park*, Jong-Il Kim*, **Young Seok Ju***, Omer Gokcumen, Ryan Mills *et al.*, *Nature Genetics* 42(5). (2010) (*Co-first author)
5. A highly annotated whole-genome sequence of a Korean individual. Jong-Il Kim*, **Young Seok Ju***, Hansoo Park, Sheehyun Kim, Seonwook Lee *et al.*, *Nature* 460 (7258). (2009) (*Co-first author)

FELLOWSHIP

EMBO (European Molecular Biology Organisation) Long Term Fellowship for Postdoctoral Research **Apr.2013–present**

Michael BRUDNO

I am a Computer Scientist who has worked extensively in designing methodologies for analyzing large-scale biological and medical data, including genomic sequencing data (both model organism and human patient), patient phenotype data, and biological networks. My Computer Science publications span Computational Biology, Machine Learning, Computational Theory, and User Interfaces.

Education

	Institution	Field of Study	Year
BA	University of California, Berkley	Computer Science & History	2000
MSc	Stanford University	Stanford University	2003
PhD	Stanford University	Stanford University	2004

Positions

2004-2005	Postdoctoral Fellow, Department of Computer Science, UC Berkeley
2005	Visiting Scientist, Massachusetts Institute of Technology (MIT)
2006-2011	Assistant Professor and Canada Research Chair (Tier 2), Dept. of Computer Science & Donnelly Centre, University of Toronto
2011-present	Senior Scientist, Genetics and Genomic Biology, The Hospital for Sick Children
2011-present	Associate Professor, Department of Computer Science, University of Toronto
2011-present	Scientific Director, Centre for Computational Medicine, Hospital for Sick Children

Select Awards

- Canadian Association for Computer Science Outstanding Young Scientist (2012)
- U. of Toronto Inventor of the Year (2012)
- MPrime Excellence in Mentorship Award
- Alfred P. Sloan Research Fellow (2010)
- Ontario Early Researcher Award (2009)

Select Publications over last 5 years (of >50 overall):

Trainees supervised by me are in bold. The complete list is at http://www.cs.toronto.edu/~brudno/

1. **Rumble SM**, Lacroute P, **Dalca AV**, **Fiume M**, Sidow A, Brudno M. SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Computational Biology*, **5**:5 2009 [Cited by 275]
2. **Lee S**, Hormozdiari F, Alkan C, Brudno M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods*. **6**:473-474 2009 [Cited by 79]
3. **Medvedev P, Stanciu M,** Brudno M. Computational methods for discovering structural variation with next generation sequencing. *Nat*. *Meth*., **6**:S13–20 2009 (Review). [Cited by 225]
4. **Fiume M, Williams V, Brook A**, Brudno M. Savant: Genome Browser for High Throughput Sequencing Data. *Bioinformatics* **26**:1938-1944 2010 [Cited by 54]
5. **Medvedev P, Fiume M, Dzamba M, Smith T**, Brudno M. Detecting Copy Number Variation with Mated Short Reads. *Genome Research* **20**:1613-1622 2010 [Cited by 62]
6. **Mezlini AM**, **Smith EJ, Fiume M, Buske O**, … Brudno M. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res*. **20**:519-29 2013 [Cited by 9]

Training of Highly Qualified Personnel

	In Progress	Completed
Masters	1	8
PhD	3	2
Postdoctoral Fellows	2	0

Current Significant Grant Funding

2010-2015	Ontario Research Fund (ORF-GL2) CAD$9,700,000; (Co-PI; subcontract: CAD$250,000)
2012-2017	National Science and Engineering Research Council Discovery Grant CAD$200,000 (PI)
2013-2016	Genome Canada Bioinformatics and Computational Biology Grant; CAD$1,000,000 (PI)
2013-2016	CIHR/NSERC Collaborative Health Research Program (CHRP) CAD$420,000 (PI)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Pan-cancer molecular archaeology: the life history of 2000 cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

David Wedge and Peter Van Loo, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK (ICGC breast and prostate cancers)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Moritz Gerstung and Peter Campbell, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK	

Background and preliminary data

The cancer genome carries within it an archaeological record of its past. We recently developed several "molecular archaeology" algorithms to disentangle the relative timings of mutations and copy number changes over the life history of a tumor from whole-genome sequencing (WGS) data, allowing detailed insight into cancer development and evolution [Nik-Zainal#, Van Loo#, Wedge# *et al.* (2012), *Cell* 149:994-1007 (#: equal contribution)]. These methods allow inference of phylogenetic relationships between competing subclones from WGS data, using a variety of analytical techniques, including the identification of subclonal copy number changes via haplotype imputation, clustering mutations into identifiable subclones and distinguishing between related and unrelated subclones by phasing mutation-mutation and mutation-SNP pairs. In addition, we can track the evolution of processes causing mutations and determine the relative order of different mutations, copy number gains and whole-genome duplications.

Further developments have extended these methods to multiple-sampling approaches, and allowed us to decrypt the phylogeny of multifocal prostate cancers and infer the presence of clonal expansions in morphologically normal prostate tissue, as well as extensive branching evolution and cancer clone mixing (Cooper#, Eeles#, Wedge#, Van Loo# *et al.*, manuscript under review). In addition, we integrated all threads of these methods to disentangle a tumor's life history and subclonal architecture, resulting in the automatic (i.e. programmatic) deduction of phylogenetic trees from WGS data [Wedge *et al.*, *in preparation*]. These analyses will benefit further from recently developed techniques for ordering series of overlapping rearrangements [Li#, Schwab#, Ryan# *et al.,* manuscript under review] and for calling subclonal mutations [Gerstung *et al.*, manuscript under review]. In addition, we have also recently derived statistical models for obtaining a temporal ordering of the acquisition of driver mutations from multiple unrelated samples [Papaemmanuil *et al.* (2013), *Blood* 122:3616-3627]. Finally, by disaggregating mutational signatures [Alexandrov *et al.* (2013), *Nature* 500:415-421], we have recently calibrated a molecular clock to the patient's age, allowing real-time chronological estimates of a tumor's life history.

Timelines & resources dedicated to project

The proposed project will greatly benefit from tools whose development has been proposed as part of other ICGC pan-cancer analyses on mutational signatures (M. Stratton and P. Campbell) and structural rearrangements (J. Korbel and P. Campbell). The most computationally expensive parts of the analysis will be the Battenberg algorithm, the Bayesian Dirichlet process and the construction of phylogenetic trees. We estimate to require approximately ~10 days of CPU time per genome, i.e. a total of ~55 CPU years for 2000 genomes.

Months 1-6: Further method development (mutation signatures and real-time molecular clock; multi-sample approaches; subclonality inference; timing driver events; taxonomy of cancer based on evolutionary history).
Months 1-6: Initial analyses (timing driver mutations using TCGA data; multi-sample pilot)
Months 7-12: Life history inference on all samples (phylogenetic tree reconstruction; life history using real-time molecular clock; taxonomy of cancer based on evolutionary history)
Months 13-15: Interpretation and writing

Research proposal

We anticipate that a pan-cancer molecular archaeology approach to obtain detailed evolutionary histories of tumors would give unprecedented insights into carcinogenesis and cancer evolution. The ICGC pan-cancer data set provides an ideal opportunity to combine the full panoply of aberrations (including rearrangements, copy number changes, point mutations and indels) from large numbers of whole genomes, and outline the life history and subclonal architecture of each tumor. All ICGC whole genomes will be analyzed within our pipeline, resulting in a timeline of development and phylogenetic tree for each case. Spontaneous deamination at CpG dinucleotides is believed to occur continuously in both healthy and neoplastic tissue, leading to C>T mutations. Using a mutational signature dominated by this process as a molecular clock will allow inferring the timing of events along a tumor's developmental history in real chronological time. The mutation signatures of each stage in a cancer's development may then be deciphered in conjunction with the evolution of mutation rates. These signatures will be augmented by the positioning of rearrangements, copy number changes and indels along the tumor's developmental lifetime (the trunk of the tree) and subsequent periods of subclonal diversification (the different branches). These analyses will allow insight into when the most recent common ancestor of all tumor cells appeared. It will greatly inform on each cancer's past, including when driver mutations appeared, when copy number changes and rearrangements occurred and when mutational processes were active. Finally, it will inform us on the extent of subclonal variegation, allowing a glimpse into the tumor's future.

Furthermore, we will combine the ICGC tumor genomes with mutation (substitution and indel) and copy number data from TCGA, aiming to time driver mutations along a cancer's lifetime. Such analyses will allow us to evaluate which cancer genes are mutated early and which are mutated late during tumor development, which drive subclonal diversification, and which drive metastasis. Finally, it will allow insight into the extent to which early driver mutations or copy number changes affect the subsequent development of a tumor.

In addition to analyzing each tumor genome in isolation, we will apply our multiple-sampling molecular archaeology approaches to the ~200 cases for which multiple genomes will be available. Our previous studies clearly demonstrated that joint analysis of multiple samples allows much greater precision in identifying subclonal clusters and their trajectory in time. This should enable us to answer questions concerning many aspects of tumor development, metastasis, progression and response to treatment. Which subclonal populations are likely to metastasize? How closely related are different metastases? Do metastases arise from a single founder cell? How different are metastases with different source and host tissues? Can we estimate growth rates of competing subclones and model tumor progression? Is there evidence for a synergistic relationship between subclones? Which subclones are resistant/sensitive to particular treatment? How do individual treatments affect mutational processes?

Combining all 2000 ICGC cases will allow us to derive a timeline of the molecular events occurring during a cancer's life. This will be an unprecedented view illustrating the typical timing of major events such as the first driver, whole genome duplication, the acquisition of a mutator phenotype and metastasis. As we can only partly reconstruct the evolutionary trajectory of each individual case, this part of the analysis will greatly benefit from combining a large number of cancer samples. In addition, a large-scale pan-cancer approach allows us to study the inter-patient and inter-tumor-type differences and similarities in life history. Finally, we also aim to derive a taxonomy of cancer based on each tumor's evolutionary history. We expect these analyses to identify (sub)groups of tumors that have traversed similar paths to cancer, with particular types of events occurring in similar succession. Subgroups may be identified within or between tumor types and potentially imply similar clinical behavior.

Legacy plans

We have recently developed techniques for calling subclonal mutations [Gerstung *et al.*, under review] and for the automatic reconstruction of phylogenetic trees [Wedge *et al.*, in preparation]. These tools will be further developed and integrated with tools for the identification of mutational signatures [Alexandrov et al (2013) Nature 500:415-421] and for the disaggregation of complex rearrangements [Li#, Schwab#, Ryan# *et al.*, under review], as well as with novel tools to derive a taxonomy of cancer based on a tumors life history, to build a package that will represent a valuable legacy of the project.

All source code for producing results and figures to be published as part of this project will be produced in R and will be freely available to the academic research community using web programming tools such as Shiny.

Finally, the derived timelines of cancer development and taxonomy of cancer developmental history will be a valuable resource to the field of cancer genomics and cancer research in general.

David C Wedge

Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SA, UK
+44 (0)1223 494887
dw9@sanger.ac.uk

Research interests

Cancer genomics, particularly the analysis of subclonality and evolution in cancer.

Education

PhD Computer Science 2002-2005
Manchester Metropolitan University (Director of Studies: David Ingram)
Thesis: *Wave Overtopping Prediction Using Global-Local Artificial Neural Networks*

MSc Software Development (distinction) 2000-2001
Huddersfield University

BA Chemistry (upper second) 1985-1989
Pembroke College, Oxford University

Employment

Senior statistician and staff scientist 2011-present
Cancer Genome Project, Wellcome Trust Sanger Institute

Postdoctoral research fellow 2010-2011
University of Manchester (Roy Goodacre's group)

Postdoctoral research associate 2009-2010
University of Manchester (Simon Hubbard's group)

Postdoctoral research associate 2006-2009
University of Manchester (Douglas Kell's group)

Selected publications

D.C. Wedge et al. Cakesuite: software tools for the analysis of subclonality in cancer, in preparation

C. Cooper*, R. Eeles*, D.C. Wedge*, P. Van Loo* et al. The Life History of Multifocal Prostate Cancer: Multiple Independent Clonal Expansions in Neoplastic and Normal Prostate Tissue, *Nature Medicine*, submitted

E.P. Murchison*, D.C. Wedge* et al. The genome of a transmissable cancer reveals the origin and history of an ancient lineage, *Science*, in press

L. Alexandrov, S. Nik-Zainal, D.C. Wedge, … , P.J. Campbell and M.R. Stratton. Signatures of mutational processes in human cancer, *Nature* (2013) 500:415-421

L. Alexandrov, S. Nik-Zainal, D.C. Wedge, P.J. Campbell and M.R. Stratton. Deciphering Signatures of Mutational Processes Operative in Human Cancer, *Cell Reports* (2013) 3:246-259

S. Nik-Zainal*, P. Van Loo*, D.C. Wedge* et al. The life history of 21 breast cancers, *Cell* (2012) 149:994-1007

* equal contribution

Curriculum Vitae Peter Van Loo

Academic achievements

- 05/2008: **PhD in Medical Sciences**, University of Leuven, Leuven, Belgium (main supervisor: Prof. Dr. Peter Marynen)
- 1999 – 2004: **Master of Science in Engineering: Cell- and Gene Biotechnology**, University of Leuven, Belgium (5 year bachelor + master program), magna cum laude
- 1998 – 2003: **Master of Science in Engineering: Electrotechnical Engineering**, University of Leuven, Belgium (5 year bachelor + master program), magna cum laude

Postdoctoral research

- Oct 2010 to date: Cancer Genome Project, **Wellcome Trust Sanger Institute**, Hinxton Cambridge, UK (Dr. Peter Campbell)
- Dec 2008 – Oct 2010: Human Genome Laboratory, Department of Human Genetics, **VIB and University of Leuven**, Belgium (Prof. Dr. Peter Marynen)
- Jun 2008 – Dec 2008: Department of Genetics, **Institute for Cancer Research, University of Oslo**, Norway (Prof. Dr. Anne-Lise Børresen-Dale and Prof. Dr. Vessela Kristensen)
- Funding: postdoctoral research grant from the Research Foundation – Flanders (FWO), 10/2008 – 9/2014 (3 + 3 years; salary funding + bench fee).

Scientific leadership experience

- Oct 2013 to date: 10% faculty position, University of Leuven, Belgium
- Directed pan-cancer copy number analysis study (developed this line of research, recruited a team of interested students, postdocs and PIs from different institutions to the project, directed the research and wrote the manuscript)
- Currently line managing one visiting researcher at the Wellcome Trust Sanger Institute
- Co-grantholder of 3 research project grants (€240k, €320k and NOK 7.5M, or totalling approx $2M)

Publication metrics

- **37 publications in international journals**, including **Nature**, **Cell** (2), **Science**, **Nature Genetics** (2), **Nature Biotechnology** (3), **PNAS** (2) and the **Journal of Clinical Oncology**, **1286 citations**[*]
- 9 first author publications in international journals, including Cell, Nature Biotechnology, PNAS and the Journal of Clinical Oncology, 558 citations[*]
- 28 publications without main PhD supervisor, including 4 first author publications
- 26 publications without current supervisor, including 7 first author publications
- **H-factor: 16**[*]
- 12 publications in 2012 (387 citations[*]), 7 publications in 2013.

Selected publications

- Nik-Zainal, S.[#], **Van Loo, P.[#]**, Wedge, D.C.[#], *et al*. (2012). The life history of 21 breast cancers. *Cell*, 149:994-1007, 105 citations[*].
- **Van Loo, P.[#]**, Nordgard, S.H.[#] *et al*. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 107:16910-16915, 43 citations[*].
- Aerts, S.[#], Lambrechts, D.[#], Maity, S.[#], **Van Loo, P.[#]**, Coessens, B.[#], *et al*. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537-544, 293 citations[*].

Selected presentations

- Fourth International Symposium on Adrenal Cancer, Paris, France, 23/02/2013, "Studying intra-tumor heterogeneity by massively parallel sequencing: mechanisms and clinical implications", **keynote presentation**
- The European Cancer Congress (ECC 2013), Amsterdam, The Netherlands, 27 September – 1 October 2013, "The landscape of tumour suppressors across primary tumours", presentation in presidential session

[*] All citations mentioned are calculated using Scopus, **excluding** self-citations of all authors.
[#] Equal contribution

Moritz Gerstung

Current address
Wellcome Trust Sanger Institute
Hinxton CB10 1SA
UK
Email: mg14@sanger.ac.uk

EDUCATION & TRAINING

Date	Institution	Degree/position	Field of study
Since Jun 2012	Wellcome Trust Sanger Institute, UK	Postdoctoral fellow	Cancer genomics
Apr 2008 – Apr 2012	ETH Zurich, Switzerland	PhD	Computational biology
Feb 2004 – Oct 2004	University of Melbourne, Australia	Visiting student	Physics
Oct 2001 – Dec 2007	University of Freiburg, Germany	MSc	Physics

PEER-REVIEWED PUBLICATIONS

Journal articles

- E. Papaemmanuil, M. Gerstung, L. Malcovati, et al. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, ***122***:3616-27.
- T. Kockmann, M. Gerstung, T. Schlumpf, et al. (2013). The BET protein FSH functionally interacts with ASH1 to orchestrate global gene activity in Drosophila. *Genome Biol,* **14**:R18.
- S. Kawamura, M. Gerstung, A. T. Colozo, et al. (2013). Kinetic, energetic, and mechanical differences between dark-state rhodopsin and opsin. *Structure,* **21**:426-37.
- M. Gerstung, C. Beisel, M. Rechsteiner, P. et al. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun*, 2
- S. Ambatipudi, M. Gerstung, M. Pandey, et al. (2012). Genome-wide expression and copy number analysis identifies driver genes in gingivobuccal cancers. *Genes Chromosomes Cancer*, 51:161-73.
- M. Gerstung, N. Eriksson, J. Lin, B. Vogelstein, and N. Beerenwinkel (2011). The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS One*, 6:e27136.
- M. Gerstung, H. Nakhoul, and N. Beerenwinkel (2011). Evolutionary Games with Affine Fitness Functions: Applications to Cancer. *Dynamic Games and Applications*, 1:370-385.
- S. Ambatipudi, M. Gerstung, R. Gowda, et al. (2011). Genomic profiling of advanced-stage oral cancers reveals chromosome 11q alterations as markers of poor clinical outcome. *PLoS One*, 6:e17250.
- D. Enderle, C. Beisel, M. B. Stadler, M. Gerstung, P. Athri, and R. Paro (2011). Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. *Genome Res*, 21:216-26.
- S. M. Pathare, M. Gerstung, N. Beerenwinkel, et al. (2011). Clinicopathological and prognostic implications of genetic alterations in oral cancers. *Oncol Lett*, 2:445-451.
- M. Gerstung and N. Beerenwinkel (2010). Waiting time models of cancer progression. *Math Pop Studies*, 17:115--135.
- M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, 25:2809-15.
- M. Gerstung, J. Timmer, and C. Fleck (2009). Noisy signaling through promoter logic gates. *Phys Rev E*, 79:011923.
- F. Geier, J. U. Lohmann, M. Gerstung, A. T. Maier, J. Timmer, and C. Fleck (2008). A quantitative and dynamic model for plant stem cell regulation. *PLoS One*, 3:e3553.
- E. Anastasiou, S. Kenz, M. Gerstung, D. MacLean, J. Timmer, C. Fleck, and M. Lenhard (2007). Control of plant organ size by KLUH/CYP78A5-dependent intercellular signaling. *Dev Cell*, 13:843--856.

SELECTED TALKS

- The impact of driver mutations on gene expression, blood counts and survival in Myelodysplastic syndromes. EMBL Conference Cancer Genomics, Heidelberg, Nov 04, 2013.
- deepSNV – analysis of subclonal mutations in tumours. Bioconductor workshop, Zurich, Dec 13, 2012.
- Revealing Intra-Tumor Heterogeneity with Ultra-Deep Sequencing, Keystone Symposium: Changing Landscape of the Cancer Genome, Boston, MA, Jun 20-24, 2011.

CURRICULUM VITAE

NAME	POSITION TITLE
Dr Peter J Campbell	Head of Cancer Genetics & Genomics, Wellcome Trust Sanger Institute

EDUCATION/TRAINING

FIELD OF STUDY	INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED
Mathematics and Statistics	University of Otago, New Zealand	BSc Hons (1st Class)	1994
Medicine	University of Otago, New Zealand	MB ChB (Distinction)	1995
Haematology	Royal Australasian College of Physicians	FRACP	2003
Haematology	Royal College of Pathologists of Australasia	FRCPA	2003
Haematology	University of Cambridge	PhD	2006

SELECTED PEER-REVIEWED PUBLICATIONS

Papaemmanuil E, Rapado I, …, Greaves M **and Campbell PJ**. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. **Nature Genetics** 2013 (in press).

Alexandrov LB, Nik-Zainal S, …, **Campbell PJ** and Stratton MR. Signatures of mutational processes in human cancer. **Nature** 2013, 500(7463), 415-21.

Nik-Zainal S, Van Loo P, … Futreal PA, Stratton MR, **and Campbell PJ**. The life history of 21 breast cancers. **Cell** 2012, 149(5), 994-1007.

Nik-Zainal S, Alexandrov LB, … Futreal PA, **Campbell PJ** and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. **Cell** 2012, 149(5), 979-993.

Papaemmanuil E, Cazzola M, …Futreal PA, Stratton MR, **and Campbell PJ**. Somatic *SF3B1* mutation in myelodysplasia with ring sideroblasts**. N Engl J Med** 2011, 365(15):1384-95.

Stephens PJ, Greenman CD, …, Futreal PA **and Campbell PJ**. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. **Cell** 2011, 144(1), 27-40.

Greenman CD, Pleasance ED, …Futreal PA, Stratton MR, **and Campbell PJ.** Estimation of rearrangement phylogeny for cancer genomes. **Genome Res**. 2012, 22(2), 346-61.

Campbell PJ, Yachida S,… Iacobuzio-Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. **Nature** 2010, 467(7319), 1109-13.

Pleasance ED, Stephens PJ, … Stratton MR, Futreal PA, **and Campbell PJ**. A small cell lung cancer genome with complex signatures of tobacco exposure. **Nature** 2010, 463(7278), 184-90.

Campbell PJ, Stephens PJ, , … Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics** 2008, 40(6), 722-9.

Campbell PJ, Pleasance ED, … Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. **Proc Natl Acad Sci USA** 2008, 105(35), 13081-6.

Campbell PJ, Scott LM, … Harrison CN, Green AR. Definition of subtypes of essential thrombocythaemia and relation to polycythaemia vera based on JAK2 V617F mutation status: a prospective study. **Lancet** 2005, 366(9501), 1945-1953.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Multi-platform based pathway analyses incorporating whole genome sequencing of 20+ TCGA/ICGC cancer types.	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Christopher Benz, MD, Buck Institute for Research on Aging; Joshua Stuart, PhD, University of California Santa Cruz.	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Christina Yau, PhD, Buck Institute for Research on Aging; Ted Goldstein, PhD, University of California Santa Cruz.	
Background and preliminary data	
We recently completed a comprehensive study of >3,000 TCGA tumors representing 12 different cancer types (acute myelogenous leukemia, glioblastoma multiforme, lung squamous cell carcinoma, lung adenocarcinoma, and cancers of the head and neck, breast, ovary, endometrium, kidney clear cell, bladder, color and rectum), and performed an integrated analysis of the multi-dimensional data obtained from 6 independent TCGA assay platforms (DNA exome sequencing, DNA methylation, DNA copy number, RNA sequencing for mRNA and microRNA expression, and reverse phase protein arrays). Unsupervised clustering analysis integrating the individual platform data revealed 11 basic cancer subtypes within the PanCan-12 tumor set, indicating that by this new subtyping 10% of cancers should be reclassified outside of their standard tissue of origin pathology designation. As part of this PanCan-12 study we also used only the DNA copy number and mRNA expression data to determine inferred pathway activities for all tumors using the PARADIGM algorithm (which employs prior pathway knowledge); and analysis of these inferred pathway activities produced an unsupervised clustering solution of multiple cancer subtypes that were highly concordant with the subtypes determined by the integrated platform analysis. While the new cancer subtypes identified from the PanCan-12 tumor set each possessed different frequencies of significantly mutated genes (SMGs) based on exome sequencing, whole genome sequencing (WGS) was not included in the analysis and the SMGs alone were not able to delineate either the basic PanCan-12 subtypes or their cancer-driving pathways. We now propose to combine TCGA and ICGC resources to expand the PARADIGM pathway analysis by including at least 8 more tumor types (generating PanCan-20+ tumor set), each additional tumor type containing >100 tumor cases with at least DNA copy number and mRNA expression data, and also to broaden the PARADIGM pathway analysis by incorporating the anticipated 2000 TCGA/ICGC WGS data available for a significant portion of the PanCan-20+ tumor set.	
Timelines & resources dedicated to project	

Define the PanCan-20+ tumor set before February 2014; obtain DNA copy number and mRNA expression data on all PanCan-20+ tumors before April 2014 (data freeze); run PARADIGM to obtain inferred pathway activities by June 2014; obtain 2000 WGS dataset and map to PanCan-20+ tumor set, run PARADIGM program on this mapped WGS subset of PanCan-20+ tumors; integrate and correlate PARADIGM pathway activities with WGS abnormalities by Sept 2014; also attempt to incorporate mutation data, especially from non-coding analysis, into subtyping using heat diffusion approaches. Complete analysis by Dec 2014; write and submit manuscript in early 2015.

Research proposal

The PanCan-20+ tumor set will be comprised of those 12 different tumor types and 3,527 individual TCGA cases with DNA copy number and mRNA expression data from which PARADIGM inferred activities were determined for the PanCan-12 analysis described above. ICGC cases falling within those tumor types for which WGS data are available will be included. The addition (8+ other) tumor types from TCGA with soon to be available DNA copy number and mRNA expression data for PARADIGM analysis will include (# cases): thyroid (500), melanoma (360), stomach (340), low grade glioma (307), prostate (261), kidney papillary (172), cervical (166), liver (152), sarcoma (114). We would like to include pancreatic cancer in the PanCAn-20+ tumor set, but there are currently only 67 from TCGA and perhaps as many as 170 currently listed in the ICGC portal (http;//dcc.icgc.org), but it remains unclear how many of these will have both DNA copy number and mRNA expression (e.g. RNAseq) data. For PARADIGM analysis, we will proceed as we did for the PanCan-12 analysis, using only the gene expression and DNA copy number data, where 3,527 tumors yielded 17,365 pathway features and 80 differentially regulated hubs (>15 downstream targets), after correction for false discovery (FDR $p<0.05$), within at least one cancer subtype for which its differential activation was most significant. To assess the pattern of regulatory hub activation across tumor types, we will discretize the difference in means; resulting rank-transformed data will bve represented in a hub x cancer type matrix including all cancers. Hierarchical clustering will then be performed and a heatmap of relative hub activity is generated, with co-variate bars showing the presence/absence of exome mutations for specific genes. Cytoscape plots will be used to illustrate the most activated or most repressed pathways. There are additional algorithms available that have previously been used to analyze TCGA data and relate PARADIGM pathway activities to exome mutations include PARADIGM-Shift and DIPSC (Differential Pathway Signature Correlation), and these can be adapted for use with WGS calls. For example, mutations in cis-regulatory elements (promoters, enhancers, super-enhancers, etc.) may be analyzed by PARADIGM-Shift to predict gain- or loss-of-function of their regulated genes. Other non-coding DNA abnormalities (e.g. lincRNAs, etc.) will be correlated with all differentially regulated PARADIGM pathway activities by DIPSC to identify potential downstream correlates in the form of pathway surrogates for these WGS abnormalities.

Finally, as non-coding analysis is one of the thrusts of the new WGS Pan-Cancer project, we will adapt our tools to both interpret and incorporate this new information. We have developed a heat-diffusion approach inspired by HotNet, called TieDIE, that identifies clustered pathways in protein-protein interaction networks from a set of mutated genes. TieDIE also incorporates multiple types of integrated data including transcriptional data to focus its search on mutated subnetworks that connected to transcription factors identified as differentially altered in tumors. A recent paper has shown how to use these heat diffusion approaches for subtyping tumors (e.g. Hofree *et al.,* 2013 *Nature Genetics*, TCGA Pan-Cancer Project). We will make use of the non-negative matrix factorization approach described in that work together with our novel multi-diffusion strategy to derive mutation-directed subtypes. This will add an important genomic component to a new definition of tumour biology.

Legacy plans

Our PARADIGM code is currently available on the Firehose platform that is run as part of the TCGA central pipeline. We are aware that ICGC may be setting up a project-wide compute platform that may also be compatible with this set up (e.g. GenomeBridge / FireBridge run from the Broad Institute). If this is made available, we will work to port our routines to that new environment but realize this is dependent on a cloud provider. Therefore we will also make our code available on github along with testing examples and tutorial documentation as is our standard practice. Internally we support a Galaxy-based compute workflow system and are aware that the OICR set up run by Lincoln Stein may also be compatible. We will work with Lincoln's group to make sure we can port our code to OICR's environment as well, which we anticipate will be more straightforward than the work we've already performed to establish Firehose-based methods.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Christopher C. Benz, M.D.	Professor & Program Director
eRA COMMONS USER NAME (credential, e.g., agency login) CHRISBENZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of California, Los Angeles	B.S.	1968	Biochemistry
University of Michigan, Ann Arbor	M.D.	1972	Medicine
VGH/CCABC & UBC, Vancouver, B.C., Canada	Residency	1972-1978	Int.Med., Heme-Onc
Yale Univ. School of Med., New Haven, CT	Subspecialty	1978-1979	Oncology

Personal Statement.

For 30 years my translational research efforts have focused on identifying molecular strategies to improve breast cancer diagnostics and therapeutics. I have published over 250 peer-reviewed manuscripts and serve on multiple national and international review and oversight committees, including the National Cancer Institute's DTP/DCTD Biological Resources Branch Oversight Committee, the American Association of Cancer Research's Task Force on Cancer and Aging, and the national Steering Committee for the NCI/NHGRI-funded The Cancer Genome Atlas (TCGA) program. For over 12 years I have also led an interdisciplinary Buck Institute team of cancer scientists working side-by-side with protein chemists and mass spectroscopists to functionally dissect endogenous breast cancer pathways involving estrogen receptor (ER), p53, and ErbB2/HER2. I continue to care for breast cancer patients at the UCSF Carol Franc Buck Breast Care Center; and since 1993 I have been a UCSF Breast SPORE (P50) project investigator bringing a new class of targeted nanoparticles into the clinic. For the past 4 years I have co-led (with David Haussler and Joshua Stuart) the Buck-UC Santa Cruz Genome Data Analysis Center for the TCGA.

Positions and Honors.

1979-1980 Postdoctoral Associate, Oncology-Pharmacology, Yale University School of Medicine
1981-1982 Instructor, Department of Medicine, Yale University School of Medicine
1982-1983 Assistant Professor of Medicine, Yale University School of Medicine
1983-1988 Assistant Professor of Medicine (in residence), University of California, San Francisco
1984-1987 Director, U.C.S.F. Hormone Receptor Laboratory
1988-1994 Associate Professor of Medicine, U.C.S.F.
1994- Professor of Medicine in Residence, U.C.S.F. (Adjunct Professor, 9/2000-present)
1994-1995 Acting Director, U.C.S.F. Cancer Research Institute
1995- Member, Joint UCSF-UCB Graduate Group in Bioengineering
1997-1998 Visiting Scientist/Professor of Mol. Medicine, Univ. Basel & Friedrich Miescher Inst., Basel CH
2000- Director; Cancer & Developmental Therapeutics Program; Buck Institute for Age Research

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Stuart, Joshua M	Professor
eRA COMMONS USER NAME (credential, e.g., agency login) Joshstuart	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Colorado, Boulder, CO	B.A.	12/96	Molecular Biology
University of Colorado, Boulder, CO	B.S.	12/96	Computer Science
University of Colorado, Boulder, CO	R.A.	08/98	Chaos Theory
Stanford University, Stanford, CA	Ph.D.	01/2004	Biomedical Informatics

A. Personal Statement

Dr. Stuart has an expertise in developing computational models to integrate multiple sources of information and a background in machine-learning applied to high-throughput datasets. He has recently developed pathway-based models to integrate multiple sources of gene activity to predict alterations and clinical outcomes in tumor samples. He co-directs the UCSC-Buck institute genome data analysis center, is a co-leader of the pan-cancer analysis working group for the Cancer Genome Atlas project, and leads the pathway analysis for a prostate cancer Stand Up To Cancer Dream Team.

B. Positions and Honors

Positions and Employment

1993-1996 Laboratory Research Assistant, Dept. of Molecular Biology (Dr. G. Stormo), University of Colorado, Boulder, Colorado.
1994 University of Colorado Health Science Cancer Fellowship, Boulder Colorado, Summer.
1996-1997 Research Assistant, Dept. of Computer Science (Dr. L. Bradley), UC, Boulder.
2000 Teaching Assistant in Biomedical Informatics, Stanford, University, Stanford, CA.
2003-2009 Assistant Professor, Dept. of Biomolecular Engineering, University of California, Santa Cruz.
2009-*present* Associate Professor, Dept. of Biomolecular Engineering, University of California, Santa Cruz.

Honors

2013-*present* Jack Baskin Endowed Chair, UCSC School of Engineering.
2009-2014 NSF CAREER Award.
2006-present Alfred P. Sloan research fellowship.
2006 University of Colorado Kalpana Chawla Outstanding Recent Graduate Award
1996 *magna cum laude,* MCD Biology, University of Colorado.
1995-1996 Achievement Rewards for College Scientists (ARCS) scholarship recipient for research in Dr. G. Stormo's laboratory.

C. 15 Selected Peer-reviewed Publications

1. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. "The Cancer Genome Atlas Pan-Cancer analysis project." *Nature Genetics*. 2013 45(10). *In press*.

2. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, and Stuart JM. “Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE).” *Bioinformatics* 2013. *To appear.*
3. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, The Cancer Genome Atlas Research Network, Friend SH, Stuart JM, Liang H, Margolin AA. “Enabling Transparent and Collaborative Computational Analysis of 12 tumor types within The Cancer Genome Atlas.” *Nature Genetics*. 2013. *In press.*
4. International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group. “Computational approaches to identify functional genetic variants in cancer genomes.” *Nature Methods.* 2013 Jul 30;10(8):723-9.
5. Wong CK, Vaske CJ, Ng S, Sanborn JZ, Benz SC, Haussler D, Stuart JM. “The UCSC Interaction Browser: multidimensional data views in pathway context.” *NAR* 2013 Jul 1;41:W218-24
6. Cancer Genome Atlas Research Network. “Integrative Analysis of genomic and molecular alterations in clear cell renal cell carcinoma.” *Nature.* 2013. *In press.*
7. Cancer Genome Atlas Research Network. “Integrated genomic characterization of endometrial cancer.” *Nature*. 2013. In press.
8. Schulze CJ, Bray WM, Woerhmann MH, Stuart J, Lokey RS, Linington RG. "Function-first lead discovery: mode of action profiling of natural product libraries using image-based screening.” Chem Biol. 2013 Feb 21;20(2):285-95
9. Perou C et al. and The Cancer Genome Atlas. “Comprehensive molecular portraits of human breast tumors.” *Nature.* 2012 Oct 4;490(7418):61-70.
10. Ellis MJ *et al.* “Whole-genome analysis informs breast cancer response to aromatase inhibition.” *Nature.* 2012 Jun 10;486(7403):353-60.
11. Heiser LM, *et al.* “Subtype and pathway specific responses to anti-cancer compounds in breast cancer.” *Proc. Natl. Acad. Sci.* 2012 Feb 21;109(8):2724-9.
12. Spellman P *et al.* The Cancer Genome Atlas. “Integrated Genomic Analyses of Ovarian Carcinoma.” *Nature. 2011.* Jun 29; 474. 609-15.
13. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM, “Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.” *Bioinformatics*. 2010.
14. Woehrmann MH, Gassner NC, Bray WM, **Stuart JM***, Lokey S*. ”HALO384: a halo-based potency prediction algorithm for high-throughput detection of antimicrobial agents.” *J Biomol Screen*. 2010 Feb;15(2):196-205.
15. Stuart JM*, Segal E*, Koller D, Kim SK. A Gene Coexpression Network for Global Discovery of Conserved Genetic Modules *Science* 2003 302:249-55. No PMCID.
***** equal contributions

D. Research Support

ACTIVE

1R01CA180778-01 (Stuart) 7/1/2013 – 6/30/2018 1.0 calendar
National Institutes of Health/NIGMS $984,593
BigData: Mid-Scale:DCM: DA: ESCE: Discovering Molecular Processes and Patient Outcome Patterns in Large-Scale Cancer Genomics Datasets Using a Biomedical Evidence Graph
Create a resource of interpretive levels of data derived from next generation sequencing data deposited in the Cancer Genomics Hub (CGHub). Automated analysis pipelines will create gene- and pathway-level inferences from raw sequence reads on tumor samples. The information will be connected to phenotypic metadata to enable crowd-sourced competitions to identify best-of-breed algorithms for predicting patient outcomes.

DBI 0845783 (Stuart) 8/15/2009 – 7/31/2014 0.5 calendar
National Science Foundation $182,356
CAREER: Development of the UCSC Interaction Browser for Integrative Genomics
The goal is to create new algorithms for discovering causal genetic interactions and the UCSC Interaction Browser, an online functional genomics resource for investigating networks of gene-associated relationships.

5U24 CA143858-04 (Haussler, Stuart) 9/28/09 – 7/31/14 1.0 calendar
NCI/NIH $1,017,820
UCSC-Buck Institute Genome Data Analysis Center for the Cancer Genome Atlas Research Network
To develop and apply methods for high throughput production-ready processing and analysis of large-scale next-generation sequencing data produced by the CGA project.

SU2C-AACR-DT0812 subaward (Small, Stuart) 12/1/2013 – 9/30/2016 1.0 calendar
UCSF/SU2C/AACR/Prostate Cancer Foundation $443,332
SU2C Prostate Dream Team Translational Cancer Research Grant: Targeting Adaptive Pathways in Resistant CRPC
Project supports identifying pathways in prostate cancer underlying androgen inhibition resistant disease. Dr. Stuart's lab will develop novel algorithms and deploy a data structure called to link together findings across labs.

OVERLAP: There is no duplication of financial support or overlap of effort between projects.

PENDING

Grant number pending (Stuart) 9/1/13-8/31/18 1.0 calendar
NIH National Institutes of Health $500,000
New Integrative Pathway Analysis Methods to Predict Biomedical Outcomes
Extend machine-learning and probabilistic graphical modeling approaches developed in the field of cancer genomics to the analysis of a broad range of human and model organism datasets. Novel methods for proposing genetic perturbations using a formal computational analysis will be developed and tested for their ability to suggest pluripotent and lineage-committing factors in a neural progenitor differentiation assay. The methods developed will contribute significant theoretical advances as well as reveal common mechanisms of stem cells and tumor biology to shed light on new treatment options for cancer.

Completed

SU2C-AACR-DT0409 (Slamon, Gray) 8/1/2009 – 9/30/2012 0.5 calendar
American Association for Cancer Research $35,000
An Integrated Approach to Targeting Molecular Breast Cancer Subtypes
Stand Up To Cancer Dream Team Award: Personalizing treatment of triple negative, metastatic breast cancer. The goals of the project are to develop tools to detect pathway perturbations in triple-negative breast cancer metastatic tumor samples, develop the UCSC BioIntegrator for predicting clinical outcomes from multiple data sources on these large patient cohorts, and develop the UCSC Cancer Browser to visualize high-throughput results for these patients. This grant funds 0.5 graduate student researchers and 1 summer month salary for Dr. Stuart.

1U54 HG006097-03 subaward (Mitchison, Sorger) 9/30/2010 – 7/31/2013 0.5 calendar
National Institutes of Health / NHGRI $94,812
Harvard LINCS Project: Pharmacological Response Signatures in Disease
The aims are to connect drug response measurements with genome-wide characterizations of cancer cell lines using integrated pathway modeling.

RN1-00540-1-004 co-PI (Forsberg, Stuart) 3/1/08-5/31/13 0.5 calendar
California Institute of Regenerative Medicine 374,470
Mechanisms of Stem Cell Fate Decisions
The goal is to develop computational tools to identify genetic markers of stem cells from genome-wide expression data. The work pays for ½ of a graduate student researcher in the Stuart lab.

1110725 (Stuart) 7/1/2010 – 6/30/2013 0.5 calendar
Department of Energy $67,673
Connecting genomes to physiology and response in marine photosynthetic eukaryotes: Systems biology of the green alga Micromonas
The goal is to apply pathway analysis to one of the most abundant marine algal communities. The work should uncover mechanisms underlying robustness and sensitivity of carbon cycling to changing environmental conditions in the open ocean.

Overlap of Funded Research with Current Project

The TCGA grant, SU2C grants, and MBARI contract each support aspects of integrated pathway analysis. For TCGA, we are supported to develop the PARADIGM software, incorporate it into NCI's centralized pipeline run at the Broad Institute called Firehose, and to make the analysis available for each tumor type. For the SU2C grant, we were funded to develop methods to identify signatures based on these PARADIGM pathway inferences and to collect datasets from contributing centers for this analysis. The MBARI project supports work to translate our pathway analysis to other species for use in projects outside of cancer genomics. The current grant will fund new algorithm development specific for integrating many more types of omics and signaling data into our pathway and visualization analysis.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Christina Yau	Senior Scientist, Buck Institute for Research on Aging Asst. Adjunct Professor, UCSF
eRA COMMONS USER NAME (credential, e.g., agency login) CHRISTINAYAU	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
UC Berkeley	B.S	2000	Bioengineering
Joint Program at UC Berkeley and UCSF	Ph.D	2008	Bioengineering
UC Berkeley		2008-2009	Biostatistics
Buck Institute for Research on Aging		2008-2010	Breast Cancer Biology

Personal Statement

My graduate training was in biostatistics and bioinformatics. Since then, I have published more than 20 peer reviewed studies, the last 10 being primarily bioinformatics studies relating to breast cancer. However, for the past 4 years, I have been heavily involved in the TCGA as part of the Buck Institute-UCSC GDAC, working with Dr. Benz and collaborating closely with Drs. Haussler and Stuart and all other UCSC bioinformaticians; in this role I have contributed substantially to all TCGA Nature publications since 2012 focused on 5 other cancer types in addition to breast cancer. Most recently, I am first author (shared) on the submitted Pan-Cancer integrated analysis of 12 different cancer types using 6 different high dimensionality TCGA assay platforms.

Positions and Honors

1998-2000	Undergraduate Lab Assistant, Maboudian Group, UC Berkeley
1999-2000	Undergraduate Student Instructor, Biology 1B, UC Berkeley
2000	Graduate Student Instructor, Biology 1B, UC Berkeley
2001-2003	Graduate Student Researcher, UCSF
2003-2008	Graduate Student Researcher, Buck Institute for Research on Aging
2008	UCSF, Breast Oncology Program Scientific Retreat, Best Poster Presentation Award
2008	American Federation for Aging Research AFAR-GE Healthcare Junior Investigator Award
2008-2009	Joint Postdoctoral Research Fellow, Buck Institute and UC Berkeley
2009	AACR-Susan G. Komen for the Cure Scholar-in-Training Award, AACR 100th Annual Meeting
2009	AACR Translational Research Scholar Award, San Antonio Breast Cancer Symposium
2009-2010	Postdoctoral Research Fellow, Buck Institute for Research on Aging
2010-2012	Staff Scientist, Buck Institute for Research on Aging
2010-2012	Staff Research Associate, UCSF (Dept. Surgery)
2012-present	Senior Scientist, Buck Institute for Research on Aging
2012-present	Asst. Adjunct Professor, UCSF (Dept. Surgery)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Theodore C Goldstein TGOLDS	Research Associate, Biomolecular Engineering UC Santa Cruz

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
UC Santa Cruz	B.A	1983	Computer Science
UC Santa Cruz	M.S.	2011	Bioinformatics
UC Santa Cruz	Ph.D.	2013	Biomolecular Engineering & Bioinformatics

Personal Statement

I am a researcher on the SU2C/PCF West Coast Prostate Cancer Dream Team at UCSC, where I am developing MedBook, a social network and bioinformatic system that will unify cancer patients, doctors, clinical and academic researchers in the war against cancer. I am also Director of Research at Guttman Initiatives, a philanthropic development and design firm where I advise on the development of predictive bioinformatic systems for chronic and transitional-care patient monitoring systems. I was previously Vice President of Developer Tools at Apple, Inc. My team created the Mac OS X and iOS runtime systems and the Xcode developer tools for Apple's Intel, iPhone, and iPad products which Transformed and reignited Apple's application community and created the tools for Intel Mac, iPhone and iPad by creating powerful tools for developers and end-users which have enabled Apple to become the most valuable company in the world. Prior to Apple, I led the electronic commerce and smart card efforts at Sun Microsystems, where I created Java Card which is used for all federal id systems, ecommerce, and medical information. JavaCard unified the smart card industry, a platform with over 2 billion units deployed (in bank cards, transportation cards, and GSM and other telephone applications. I was recognized and awarded Sun Microsystems' President's award. I have ten patents in computer and information technologies. I am Currently bringing the techniques of advanced user centered design and massively scalable computing to bioinformatics, biomolecular engineering and health care informatics.

Positions and Honors

2013	Guttman Initiatives, *Director of Research*
2008-2013	Baskin School of Engineering, UC Santa Cruz *Student and Researcher*
2007	Theranos Inc., *Vice President of Software*
2002-2007	Apple, Inc. *Vice President of Development Technologies*
2001	ActivCard, Inc. *Vice President of Business and Technology*
2000-2001	Catalytic Consulting, *Consultant*
1998-2000	Brodia.com, *Vice President of Engineering & CTO*
1998	Sun Microsystems President's Award
1990-1998	Sun Microsystems Laboratories & JavaSoft, *Principal Investigator and CTO Ecommerce*
1986-1990	ParcPlace Systems / Xerox Palo Alto Research Center, *Manager*
1984-1986	CoDesign, *Engineer*
1983	The Systems Group (for Intel), *Engineer*

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm local time). Explanatory notes follow the form

Title of abstract

The interplay between non-coding regulatory mutations and the cancer epigenome

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Benjamin P. Berman & Peter W. Laird, USC Epigenome Center, TCGA

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Huy Dinh & Hui Shen, USC Epigenome Center, TCGA	

Background and preliminary data

Cis-regulatory logic is encoded by short DNA sequences controlling the binding of sequence-specific transcription factors, and the structural modification of chromatin by epigenetic modifiers. Our understanding of chromatin state architecture in the human genome has expanded dramatically in the past several years with large-scale sequencing efforts using ChIP-seq, DNase-seq, Hi-C, and other techniques. These methods have been applied primarily in the context of cell lines, whereas the bulk of patient chromatin state information comes from DNA methylation profiling. Our group and others have established that both focal chromatin states (promoters, enhancers, insulators), as well as long-range topological domains, are reflected in DNA methylation patterns in normal tissues and tumors.

It has long been known that somatic chromosomal translocations occurring in non-coding regions can drive cancer by the incongruous juxtaposition of non-coding cis-regulatory sequences adjacent to oncogenic genes. Furthermore, based on the single-nucleotide specificity of most transcription factors, it has also been hypothesized that somatic point mutations in non-coding control sequences could drive oncogenesis. Interest in this mechanism has surged since the recent reports of whole-genome sequencing to identify TERT promoter mutations affecting transcription factor binding and driving cancer. It is hypothesized that similar point mutations will affect enhancers and insulators. These regulatory mutations are mediated via changes to chromatin structure that affect DNA methylation state, and thus ICGC samples with matched WGS and DNA methylation data present a tremendous opportunity to understand how non-coding regulatory mutations drive oncogenesis, and when and how they may be targeted by epigenetic therapies.

This interplay between the genome and epigenome is not unidirectional. Just as regulatory mutations affect epigenetic state, we know that specific epigenetic states influence genome mutation rates. Regions of open chromatin can be hotspots for double stranded breaks and oncogenic translocations such as TMPRSS2:ERG, while repressive topological domains are associated with elevated point mutation rates. Identifying the types of chromatin state that are associated with specific mutagenic mechanisms will allow us to understand how the epigenomic features of the tumor cell of origin can influence the rate of driver mutations non-uniformly across the genome.

Our group has developed computational tools to identify chromatin state changes in patient samples using DNA methylation data taken directly from the sample, combined with reference chromatin annotations from databases such as ENCODE, NIH Reference Epigenome Mapping Consortium, BLUEPRINT, and Cistrome. We have preliminary data strongly suggesting that ICGC DNA methylation platforms such as the Infinium HM450 array and whole-genome bisulfite sequencing (WGBS) can be used to identify chromatin states corresponding to transcription factor binding sites, enhancers, insulators, topological domains, and late-replication domains. Using the large number of WGS cases that will be available, we expect to be able to link somatic non-coding driver mutations to specific chromatin state changes. This will allow us to understand the key intermediate link between these mutations and the transcription changes they ultimately affect. Associating these regulatory mutations with binding by specific upstream transcription factors or chromatin modifiers (using reference ChIP-seq databases), will allow the identification regulatory pathways that can be compromised by mutation of either upstream *trans* factors or cis-regulatory control sequences.

Timelines & resources dedicated to project

We can begin the inference of chromatin states from DNA methylation profiles (Aim 1) before the completion of WGS variant call (VCF) files are available. We will incorporate any available DNA methylation datasets during the analysis timeframe. We are currently aware of large numbers of cases on the Illumina Infinium platforms, as well as whole-genome bisulfite sequence (WGBS) samples in various stages of completion at TCGA, DKFZ (Germany) and CNAG (Spain). In addition, a number of other ICGC project descriptions list plans for methylation profiling: using WGBS for a subset of HER2 Breast cases (Sanger Institute), and using an unspecified platform for Malignant Lymphoma (DKFZ), Prostate (Fondation Synergie/CNAG), and Liver (RIKEN). We will work with whichever DNA methylation platforms are used in WGS-matched samples. However, some platforms without single-base resolution (MRE-seq, meDIP-seq) will be left out of certain analyses in Aim 3 that require it.

The timelines for Aims 2-4 will be heavily dependent on the core variant calling of WGS data by the Consortium. We will begin working on analysis methods as soon as preliminary VCF files become available, but we will need significant numbers of samples to make statistically significant associations.

Research proposal

Aim 1: Identify inferred chromatin states within individual ICGC samples by combining DNA methylation data from the samples themselves with reference epigenome data from ENCODE, REMC, and Cistrome. Specifically, we will identify inferred promoters, enhancers, insulators, transcription factor binding sites, and early- and late-replicating domains. We will identify those chromatin states that are recurrent and unique to specific subsets of tumor samples. This will allow testing for association with non-coding mutations in Aim 2.

Aim 2: Use the chromatin states identified in Aim 1 to test associations to various types of non-coding somatic mutation. (2A) Identify point mutations associated with a localized chromatin state (100-10,000 bp). The size here is dictated by the size of focal chromatin states such as promoters, enhancers (and super-enhancers), insulators. (2B) Identify point mutations associated with long-range chromatin domains (100kb-10Mb). This can be the result of disrupting a chromatin boundary element, which may allow the spreading of a chromatin state from one or the other side of the boundary. (2C) Identify balanced translocations associated with changes to long-range chromatin domains on one or both sides of the breakpoints.

Aim 3: Untangle the effects of cancer-specific regulatory mutations from the general rules of the sequence/methylation in normal tissues. It is known that local sequence polymorphisms can have a strong effect on methylation in normal tissues (a trivial but important case is the large number of CpG->CpA SNPs in the genome, where the A allele abolishes methylation present at the adjacent cytosine). Using the matched normal samples from ICGC groups, we will identify these non-coding “meQTLs”. For HM450 data, we will search for association between homozygous SNPs across individuals. We will confirm using WGBS samples, where allele-specific linkages can be established within an individual sample at heterozygous SNPs. We will use our Bis-SNP tool, which identifies SNPs for allele-specific methylation analysis.

Aim 4: Identify chromatin states in normal cells that are associated with mutations in the tumor. Using the chromatin states inferred from matched normal tissues in Aim 1, we will investigate the type and frequency of somatic mutations within each particular state. We will focus on those chromatin states that are variable across individuals, and thus able to most strongly suggest causality. Point mutations will be stratified by dinucleotide contexts to indicate relationships with particular mutagen and/or particular DNA repair deficiencies. Structural breakpoints will be stratified by presence of microhomology to indicate likely replication-based vs. double-stranded break processes.

Legacy plans

All code under development will be available via a public code repository such as GitHub. Release versions of analysis pipelines will be sufficiently documented to allow replication by third parties. This will include vignettes/tutorials of any analyses performed for Consortium publications. Executables will also be made available, along with instructions for installing them along with any necessary 3rd party libraries. A test input dataset will be made available, along with documentation of the commands required to begin an analysis run, and the expected output files. If a centralized workflow execution engine such as FireHose is implemented by the Consortium, we will provide wrappers for our code and/or virtual machines when feasible.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Benjamin P. Berman, Ph.D.	Assistant Professor of Bioinformatics & Preventive Medicine Director of Sequencing Informatics, USC Epigenome Center
eRA COMMONS USER NAME (credential, e.g., agency login) BENBERMAN	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of California, Berkeley	B.A.	05/96	Computer Science
University of California, Berkeley	Ph.D.	12/06	Molecular & Cell Biology (bioinformatics track)

Summary research statement

Dr. Berman is an expert in cancer epigenetics, non-coding regulatory sequence annotation, and next-generation sequencing. As a doctoral student with Dr. Michael Eisen and Dr. Gerry Rubin, he published the first demonstration that distal enhancer elements could be identified genome-wide using a purely computational approach (Berman et al. ***PNAS*** 99,757; 2002), as well as a complete map of gene expression in *Drosophila* development (***Genome Biology***; 2007). Since 2008, he has led the development of next-generation sequence analysis pipelines for The USC Epigenome Center and the Norris Comprehensive Cancer Center, which are used to analyze close to a thousand samples a year from more than 50 different labs. He led an effort to sequence the first complete, single-basepair methylome of a primary colorectal tumor, discovering that cancer hypomethylation was focused within long domains of late replication coinciding with topological domains attached to the nuclear lamina (***Nature Genetics*** 44,40; 2012). He is a member of the The Cancer Genome Atlas (TCGA) epigenome data production group at USC, and has participated in the analysis of a number of TCGA cancer projects (***Cancer Cell*** 17,510; 2010, ***Nature*** 487,330; 2012, ***Nature*** 490,61; 2012, ***Nature*** 499:43; 2013, ***Nature Genetics*** 45:1113; 2013). His group developed an open-source toolkit for bisulfite sequence analysis (***Genome Biology*** 13,R61; 2012), and is currently developing novel methods for an intergrative analysis of dozens of complete methylomes being sequenced by his group. In 2012, he co-developed a novel whole-genome sequencing method, NOMe-seq, to detect nucleosome occupancy and DNA methylation patterns within individual molecules (***Genome Research*** 22,2497; 2012). NOMe-seq was recently named one of the top 10 innovations of 2013 (***The Scientist*** 27,38394; 2013). Dr. Berman is a member of the USC High Performance Computing Faculty Advisory Committee and the USC/Norris Cancer Center Bioinformatics Committee, and is an awardee of the 2014 STOP CANCER Research Career Development Award.

Positions and Employment

1996-1997 Software Engineer, Apple Computer, Inc.
1998-2000 Bioinformatics Software Engineer, Lawrence Berkeley National Laboratory
2001-2006 PhD Research Assistant, UC Berkeley and Howard Hughes Medical Institute (Co-advisors: Gerald M. Rubin and Michael B. Eisen)
2007-2007 Postdoctoral Fellow, University of Southern California (advisors: Chris Haiman& Gerry Coetzee)
2008-2011 Senior Research Associate, USC Epigenome Center, Keck School of Medicine of USC
2011-present Assistant Professor, Bioinformatics Division, Department of Preventive Medicine, USC

Honors and Awards

2001-2003 NIH/Berkeley Genomics Pre-doctoral Trainee
2007-2008 NIH Molecular Epidemiology Post-doctoral Trainee
2010-2014 Forbeck Scholar Award, William Guy Forbeck Research Foundation
2013-2014 American Cancer Society Junior Faculty Research Award
2013-2016 STOP CANCER Research Career Development Award

Program Director/Principal Investigator (Last, First, Middle):

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Peter W. Laird	Professor of Surgery, Biochem & Molecular Biology Director, USC Epigenome Center
eRA COMMONS USER NAME (credential, e.g., agency login) PETER_W_LAIRD	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Leiden, The Netherlands	B.Sc.	12/82	Biology & Chemistry
University of Leiden, The Netherlands	M.Sc.	03/84	Molecular Biology
Netherlands Cancer Institute / University of Amsterdam (with Dr. Piet Borst)	Ph.D.	03/88	Molecular Biology

Summary Research Statement

Dr. Peter W. Laird is an internationally recognized expert in cancer epigenetics and in DNA methylation analysis technology. As a postdoctoral fellow with Dr. Rudolf Jaenisch, he published the first demonstration of a causal role for DNA methylation in oncogenesis, using a mouse model (Laird et al., ***Cell*** 81, 197,1995). This work was cited as a "milestone" in cancer by Nature magazine (***Nature Cancer Milestones***, April, 2006). He has invented several DNA methylation analysis techniques, including COBRA (1997) and MethyLight (1999). His recent work includes the development of a new mouse model for invasive colorectal cancer (***Cancer Research*** 66:8430, 2006), the report of a very strong association between DNA methylation and *BRAF* mutation in colorectal cancer (***Nature Genetics*** 38,787; 2006), the discovery that embryonic stem cell Polycomb repressor targets are predisposed to abnormal DNA methylation in cancer (***Nature Genetics*** 39,157; 2007, ***Genome Research*** 22,271; 2012), the discovery of a novel epigenetic subtype of glioblastoma (G-CIMP), associated with *IDH1* mutation and better survival ***Cancer Cell*** 17,510; 2010, and the single-basepair resolution analysis of the entire methylome of primary colorectal cancer, leading to the discovery that focal CpG island hypermethylation is concentrated within long regions of hypomethylation that coincide with late replication and attachment to the nuclear lamina (***Nature Genetics*** 44,40; 2012). He also oversees all epigenomic data production for The Cancer Genome Atlas (TCGA) (***Nature*** 455,1061; 2008, ***Nature*** 474,609; 2011, ***Nature*** 487,330; 2012, ***Nature*** 489,519; 2012, ***Nature*** 490,61; 2012, ***Nature*** 497,67; 2013, ***Nature*** 499:43; 2013, ***NEJM*** 368:2059, 2013; ***Nature Genetics*** 45:1113; 2013, ***Nature Genetics*** 45:1134; 2013, ***Cell*** 155:462; 2013). He was appointed Founding Director of the USC Epigenome Center in 2007. He has been Co-Leader of the Epigenetics and Regulation Program of the USC/Norris Comprehensive Cancer Center since 2004.

Positions and Employment

1988-1991 Postdoctoral Fellow, The Netherlands Cancer Institute, Amsterdam, The Netherlands, with Dr. Anton Berns

1991-1996 Postdoctoral Fellow, The Whitehead Institute for Biomedical Research/Massachusetts Institute of Technology, Cambridge, MA, with Dr. Rudolf Jaenisch

1996-2002 Assistant Professor of Surgery and of Biochemistry & Molecular Biology, Keck School of Medicine USC, Los Angeles, CA

2002-2012 Associate Professor of Surgery and of Biochemistry & Molecular Biology, Keck School of Medicine USC, Los Angeles, CA

2012-Present Professor of Surgery and of Biochemistry & Molecular Biology, Keck School of Medicine USC, Los Angeles, CA

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identification of mechanisms of structural variation and copy number alterations in cancer whole genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Rameen Beroukhim, Matthew Meyerson

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Jeremiah Wala, Marcin Imielinski	

Background and preliminary data

Copy number changes and structural variations are major mutational forces in cancer and drive tumorigenisis through large-scale changes to gene copy number and by the creation of oncogenic fusion genes. These structural variations arise from varied rearrangement mechanisms, including non-homologous end-joining, microhomology mediated breakpoint repair, and non-allelic homologous recombination, and with each process governed by the local sequence context at the breakpoints. A complete understanding of the selection pressures that give rise to copy number alterations in cancer requires identifying both the genes affected by these rearrangements and the underlying physical events that gave rise to the alterations. Previous approaches to breakpoint detection in cancer have only begun to attempt to define the underlying rearrangement mechanisms, and to our knowledge no pan-cancer analysis of breakpoint mechanisms and motifs has been undertaken.

In our initial approach, we identified a large number of copy-number breakpoints that do not have an associated rearrangement signature as analyzed by current methods. These breakpoints tend to lie in or near highly redundant regions of the genome, making their detection difficult with standard approaches that rely on signatures of discordant paired-end reads. The low-sensitivity of current rearrangement detection methods at regions of the genome with poor mappability would bias analyses towards the detection of mechanisms that occur at unique and complex regions of the genome. To increase detection sensitivity across the genome, we employ an integrated approach that combines copy level changes with whole genome sequencing reads to enhance the rearrangement signal at these breakpoints. By increasing the number of detected rearrangements, we aim to provide a comprehensive and unbiased landscape of the mechanisms of structural variation across a large cohort of whole-genome sequenced tumor normal pairs, and to use this data to inform our analysis of the significance of copy-number changes.

Timelines & resources dedicated to project

Timeline: Initial detection of rearrangement breakpoints in 2000 genomes by Aug 2014. Characterization of breakpoint motifs and rearrangement mechanisms by Oct 2014. Significance analysis of copy number changes by Dec 2014. Manuscript preparation / submission in May 2015.

Resources: tumor and normal whole genome read data, copy number segmentation, jumping library data for rearrangement detection validation

Research proposal

We will perform breakpoint detection and motif identification across the 2000+ ICGC/TCGA whole genome tumor-normal pairs. We will use the breakpoint analysis to identify the size and distribution of copy number events and determine the significance of the genes involved in these events for driving tumor selection. To accomplish these, we propose the following three-phase approach:

1) Genome-wide identification of rearrangement events.
Copy-number breakpoints derived from tumor/normal read count ratios provide a strong prior on the locations of rearrangement breakpoints. Additionally, the relative copy levels at these breakpoints indicate the likely orientation of the rearrangement event. We have an initial pipeline that integrates this copy-level information with sequencing reads to identify the most likely rearrangement partners for these copy number breakpoints. The validation set for this approach will come from our 3kb insert-size jumping library data for two breast cancer cell lines, which provides an orthogonal and sensitive approach for identifying rearrangements. We will then scale the detection pipeline to perform unbiased genome-wide detection of rearrangement events.

2) Identification of recurrent sequence motifs at breakpoints and correlations analysis between rearrangement mechanisms and tumor-type and genomic alterations.
Varied methods exist for analyzing sequence motifs across genome loci, and we will use our breakpoint calls from Aim 1 along with available motif detection software to characterize the sequence context around rearrangement breakpoints. These sequence motifs will be used to infer the underlying mechanism behind the rearrangement, with the aim of providing a comprehensive landscape of rearrangement forces across the 2000 cancer genomes. Additionally, we will correlate the breakpoint motifs with tumor type, and identify genomic alterations (e.g. mutations, copy level changes) that predispose to certain classes of rearrangement events.

3) **Identification of recurrent and functionally significant copy number alterations**
We have also developed techniques to assess the significance of somatic copy number changes from absolute allelic copy levels in order to identify tumor suppressors and oncogenes. We will combine these methods with our rearrangement detection pipeline to identify regions in which the rate of copy number alterations vary from predicted rates, indicating positive or negative selection. We will also further develop novel techniques that control for confounders and uncover correlations between genetic events (implying functional relationships) and with patient data, with the ultimate goal of improving prognostics and informing treatment.

These analyses will contribute to our still rudimentary knowledge of the physical mechanisms that shape the cancer genome, and obtaining a better understanding of the mechanisms of copy-number alterations in cancer will inform our significance analyses aimed at identifying recurrently amplified or deleted driver genes. The extension of our rearrangement and copy level analyses to whole-genome sequencing data will also allow us to investigate recurrent alterations in non-coding regions of the genome, including trans-regulatory and enhancer elements, and identify focal events that are below the detection threshold of SNP-array based copy number pipelines.

Legacy plans

We will provide our rearrangement detection software, genome-wide rearrangement annotations, sequence motif definitions and genome-wide copy number significance profiles to the community.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Beroukhim, Rameen	Assistant Professor of Medicine
eRA COMMONS USER NAME (credential, e.g., agency login) rberoukhim01	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
University of California, Berkeley	A.B.	1991	Physics and Philosophy
University of Cambridge, England	M.Phil.	1992	Molecular Biology
University of Cambridge, England	Ph.D.	1996	Molecular Biology
University of California, San Francisco	M.D.	2000	Medicine

Positions and Employment

2000-2001 Intern, Internal Medicine, University of California, San Francisco, CA
2000-2002 Fellow, Molecular Medicine, University of California, San Francisco, CA
2001-2002 Junior Assistant Resident, Internal Medicine, University of California, San Francisco, CA
2002-2006 Clinical Fellow, Department of Medicine, Harvard Medical School, Boston, MA
2002-2006 Clinical Fellow, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2005-2009 Visiting Postdoctoral Scientist, Cancer Program, Broad Institute, Cambridge, MA
2006- Instructor in Medicine, Harvard Medical School, Boston, MA
2006- Instructor, Department of Medical Oncology, Dana-Farber Cancer Institute (DFCI), Boston, MA
2006- Attending Staff Oncologist, Department of Medical Oncology, DFCI, Boston, MA
2006- Associate Physician, Internal Medicine, Brigham and Women's Hospital, Boston, MA
2009-2010 Instructor, Department of Cancer Biology, DFCI, Boston, MA
2009- Consultant, Novartis Institutes for Biomedical Research
2009- Associate Member, Broad Institute, Cambridge, MA
2010- Assistant Professor of Medicine, DFCI and Harvard Medical School, Boston, MA

Honors

1990 Phi Beta Kappa
1991 High Honors in Physics, University of California, Berkeley, CA
1991 Winston Churchill Scholarship, Winston Churchill Foundation
1992 Glaxo Dorothy Hodgkin Scholarship
1992 Overseas Research Studentship (British Government award
1996 Max Perutz Prize
2004 Postdoctoral Traineeship Award
2006 DF/HCC Prostate SPORE Career Development Award
2007 Physician Research Training Award
2009 V Foundation Scholar
2012 Sontag Scholar

Selected Peer-reviewed Publications (Selected from 96 peer-reviewed publications)

1. **Beroukhim, R.**, et al (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences of the United States of America, 104(50), 20007-20012. PMCID: PMC2148413
2. **Beroukhim, R.**, et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899-905. PMCID: PMC2826709
3. Mermel CH, Schumacher SE, Hill B, Meyerson ML, **Beroukhim R**[1], Getz G[1]. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMCID: PMC3218867
4. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, **Beroukhim R**. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013; 45:1134-40. NIHMSID: 517488.

Curriculum vitae for Matthew Meyerson, M.D., Ph.D.

Education and Training

1985	A.B., Chemistry and Physics, Harvard College
1993	M.D., Harvard Medical School
1994	Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996	Resident, Clinical Pathology, Massachusetts General Hospital
1995-1998	Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

1998-2005	Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006	Associate Member, Broad Institute of Harvard and MIT
2005-	Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
2005-2009	Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006-	Senior Associate Member, Broad Institute of Harvard and MIT
2009-	Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

Awards and Honors

1999	Pew Scholar in the Biomedical Sciences
2004	Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005	Clinical Investigator Award, American Lung Association
2009	Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010	Team Science Award, American Association for Cancer Research
2011	Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012	Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …, Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-13795.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Wala, Jeremiah	Graduate Student, Harvard University
eRA COMMONS USER NAME (credential, e.g., agency login) jeremiahwala	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Cornell University	B.S.	2009	Engineering Physics
Cornell University	M.S.	2010	Applied Physics
Harvard University	M.D.	2018 (exp)	
Harvard University	Ph.D.	2016 (exp)	Cancer Genomics

Research Experience

2006-2009 Electrospinning and microfluidics, Cornell University, Ithaca, NY
2009-2010 Medical computer vision (thoracic CT), Cornell University, Ithaca, NY
2011-2012 Radiotherapy treatment planning optimization, Massachusetts General Hospital, Boston, MA
2012- Cancer genomics, Broad Institute, Cambridge, MA

Honors

2005 Cornell Tradition Fellowship, Cornell University
2009 *Magna cum laude*, Cornell University
2009 Hartman Prize in Experimental Physics, Cornell University
2010 Cuykendall Memorial Teaching Award, Cornell University

Peer-reviewed Publications

1. Craft D, McQuaid D, **Wala J**, Chen W, Salari T, Bortfeld T. Multicriteria VMAT optimization. Medical Physics. 2012; 57(17), 686-696. PMID: 22320778
2. Salari E, **Wala J**, Craft D. Exploring trade-offs between VMAT dose quality and delivery efficieny using a network optimization approach. Physics in Medicine and Biology. 2012; 57(17), 5587-5600. PMID:
3. **Wala J**, E Salari, W Chen, D Craft. Optimal partial-arcs in VMAT treatment planning. 2012. Physics in Medicine and Biology. 2012; 57:5861-5874
4. **Wala J**, D Craft, J Paly, A Zietman, J Efstathiou. Maximizing dosimetric benefits of IMRT in the treatment of localized prostate cancer through multicriteria optimization planning. 2013. Medical Dosimetry. 2013; 38(3):298-303.
5. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, **Wala J**, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013; 45:1134-40. NIHMSID: 517488.
6. Berger A, Imielinski M, Duke F, **Wala J**, Kaplan N, Shi G, Andres D, Meyerson M. Oncogenic RIT1 mutations in lung adenocarcinoma. 2014. Oncogene. In Press

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Imielinski, Marcin	Research Fellow in Pathology
EMAIL ADDRESS marcin@broadinstitute.org	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Rutgers College	B.S., B.A.	5/00	Computer Science, Biological Sciences
University of Pennsylvania School of Medicine (UPenn)	M.D., Ph.D.	5/08	Medicine, Genomics and Computational Biology
Massachusetts General Hospital (MGH), Harvard Medical School (HMS)	Resident	6/11	Clinical Pathology
Brigham and Women's Hospital (BWH), MGH, HMS	Fellow	6/12	Molecular Genetic Pathology

A. Personal Statement

I am an M.D. with clinical training in molecular genetic pathology and a Ph.D. computational biologist with broad experience in genomics and systems biology. I'm fascinated by the potential of integrated 'omics and big data analytics to transform cancer medicine and reveal fundamental features of tumor biology.

B. Positions and Honors

2000-2008 M.D. / Ph.D. Student, Genomics and Computational Biology Program, UPenn
2007-2010 Research Associate, Center for Applied Genomics, Children's Hospital of Philadelphia
2008-2011 Resident in Pathology, MGH / HMS
2011-2012 Clinical Fellow in Molecular Genetic Pathology, BWH / MGH / HMS
2010-Present Postdoctoral fellow in Dr. Matthew Meyerson lab, DFCI / Broad Institute
2012-Present Research Fellow, Department of Pathology, MGH / HMS

Honors: National Merit Scholar (1995), Presidential Scholar, Rutgers College (1995), Henry Rutgers Scholar, Rutgers College (1999), Best Student Poster Award, 5th International Conference for Systems Biology, Heidelberg, Germany (2004), BioAdvance Fellowship in Bioinformatics (2004), Best Paper Award in Session, 26th American Control Conference, New York, NY (2007), Trainee Research Award, American Society for Human Genetics (2009), Best Abstract, Pathology, MGH Clinical Research Day (2010), Most downloaded article in July 2010 for journal "Chaos" (2010), AACR Scholar-in-Training Award (2012), Best Poster in Anatomic Pathology, HMS Pathology Retreat (2013), Top 5 Abstract, Dana-Farber / Harvard Cancer Center Lung Cancer Research Symposium (2013)

C. Selected Peer-reviewed Publications (Selected from 36 peer-reviewed publications)

1. Imielinski, M. et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. *J Clin Invest* (2013), in press.
2. Berger, A.H. et al. Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* (2013), in press.
3. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150, 1107–1120 (2012).
4. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* 150, 251–263 (2012).
5. Imielinski, M. et al. Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol Cell Proteomics* 11, M111.014910 (2012)
6. Imielinski, M. & Belta, C. Deep epistasis in human metabolism. *Chaos* 20, 026104 (2010).
7. Imielinski, M. et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* 41, 1335–1340 (2009).

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

An integrated nexus of >15,000 genome sequences and analysis tools facilitates more efficient cancer somatic driver gene discovery

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Eric Boerwinkle1,2, Richard Gibbs1

1Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 2Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
David Wheeler Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX	

Background and preliminary data

In order to increase sample size and control costs, there is growing enthusiasm for sequencing more cancer (i.e. tumor) samples without sequencing surrounding normal tissue or circulating lymphocytes from the same individual. Instead, one would analyze the spectrum of variants in the cancer samples and compare this to databases of DNA sequences or variant sites from presumed non-cancerous samples, as well as the very large and growing collection of variant sites from tumors. To support such practices large reference samples will be needed from ethnically diverse populations with deep phenotype data, including cancer history. Careful analysis of large cancer genome sequence collections must be compared with the mutational 'background' in these reference samples. We have created a research commons based upon the analysis of multiple, large, deeply phenotyped (including cancer), longitudinal cohort studies belonging to the CHARGE consortium (N=~50,000 study participants). The current sample consists of 12,000 individuals with whole exome sequencing (WES) and 3,600 individuals with whole genome sequencing. Additional sequencing is ongoing. We are currently analyzing loss of function and predicted damaging variants in 300 tumor/normal pairs with genome sequence data and contrasting the results with a comparison of the 300 tumor sequences with the whole exome sequence data from 12,000 individuals. Such comparison takes into account the population-specific site frequency spectrum as well as local genome admixture characteristics.

Timelines & resources dedicated to project

Research proposal

We propose to use these approaches to compare the mutational profiles of the TCGA/ICGC collection of 2,000 WGS samples with our growing resource of reference samples. The comparison will consider mutational spectra and functional class. *In silico* comparison will be followed up, where needed, with experimental validation. Formal analysis of this contrast will argue for or against the ability of a cost-saving approach where only cancer samples are sequenced. A cloud computing resource would facilitate such an effort in two ways: 1) Computational efficiency, and 2) Data sharing among collaborators that would compare "their" cancer to this central database of normal genomes/exomes.

Legacy plans

Any software, algorithms, visualization tools, statistical analyses derived in this work will be made freely available to the research community.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Boerwinkle, Eric	Professor and Director Division of Epidemiology; Director Human Genetics Center
eRA COMMONS USER NAME (credential, e.g., agency login) **EBOERWINKLE**	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Cincinnati, OH	B.S.	05/80	Biology
University of Michigan, Ann Arbor, MI	M.A.	05/84	Statistics
University of Michigan, Ann Arbor, MI	M.S.	05/85	Human Genetics
University of Michigan, Ann Arbor, MI	Ph.D.	05/85	Human Genetics

A. Positions and Honors

Positions and Employment

1996-present	Professor, Human Genetics Center, School of Public Health, UTHSC-Houston
1996-present	Center Director and Professor, Institute of Molecular Medicine, University of Texas
1997-present	Director, Human Genetics Center, School of Public Health, UTHSC-Houston
2003-present	Director, Division of Epidemiology, School of Public Health, UTHSC-Houston
2004-present	Kozmetsky Family Chair in Human Genetics, Institute of Molecular Medicine, UTHSC-Houston

Honors

1991-1996	Research Career Development Award from the National Institutes of Health
1991-1996	Established Investigatorship of the American Heart Association
1999-2010	MERIT Award, National Institutes of Health
2003	President's Scholar Award
2004	Kozmetsky Family Chair in Human Genetics
2005	Ancel Keys Lecture and Award, American Heart Association

B. Selected Peer-reviewed Publication (Selected from 653 peer-reviewed publications)

Most relevant to the current application

1. Ellinor P.T., Lunetta K.L., Albert C.M., Glazer N.L., Ritchie M.D., Smith A.V., Arking D.E., Müller-Nurasyid M., Krijthe B.P., Lubitz S.A., Bis J.C., Chung M.K., Dörr M., Ozaki K., Roberts J.D., Smith J.G., Pfeufer A., Sinner M.F., Lohman K., Ding J., Smith N.L., Smith J.D., Rienstra M., Rice K.M., Van Wagoner D.R., Magnani J.W., Wakili R., Clauss S., Rotter J.I., Steinbeck G., Launer L.J., Davies R.W., Borkovich M., Harris T.B., Lin H., Völker U., Völzke H., Milan D.J., Hofman A., Boerwinkle E., et al. (2012) Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nat Genet 44(6):670-675. PMCID: PMC3366038
2. Liu X., Jian X., Boerwinkle E. (2011) dbNSFP: A lightweight database of human non-synonymous SNPs and their functional predictions. Hum Mutat 32(8):894-899. PMCID: PMC3145015 [Available on 2012/8/1]
3. Bamshad M.J., Shendure J.A., Valle D., Hamosh A., Lupski J.R., Gibbs R.A., Boerwinkle E., Lifton R.P., Gerstein M., Gunel M., Mane S., Nickerson D.A.; on behalf of the Centers for Mendelian Genomics. (2012) The Centers for Mendelian Genomics: A new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am J Med Genet A 2012 May 24. doi: 10.1002/ajmg.a.35470. [Epub ahead of print]
4. Chanda P., Yuhki N., Li M., Bader J.S., Hartz A., Boerwinkle E., Kao W.L., Arking D.E. (2012) Comprehensive evaluation of imputation performance in African Americans. J Hum Genet 2012 May 31. doi: 10.1038/jhg.2012.43. [Epub ahead of print]

stigator

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Richard A. Gibbs, Ph.D.	Director, Human Genome Sequencing Center Wofford Cain Professor Department of Molecular and Human Genetics

EDUCATION/TRAINING (*Begin with baccalaureate or other initial professional education, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
University of Melbourne, Australia	B.Sc. (Hons)	1979	Genetics
University of Melbourne, Australia	Ph.D.	1986	Molecular Genetics

A. Positions and Honors.

1995 - Present Director, BCM Human Genome Sequencing Center, Houston, TX
1998 - Present Professor, Molecular & Human Genetics, BCM, Houston, TX
2000 - Present Wofford Cain Professor of Molecular and Human Genetics

Honors

2008 Texas Genetics Society, Geneticist of the Year
2011 Elected member ASHG Board
2011 Elected Member Institute of Medicine

B. Selected peer-reviewed publications, in chronological order (earliest to recent).

1. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio, D.D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., et. al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.*, **362**,1181-1191, 2010. PMID: 20220177
2. English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS.One.*, **7**, e47768-2012. PMCID: PMC3504050
3. Wang, Y., Lu, J., Yu, J., **Gibbs, R. A.,** and Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.*, 2013. PMID: 23296920
4. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, Bekheirnia MR, Leduc MS, Kirby A, Pham P, Scull J, Wang M, Ding Y, Plon SE, Lupski JR, Beaudet AL, Gibbs RA, Eng CM. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013 369:1502-11
5. Lupski, J. R., Gonzaga-Jauregui, C., Yang, Y., Bainbridge, M. N., Jhangiani, S., Buhay, C. J., Kovar, C. L., Wang, M., Hawes, A. C., Reid, J. G., Eng, C., Muzny, D. M., and **Gibbs, R. A.** Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. *Genome Med.*, **5**, 57-2013. PMID: 23806086; PMCID: PMC3706849
6. Morrison, A. C., Voorman, A., Johnson, A. D., Liu, X., Yu, J., Li, A., Muzny, D., Yu, F., Rice, K., Zhu, C., Bis, J., Heiss, G., O'Donnell, C. J., Psaty, B. M., Cupples, L. A., **Gibbs, R.,** and Boerwinkle, E. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat.Genet.*, 2013. PMID: 23770607
7. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499, 43-49, 2013. PMID: 23792563

BIOGRAPHICAL SKETCH

NAME David A. Wheeler, Ph.D.	POSITION TITLE Associate Professor of Molecular and Human Genetics
eRA COMMONS USER NAME: WHEELER	

EDUCATION/TRAINING (*Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Maryland, College Park, MD	B.S.	1972	Biochemistry/Zoology
The George Washington University, Washington, DC	M.S.	1976	Biochemistry
The George Washington University, Washington, DC	Ph.D.	1983	Genetics

A. Personal Statement

Dr. Wheeler develops methods for discovery of genome variation in human and animal populations using DNA sequencing technologies with the goal of relating polymorphism to human disease, especially cancer. His work in this area involves large-scale multi-center national and international projects such as TCGA and ICGC and other cancer sequencing projects that aim to comprehensively catalogue all mutations leading to cancer. Dr. Wheeler is a recognized expert in mutation discovery and analysis in the cancer genome.

B. Positions and Employment

2004-2006	Co-Director for Bioinformatics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX
2004-2013	Associate Professor, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
2006-present	Director, Cancer Genomics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX
2010-present	Assistant Director, Human Genome Sequencing Center, Baylor College of Medicine, Houston TX**.**
2013-present	Professor, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

C. Selected Peer-reviewed Publications

1. **Wheeler DA**, Srinivasan M, Egholm M, Shen Y, Chen L et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-6. PMID: 18421352
2. Ding L, Getz G, **Wheeler DA**, Mardis ER, McLellan MD et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455:1069-1075. PMCID: PMC2694412
3. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, **Wheeler DA**, Gibbs RA, Yu F (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20:273-80. Epub 2009 Dec 17. PMCID: PMC2813483
4. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, ..., **Wheeler DA**, Pearson JV, McPherson JD, Gibbs RA, Grimmond SM. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491: 399-405.
5. Powell BC, Jiang L, Muzny DM, Treviño LR, Dreyer ZE, Strong LC, **Wheeler DA**, Gibbs RA, Plon SE. (2012). Identification of TP53 as an Acute Lymphocytic Leukemia Susceptibility Gene Through Exome Sequencing. Pediatric Blood and Cancer 60: E1-3.
6. The Cancer Genome Atlas Research Network. (2013). Integrative analysis of genomic and molecular alterations in clear cell renal cell carcinoma. Nature 499: 43-49.

D. Research Support

1U24CA143843-04 ($10M, PI: Wheeler) 09/29/09 – 07/31/14 NCI:The BCM Tumor Genome Characterization Center . The major goals of this project are to analyze sets of tumors plus, when appropriate, matched normal tissue to characterize and enumerate the somatic changes occurring in 500 patients for each of 20-25 tumors types over the next 5 years.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Effect of whole genome rearrangements on chromosomal domains and gene regulation in cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Lynda Chin, Department of Genomic Medicine, M. D. Anderson Cancer, The University of Texas, Houston, 77054

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Kadir C. Akdemir, Department of Genomic Medicine, M. D. Anderson Cancer, The University of Texas, Houston, 77054	

Background and preliminary data

Mammalian genomes are organized into discrete chromosomal structures – termed topologically associating domains (TADs) – spanning several megabases and within which stretches of DNA preferentially contact each other. As DNA sequences often establish long-range looping within their own TADs, the probability of any given promoter contacting its cis-regulatory elements is much higher for sequences in the same TAD. Maintaining the structural integrity of spatial organization of the genome is therefore critical for a cell's proper transcriptional regulation. The importance of TADs has been suggested in a number of studies. For instance, localization of TADs along the chromosome appears to be conserved between different cell-types as well as during cellular differentiation. Boundaries of TADs are often enriched in housekeeping genes, binding sites of genomic insulator - CTCF, short interspersed elements (SINEs) and/or tRNA genes. Deletion of a boundary between two TADs in the X-chromosome inactivation center led to their partial fusion and resulted in long-range transcriptional misregulation. Despite growing interest in topological configurations of chromatin, the effects of copy number variations (CNVs) on TAD boundaries and consequent impact in cancer development or progression have not been explored.

In this study we will to investigate the following questions:

Are TAD boundaries (inter-TAD) or TAD domains (intra-TAD) affected by the CNVs in tumor cells compared to matched normal cells?
Are elicited regions associated with deregulation (transcriptionally and/or epigenetically) of genes contained by the TADs?

To date the study of higher order chromatin structure and their impact during development and disease states have relied on model-based systems. We can leverage the ICGC pan-cancer dataset to conduct the first ever survey to study the impact of disruptions in topological structures (changes in/between TAD domains) in the human genome on gene regulation and tumorigenesis. We would like to utilize whole genome and transcriptome datasets as primary resources for this study.

Timelines & resources dedicated to project

We anticipate that identification/characterization of boundary disruption and intra-domain changes would take ~2 months.
Integration of these data with gene expression and epigenomic datasets, data analysis and interpretation would take ~9-12 months.

In addition to available cloud resourced possibly provided by ICGC, we have access to a super computer hosted by University of Texas at Austin (Texas Advance Computer Center). We are planning to use our own computational resources when necessary.

International Cancer Genome Consortium

Research proposal

Surveying and characterizing copy number variations with respect to TAD formations

Higher order chromatin structure has been shown to influence transcriptional regulation. We will investigate the impact of copy number variations (CNVs) on chromosome conformation, focusing on inter-TAD and intra-TAD changes. To parallelize the variant annotations, we plan to query ICGC generated pan-cancer VCF file(s) with GEMINI software package [1]:

1) We will first focus on CNVs that could affect the boundaries between TAD domains with two possible outcomes:

a) CNVs in TAD boundaries could cause fusion of neighboring domains – by deletion of boundary element between TADs, as occurred in the example involving the X-chromosome inactivation center [2].

b) CNVs may disrupt an intact domain into two different parts – by insertion of an ectopic boundary, such as Fgf8 locus rearrangements resulting in new promoter-enhancer pairs [3].

2) We will identify CNVs that occur within TAD domains as disruptions may affect the proper distance between promoters and their enhancers without creating de novo TAD boundary or domains. An example of this is tandem duplication in the TAD containing HoxD cluster interferes with regulatory interactions and leads to down-regulation of HoxD genes while preserving the original TAD formation [4].

Given the stability of TAD positioning, we hypothesize that existing genome-wide datasets can be used to guide the proposed analysis. We plan to use TAD domains that are established for both human embryonic stem cells and differentiated fibroblasts [5] for surveying the impact of CNVs on higher order chromatin structure and gene regulation in pan-cancer datasets.

Profiling gene expression and epigenetic features around the elicited regions

1) We will further investigate whether identified regions influence the expression dynamics of genes within the same or flanking domains. Gene lists will be generated for each TAD domain and average expression of each TAD will be used to check whether there is a significant difference between affected genomes versus non-affected genomes. We will focus on not only expression patterns of coding genes but also non-coding transcripts (wherever data is available), as these elements are known to regulate distal genes. Another important question we hope to address is whether disrupted regions could lead to redirection of regulatory elements, exposing promoters to new enhancers and thereby causing mis-regulated gene expression patterns. Toward this end, we plan to focus on transcriptional changes within the elicited TAD domains and whether there is a significant correlation between possible topological reshaping and observed transcriptional alterations.

2) DNA methylation has been implicated to cause TAD domain-wide epigenetic alterations during tumorigenesis [6]. Therefore we would like to profile DNA methylation patterns of CpG islands around the altered genomic domains (wherever data is available) and compare the results with intact-domain genomes to understand whether observed changes are associated with distinct epigenetic landscapes. Additionally, we would like to utilize massive epigenomic datasets generated by Roadmap Epigenetic Consortia for localization of tissue-specific enhancers in normal tissue types, to relate the changes in topological structures and regulatory domains.

References

1.Paila, U., Chapman, B.A., Kirchner, R. & Quinlan, A.R. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput Biol* 9, e1003153 (2013).

2.Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381-5 (2012).

3.Marinic, M., Aktas, T., Ruf, S. & Spitz, F. An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Dev Cell* 24, 530-42 (2013).

4.Montavon, T., Thevenet, L. & Duboule, D. Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. *Proc Natl Acad Sci U S A* 109, 20204-11 (2012).

5.Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376-80 (2012).

6.Hon, G.C. et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res* 22, 246-58 (2012).

Legacy plans

Python package that profiles the TAD domain disruptions and possible JAVA libraries that would be used for integrating different data sources will be shared with the public as the project progress.

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center

CURRICULUM VITAE

Lynda Chin, M.D.

PRESENT TITLE AND AFFILIATION

Department Chair, Department of Genomic Medicine, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX

Professor, Department of Genomic Medicine, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX

Scientific Director, MD Anderson Institute for Applied Cancer Science, Houston, TX

EDUCATION

Degree-Granting Education

Brown University, Providence, RI, BS, Magna Cum Laude, 1988, Neuroscience

Albert Einstein College of Medicine, Bronx, NY, MD, 1993, Medicine

Postgraduate Training

Clinical Internship, Medicine, Columbia Presbyterian Medical Center, New York, NY, 1993-1994

Clinical Residency, Dermatology, Albert Einstein College of Medicine, Bronx, NY, 1994-1997

Research Fellowship, Molecular Genetics, Albert Einstein College of Medicine, Bronx, NY, 1994-1997

EXPERIENCE/SERVICE

Academic Appointments

Clinical Instructor, Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, 1997-1998

Assistant Professor, Department of Dermatology, Harvard Medical School, Boston, MA, 1998-2004

Associate Professor, Department of Dermatology, Harvard Medical School, Boston, MA, 2005-2009

Professor, Department of Dermatology, Harvard Medical School, Boston, MA, 2009-2011

Professor, Department of Genomic Medicine, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, 9/2011-present

SELECTED HONORS AND AWARDS

The Charles E. Culpeper Scholarships in Medical Science, 2001

The Goldhirsh Brain Tumor Foundation Research Award, 2002

The James S. McDonnell Foundation 21st Century Research Award, 2003

The Claire and Richard Morse Research Award, 2004

The Milstein Innovation Award, American Skin Association, 2009

Elected, Institute of Medicine of National Academies (IOM), 2012

Kadir Caner Akdemir

kcakedemir@mdanderson.org

EDUCATION

UTHSC – MD ANDERSON CANCER CENTER, Ph.D. in Bioinformatics (2008-2013)
Yeditepe University (Istanbul/Turkey), BSc. in Computer Engineering *(2002 – 2007)*
Yeditepe University (Istanbul/Turkey), BSc. in Genetics & Bioengineering *(2004 – 2008)*

EXPERIENCE

MD ANDERSON CANCER CENTER: (2013-present) Post-doctoral research fellow, Genomic Medicine Department
YEDITEPE UNIVERSITY: (Peer-Teaching Assistant – 2006-2007): C-Programming, Data Structures Labs
HARVARD MEDICAL SCHOOL: (Summer Intern - 2007- 7 weeks): Beth Israel Deaconess Medical Center Genomic Center, Bioinformatics Core. *Project: Normalization of different microarray platforms*.

AWARDS, SCHOLARSHIPS, HONORS

2012	Best Poster - MD Anderson Alumni and Faculty Association Graduate Student Award in Basic Science
2012	Trainee Choose Winner - MD Anderson Alumni and Faculty Association Graduate Student Award in Basic Science
2012	G&D Program Retreat Best Poster Award
2011	GSBS Travel Award
2010-2011	Center for Cancer Epigenetic Scholar – a year long stipend and tuition support

PEER-REVIEWED PUBLICATIONS AND MANUSCRIPTS IN PRESS

1. Genome-wide profiling reveals stimulus-specific functions of p53 during differentiation and DNA damage of human embryonic stem cells.
Akdemir KC*, Jain AJ*, Alton K, Aronow B, Xu X, Cooney A, Li W[#], Barton MC[#].
Nucleic Acids Res. 2013 Sep 27. [Epub ahead of print] PMID: 24078252

2. The postnatal role of Sox9 in cartilage.
Henry SP, Liang S, **Akdemir KC**, de Crombrugghe B.
J Bone Miner Res. 2012 Jul 6. PMID: 22777888

3. Ubp8 and SAGA regulate Snf1 AMP kinase activity.
Wilson MA, Koutelou E, Hirsch C, **Akdemir KC**, Schibler A, Barton MC, Dent SY.
Mol Cell Biol. 2011 Aug;31(15):3126-35. PMID: 21628526

4. TRIM24 links a non-canonical histone signature to breast cancer.
Tsai WW*, Wang Z*, Yiu TT, **Akdemir KC**, Xia W, Winter S, Tsai CY, Shi X, Schwarzer D, Plunkett W, Aronow B, Gozani O, Fischle W, Hung MC, Patel DJ, Barton MC.
Nature. 2010 Dec 16;468(7326):927-32. PMID: 21164480

5. Direct activation of forkhead box O3 by tumor suppressors p53 and p73 is disrupted during liver regeneration in mice.
Kurinna S, Stratton SA, Tsai WW, **Akdemir KC**, Gu W, Singh P, Goode T, Darlington GJ, Barton MC.
Hepatology. 2010 Sep;52(3):1023-32. PMID: 20564353

6. Sequencing, analysis, and annotation of expressed sequence tags for Camelus dromedarius.
Al-Swailem AM, Shehata MM, Abu-Duhier FM, Al-Yamani EJ, Al-Busadah KA, Al-Arawi MS, Al-Khider AY, Al-Muhaimeed AN, Al-Qahtani FH, Manee MM, Al-Shomrani BM, Al-Qhtani SM, Al-Harthi AS, **Akdemir KC**, Inan MS, Otu HH.
PLoS One. 2010 May 19;5(5):e10720. PMID: 20502665

* Co-first authors [#] Co-corresponding

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Profiling long intergenic non-coding RNA interactions in the cancer genome.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Lynda Chin, M.D. at The UT MD Anderson Cancer Center (UTMDACC)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Samir B. Amin, MBBS, UTMDACC Chang-Jiun Wu, M.D., Ph.D., UTMDACC	

Background and preliminary data

The primary focus of this study is to uncover emerging functional role of long intergenic non-coding RNA (lincRNA) in progression of cancer. Although functional genomics studies have enabled us to comprehensively characterize cancer genome and driver events, it is unclear at large in mechanistically linking these driver events with ostensibly heterogeneous nature of downstream gene expression program they orchestrate. Recently, lincRNAs have been shown to serve as a linker domain at gene regulatory regions and impact on normal development as well as cancer progression. However, mechanism of such interactions and target specificity, if any remains unknown. (Guttman and Rinn, 2012) Specifically, we are working on to understand *modular interactions* of lincRNAs with mutant driver proteins and chromatin remodeler complexes in altering transcriptional and epigenetic landscape of cancer genome. We have used TCGA RNA-seq, mutation and methylation data for melanoma and performed integrative analysis to uncover differential lincRNA expression in context of mutant driver cancer subtypes, e.g., In melanoma, we observed enrichment of selective lincRNAs in BRAF V600E-PTEN deletion subtype. Based on current literature, we hypothesize that subtype-specific enrichment of these lincRNAs contain sequence-specific motif or RNA structural motif and thus, coordinate DNA-protein interactions at promoter regions of known cancer genes and thus, drive carcinogenesis. In preliminary analysis in melanoma data, we observe enrichment of AluYc family transposable elements in coding region of these lincRNAs and their putative interactions in regulatory domains of known cancer genes.

The ICGC/TCGA platform will enable us to leverage matched WGS, RNA-seq and methylation data to perform such integrative analysis. We see this as a unique and perhaps the first ever large-scale effort to profile tissue/subtype-specific expression of lincRNAs and to reveal their functional impact.

Here, we specifically propose to 1) profile lincRNA enrichment across tumors (tissue-specific) and tumor subtypes, 2) identify modular lincRNA-DNA interactions at promoter regions of known cancer genes, and 3) assess impact of non-coding variants in context of subtype-specific enrichment of lincRNAs.

Timelines & resources dedicated to project

1. lincRNA expression pipeline: For ~ 1,500 samples, require ~ 3,000 CPU hours with ~ 64 GB RAM for Cufflinks based RPKM estimation. | Dependency: Preferably TopHat or Mapspliced aligned bam file to reference genome, hg19. Post-processing, if any for batch effect, low QC samples, etc.

2. motif discovery pipeline: Using HOMER and Tea repeat analysis pipeline - variable CPU hours with average of 6,000 CPU hours and ~ 64 - 128 GB RAM. | Dependency: matched WGS bam file, level 3 RNA-seq gene expression data.

3. non-coding variant annotation pipeline: Using ANNOVAR and FunSeq (Gerstein Lab) - variable CPU hours with ~ 64 GB RAM. | Dependent on VCF file, preferably VCF v 4.0.

Research proposal

1. Profiling lincRNA expression:
lincRNAs expression will be quantified using pre-built pipeline (Unix shell and R scripts). Briefly, RNA-seq bam file aligned to reference genome (hg19) will be used as an input for Cufflinks based RPKM estimation of annotated lincRNAs (Cabili MN,2011) with and without UCSC known genes. Quality checks will be performed to minimize issues with overlapping reads and reads mapping to multiple regions. Next, we will use hierarchical and consensus NMF clustering to determine subtype-specific lincRNA enrichment and also study statistical association, if any with driver events, as identified by other working groups.

2. Define modular interactions of lincRNAs:
Using 'guilt-by-association' principle and network analysis, we will look for preferential enrichment of lincRNAs at promoter regions of known and differentially expressed cancer genes. Then, we will search for putative sequence-specific motif mediating lincRNA - DNA interaction at these promoter regions using motif discovery pipeline (see Timeline section). We are currently developing an approach to find lincRNA-protein interactions at gene promoter regions, particularly in context of differentially expressed TFs and chromatin remodeler proteins.

3. Assess functional impact of non-coding variants:
We will predict functional consequences of non-coding variants in conserved regions on ~1,000 linRNAs (Guttman, 2009) using variant annotation pipeline (see Timeline section) as well as VCFs obtained from the ICGC variant calling AWG. Also, we will focus on finding recurrent variants in promoter regions of known cancer genes which disrupt putative functional interactions obtained from aim 2.

Legacy plans

All three pipelines, lincRNA expression, motif discovery and non-coding variant annotations will be maintained using git version control, and will be shared among members of respective AWG(s).

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center

CURRICULUM VITAE

Lynda Chin, M.D.

PRESENT TITLE AND AFFILIATION

Department Chair, Department of Genomic Medicine, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX

Professor, Department of Genomic Medicine, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX

Scientific Director, MD Anderson Institute for Applied Cancer Science, Houston, TX

EDUCATION

Degree-Granting Education

Brown University, Providence, RI, BS, Magna Cum Laude, 1988, Neuroscience

Albert Einstein College of Medicine, Bronx, NY, MD, 1993, Medicine

Postgraduate Training

Clinical Internship, Medicine, Columbia Presbyterian Medical Center, New York, NY, 1993-1994

Clinical Residency, Dermatology, Albert Einstein College of Medicine, Bronx, NY, 1994-1997

Research Fellowship, Molecular Genetics, Albert Einstein College of Medicine, Bronx, NY, 1994-1997

EXPERIENCE/SERVICE

Academic Appointments

Clinical Instructor, Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, 1997-1998

Assistant Professor, Department of Dermatology, Harvard Medical School, Boston, MA, 1998-2004

Associate Professor, Department of Dermatology, Harvard Medical School, Boston, MA, 2005-2009

Professor, Department of Dermatology, Harvard Medical School, Boston, MA, 2009-2011

Professor, Department of Genomic Medicine, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, 9/2011-present

SELECTED HONORS AND AWARDS

The Charles E. Culpeper Scholarships in Medical Science, 2001

The Goldhirsh Brain Tumor Foundation Research Award, 2002

The James S. McDonnell Foundation 21st Century Research Award, 2003

The Claire and Richard Morse Research Award, 2004

The Milstein Innovation Award, American Skin Association, 2009

Elected, Institute of Medicine of National Academies (IOM), 2012

Samir B. Amin

Contact Information

Graduate Student in Chin Lab
Genomic Medicine, UTMDACC
1901 East Rd. #1954
Houston, TX 77054 USA

Phone: (713) 792-8598
Fax: (713) 792-6882
E-mail: sbamin@bcm.edu
Website: www.sbamin.com

Education

Baylor College of Medicine, Houston, TX, USA
Graduate Student — August 2011 - Present
Structural & Computational Biology & Molecular Biophysics (SCBMB) Program

Medical College & M S University of Baroda, Vadodara, Gujarat India
Bachelor of Medicine, Bachelor of Surgery — November 1998 - March 2005
GPA: 3.77 / 4.0
Accredited by The Educational Commission for Foreign Medical Graduates (ECFMG®), Philadelphia, PA USA

Professional Experience

Graduate Student — January 2012 - Present
Laboratory of Dr. Lynda Chin, MD
Department of Genomic Medicine
The University of Texas MD Anderson Cancer Center, Houston, TX USA

Research Fellow — January 2008 - June 2011
Department of Medical Oncology
Dana-Farber Cancer Institute, Boston, MA USA
Advisor: Dr. Nikhil C. Munshi, MD & Dr. Kenneth C. Anderson, MD

Adjunct Research Fellow — January 2008 - June 2011
Department of Biostatistics & Computational Biology
Dana-Farber Cancer Institute, Boston, MA USA
Advisor: Dr. Cheng Li, PhD

Resident Physician — December 2005 - November 2007
Bhailal Amin General Hospital, Vadodara, Gujarat, India
Primary rotations: Medical Oncology, ICU, ER services
Supervisor: Dr. Atul Jani, MD

Awards

Prof. John J. Trentin Award for Scholastic Excellence — 2012
Baylor College of Medicine

Professional Memberships

American Society of Hematology — 2009-2011

cv trimmed v 1.5, Nov 27, 2013. www.sbamin.com/about/cv

CURRICULUM VITAE

Chang-Jiun Terrence Wu, MD. PhD

Instructor
Department of Genomic Medicine
The UT MD Anderson Cancer Center
Houston, TX, USA

1901 East Road # 1954
3SCRB6.4101.03
Houston, TX 77054
CWu7@mdanderson.org
Tel: (713)-794-5258

EDUCATION

Degree-Granting Education

Graduate Program of Bioinformatics, Boston University, Boston MA.
Ph.D. of Bioinformatics, 2003-2009
Medical College, National Taiwan University, Taiwan. MD, 1987-1994

Postgraduate Training

Clinical Internship. National Taiwan University Hospital, Taiwan, 1993-1994
Clinical Residency and Chief Residency. National Taiwan University Hospital, Taiwan, 1996-2000
Research Fellowship. Department of Biomedical Engineering, Boston University, Boston MA, 2009
Research Fellowship. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston MA, 2009-2011
Research Fellowship. Department of Genomics Medicine, UT MD Anderson Cancer Center, Houston TX, 2011-2012
Instructor. Department of Genomics Medicine, UT MD Anderson Cancer Center, Houston TX, 2012-present

CREDENTIALS

Board Certification

Diplomat of the Society of Medicine, Taiwan, 1994
Diplomat of the Society of Otorhinolaryngology, Taiwan, 2000

HONORS AND AWARDS

Honors in International Mathematics Olympics Competition. 1985

EDITORIAL AND REVIEW ACTIVITIES

Invited reviewer for program committee of ISMB/ECCB 2013

PROFESSIONAL MEMBERSHIPS

Taiwan Society of Medicine:
Member, 1994 -present
The Cancer Genome Atlas, Genome Data Analysis Center:
Scientist and Programmer, 2010 –present
The Cancer Genome Atlas, Cutaneous Melanoma Analysis Working Group:
Analysis and Reproducibility Coordinator
Manuscript writing group

--

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mutation and expression landscapes of tRNA genes in cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Lynda Chin, MD Anderson Cancer Center, ICGC
P. Andrew Futreal, MD Anderson Cancer Center, ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Tony Gutschner, MD Anderson Cancer Center **Hannah Cheung, MD Anderson Cancer Center**	**Yitzhak Pipel, Weizmann Institute of Science**

Background and preliminary data

tRNA expression levels are tissue-specific and recent reports indicate that a subset of them is differentially expressed between normal and cancerous tissues. In fact, about 50% of the tRNA genes lie on chromosomes 1 and 6, with the majority of those on chr1q and chr6p, which are frequently amplified in human cancers and may account for this differential expression. However, the notion that tRNA gene expression is altered to allow for differential codon usage and high translation efficiency of specific classes of genes for differentiation or other cellular processes remains controversial.

In our preliminary analysis of tRNA expression changes in diverse cancer types we found that the tRNA pool consists of two subsets: The "proliferation tRNAs" are up-regulated in diverse cancers and are typically repressed in senescence, and a distinct subset of tRNAs, that often carry alternative anti-codons for same amino acids, are often repressed in cancer and up-regulated in senescence. We found that the proliferation tRNAs often carry anti-codons that correspond to the codon usage of protein-coding genes that are typically induced in cancer and in cellular proliferation. Epigenetic analyses of histone modification patterns in the vicinity of the proliferation tRNA genes shows enhanced patterns of active transcription histone marks in cancer.

However, further analysis to uncover cancer-specific mutations in tRNA genes is limited due to the low coverage of tRNA genes in whole exome sequencing data because of high sequence similarities. Therefore, the pan-cancer whole genome sequencing datasets would allow us to identify any existing mutations in tRNA genes, and the RNA-Seq data would provide insight into any differential expression of tRNAs between matched normal-tumor pairs. This analysis will help to gain further insight into potential codon usage bias in human cancers.

Timelines & resources dedicated to project

Develop methods to extract and map reads for tRNA genes (3-4 months)
Determine the somatic mutation landscape for tRNA using the new tools (5 months)
Determine the relationship between overexpressed proliferative tRNAs and the levels of their target genes (4 months)

Research proposal
There are approximately 500 tRNA as well as 100 tRNA pseudogenes within the human genome with high sequence similarities among them. Our first goal is to develop novel methods to extract and correctly map these tRNA reads. This will involve, among others, developing multiple alignment methods that determine unique and identifiable regions of each tRNA gene. Once established, the somatic mutation landscape of these genes will be characterized within each tissue type and in a pan-cancer analysis. We will pay special attention to chromosomal aberrations such as gains on 1q and 6p. Next, any mutations identified in tRNA genes and promoter regions will be further analyzed with respect to their impact on tRNA expression levels. Here, RNA-Seq data will be correlated with potential mutations, e.g. tRNA gene deletions and / or amplifications. Finally, we will determine whether somatic variants within codons, introns, and intergenic regions of tissue-specific genes can be linked to any changes in their codon usage in each tissue, especially if the variants alter the GC content of the gene. Then, the expression levels of "proliferative tRNAs" will be compared with the tRNA population. If overexpressed, the levels of the target genes associated with these proliferative tRNAs will be measured. This will delineate whether changes in codon usage are due to *cis*-acting DNA elements or through tRNA translational control.

Legacy plans
Our methods for mapping tRNA genes and measuring their tRNA population distributions will be made publicly available.

International Cancer Genome Consortium

CURRICULUM VITAE

Lynda Chin, MD
Email: LChin@mdanderson.org

PRESENT TITLE AND AFFILIATION

Professor and Chair, Dept of Genomic Medicine
Scientific Director, Institute for Applied Cancer Science

EDUCATION/TRAINING

Degree-Institution and Location

09/84-06/88 BS in Neuroscience, Brown University, Providence, RI
09/89-06/93 MD in Medicine, Albert Einstein College of Medicine, Bronx, NY
07/93-06/94 Internship, Columbia Presbyterian Medical Center, NY, NY
07/94-06/97 Residency, Albert Einstein College of Medicine, Bronx, NY
07/93-06/97 Postdoctoral, Albert Einstein College of Medicine, Bronx, NY

POSITIONS AND HONORS

1996 – 1997 Chief Resident, Dermatology, Albert Einstein College of Medicine (AECOM), NY
1998 – 2004 Assistant Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute (DFCI), Boston, MA
1999 – 2004 Scientific Director, Arthur & Rochelle Belfer Cancer Genomics Center, DFCI, Boston, MA
2005 – 2009 Associate Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2008 – Member, scientific steering committee, International Cancer Genome Consortium (ICGC).
2009 – 2011 Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2009 – Associate Member, the Broad Institute of MIT and Harvard, Boston, MA
2009 – 2011 Co-director, Melanoma Program, Dana-Farber/Harvard Cancer Center, Boston, MA
2009 – 2011 Scientific Director, the Belfer Institute for Applied Cancer Science, DFCI, Boston, MA
2009 – Member, Executive Subcommittee, The Cancer Genome Atlas (TCGA), USA
2011 – Professor and Chair, Department of Genomic Medicine, UTMDACC, Houston, TX
2011 – Scientific Director, Institute for Applied Cancer Science, UTMDACC, Houston, TX

ONGOING RESEARCH SUPPORT

1P01 CA163222 NIH/NCI Fisher (PI) 12/01/11-11/30/16 Role: Project PI
1U01 CA168394 NIH/NCI Mills (PI) 05/01/12-04/31/17 Role: Project PI
R1204 CPRIT Chin (PI) 12/16/11-12/31/16 Role: PI
7P01 CA117969 NIH/NCI DePinho (PI) 04/15/06-12/31/15
7U01 CA141508 NIH/NCI Chin (PI) 08/01/09-07/31/14
5U24 CA143845 NIH/NCI Getz (PI) 08/01/09-07/31/14
U24 CA144025 NIH/NCI Kucherlapati (PI) 08/01/09-07/31/14
5U54 CA163125 NIH/NCI Chin (PI) 08/01/09-07/31/14

PUBLICATIONS (last 5)

Peer-Reviewed Original Research Articles

Cheng, C.S. et al., Nat Commun. 2013 4:2672.
Yen, J. et al. Genome Biol. 2013 14:R113.
Brennan, C.W. et al., Cell 2013 155:462.
Cancer Genome Atlas Research Network, et al. Nat Genet. 2013 45:1113.
Watson, I.R. et al., Nat Rev Genet. 2013 14:703.

International Cancer Genome Consortium

CURRICULUM VITAE

Andrew Futreal, PhD
Phone: (713) 794-4764 Email: afutreal@mdanderson.org

PRESENT TITLE AND AFFILIATION

Primary Appointment
Professor, Department of Genomic Medicine, Division of Cancer Medicine
The University Of Texas MD Anderson Cancer Center

Honorary Faculty Member
Director, Cancer Genetics and Genomics
Co-Director, Cancer Genome Project
Wellcome Trust Sanger Institute

EDUCATION

Degree-Granting Education
B.S. in Biology, UNC-Charlotte, 1987
Ph.D. in Pathology, UNC-Chapel Hill, 1993

Postgraduate Training
Postdoctoral Fellowship, 1993-1995
Post-Doctoral IRTA Fellow Award
National Institute of Environmental Health Sciences,
National Institutes of Health

EXPERIENCE/SERVICE

Academic Appointments
Honorary Faculty Member
Director, Cancer Genetics and Genomics
Co-Director, Cancer Genome Project
Wellcome Trust Sanger Institute

ONGOING RESEARCH SUPPORT

N/A STARS Award Futreal (PI)
06/06/2012 – 06/05/2015
R1205 Futreal (PI) Cancer Prevention & Research Institute of Texas (CPRIT)
11/02/2011 – 11/01/2016

PUBLICATIONS (last 5)

Peer-Reviewed Original Research Articles
Stephens PJ, et al. Nature. 2012 May 16;486(7403):400-4.
Jonasch E, et al. Mol Cancer Res. 2012 Jul;10(7):859-80.
Nik-Zainal S, et al. Cell. 2012 May 25;149(5):979-93.
Nik-Zainal S, et al. Cell. 2012 May 25;149(5):994-1007
Ong CK, et al. Nat Genet. 2012 May 6;44(6):690-3.

International Cancer Genome Consortium

CURRICULUM VITAE

Tony Gustchner, PhD
TGutschner@mdanderson.org

Education/Training

09/08 Diploma, Martin-Luther-University, Halle-Wittenberg, Germany
09/12 Dr. rer. Nat. (Ph.D.), German Cancer Research Center and Ruperto Carola University Heidelberg, Germany
10/12 Postdoc, German Cancer Research Center and Ruperto Carola University Heidelberg, Germany
08/13 Postdoc, UT M.D. Anderson Cancer Center, Houston, TX, USA

Positions

2008-2012	PhD Student, Junior Research Group "Molecular RNA Biology & Cancer", German Cancer Research Center (DKFZ) Heidelberg and Institute of Pathology, University Hospital Heidelberg, Germany
2012-2013	Postdoctoral Fellow, Junior Research Group "Molecular RNA Biology & Cancer", German Cancer Research Center (DKFZ) Heidelberg and Institute of Pathology, University Hospital Heidelberg, Germany
2013-present	Postdoctoral Fellow, Department of Genomic Medicine, UT M.D. Anderson Cancer Cancer, Houston, TX

Honors

2008-2011	DKFZ PhD Fellowship
2011	University of Heidelberg Travel Award (Excellence Initiative)
2012	Keystone Future of Science Scholarship
2013	Elisabeth-Gateff-Prize of the German Genetics Society (Ph.D. Award)

Publications (last 7)

Hämmerle M, **Gutschner** T, et al. Posttranscriptional destabilization of the liver-specific long noncoding RNA HULC by the IGF2 mRNA-binding protein 1 (IGF2BP1). Hepatology. 2013 May 31. doi: 10.1002/hep.26537.

Eißmann M…**Gutschner** T…et al., A functional yeast survival screen of tumor-derived cDNA libraries designed to identify anti-apoptotic mammalian oncogenes. PLoS One. 2013 May 22;8(5):e64873. doi: 10.1371/journal.pone.0064873.

Gutschner T, Hämmerle M, Diederichs S. MALAT1 -- a paradigm for long noncoding RNA function in cancer. J Mol Med (Berl). 2013 Jul;91(7):791-801. doi: 10.1007/s00109-013-1028-y. Epub 2013 Mar 26. Review.

Gutschner T, et al., The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 2013 Feb 1;73(3):1180-9. doi: 10.1158/0008-5472.CAN-12-2850. Epub 2012 Dec 14.

Eißmann M, **Gutschner** T, et al., Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. RNA Biol. 2012 Aug;9(8):1076-87. doi: 10.4161/rna.21089. Epub 2012 Aug 1.

Gutschner T, Diederichs S.The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. 2012 Jun;9(6):703-19. doi: 10.4161/rna.20481. Epub 2012 Jun 1. Review.

Gutschner T, Baas M, Diederichs S. Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases. Genome Res. 2011 Nov;21(11):1944-54. doi: 10.1101/gr.122358.111. Epub 2011 Aug 15.

CURRICULUM VITAE

Hannah Cheung, PhD, PMP

HCCheung@mdanderson.org
(713) 301-4311

Education

2001	B.Sc. (Hon) Molecular Genetics, University of Alberta, Canada
2008	Ph.D. Genes and Development, UT-HSC, Houston, Texas
2008-13	Postdoctoral Training, Baylor College of Medicine, Houston, Texas
2013	PMP certified July 15, 2013, PMP Number: 1642228

Positions

2013-present	Research Scientist, MD Anderson Cancer Center, Futreal Lab
2013	Freelancer, Cactus Communications, Inc.
2008-2013	Postdoctoral Associate, Baylor College of Medicine, Plon Lab
2002-2008	Graduate Student, UT-HSC, Cote Lab
2001-2002	Research Technologist, von Borstel Lab
1999-2000	Research Intern, Hao/Roa Lab
1998	Research Assistant, Reha-Krantz Lab

Relevant Publications

Izaguirre, D.I., Zhu, W., Hai, T., **Cheung, H.C.**, Krahe, R. and Cote, G.J. PTBP1-dependent Regulation of USP5 Alternative RNA Splicing Plays a Role in Glioblastoma Tumorigenesis *Molecular Carcinogenesis* (2012) 51: 895-906.

Cheung, H.C., Hai, T., Baggerly, K.A., Tsavachidis, S., Krahe, R., and Cote, G.J. Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cells. *Brain* (2009) 132: 2277-88.

Cheung, H.C., Baggerly, K.A., Tsavachidis, S., Bachinski, L.L., Neubauer, V.L., Nixon, T.J., Aldape, K.D., Cote, G.J., and Krahe, R. Global analysis of aberrant pre-mRNA splicing in glioblastoma using exon expression arrays *BMC Genomics* (2008) 9:216.

Research and Training Support

2009 - 2011	Early Career Award, Thrasher Research Fund
2006 - 2008	Rosalie B. Hite Fellowship

International Cancer Genome Consortium

CURRICULUM VITAE

YITZHAK PILPEL, PhD
THE GEN MAY PROFESSIONAL CHAIR
Email: pilpel@weizmann.ac.il Webpage: http://longitude.weizmann.ac.il

Education

1990-1993 B.Sc. in Biology at the Tel Aviv University.
1993-1994 Studies towards M.Sc. at the Feinberg Graduate School of the Weizmann Institute of Science. Transferred to the direct path to Ph.D.
1995-2000 Ph.D. studies with distinction at the Weizmann Institute of Science under Prof. Doron Lancet and Prof. Ephraim Katchalski-Katzir. Thesis title: "Structural and evolutionary genomics of molecular recognition repertoires"
2000-2002 Post-doc, Department of Genetics and Center for Computational Genetics, Harvard Medical School

Positions

1999-2002 Post-doctoral research fellow with Dr. George M. Church, Department of Genetics, and Lipper Center for Computational Genetics, Harvard Medical School
2000-2001 Consultant for Pfizer Inc., Cambridge, MA. Subject: Sequence, structure and function in G protein-coupled receptors.
2003-2008 Senior Scientist, department of Molecular Genetics, Weizmann Institute of Science
2008- 2009 Visiting Professor, the Department of Systems Biology, Harvard Medical School
2008 – present Associate Professor with tenure at the department of Molecular Genetics, Weizmann Institute of Science
2009 – 2011 Consultant for Evogene, Rehovot, Israel

Awards and Honors

1998 First prize in the National competition for multidisciplinary Ph.D. study, Hebrew University
2000 Prize of Distinction for outstanding Ph.D. studies at the Feinberg Graduate School, Weizmann Institute of Science.
2005 EMBO Young Investigator award
2006 James Heineman Award
2007 Levinson Prize in Biology Weizmann Institute of Science, Scientific Council
2008 European Research Council grant award
2010 Hestrin Prize. The Israeli Society for Biochemistry and Molecular Biology
2011 EMBO member
2012 Michael Bruno Memorial Award
2013 IBM faculty Award

Relevant Publications

Gingold H, Dahan O, **Pilpel Y**.Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. Nucleic Acids Res. 2012 Nov 1;40(20):10053-63 Epub 2012 Aug 31.
Gingold, H. and Pilpel Y. Determinants of translation efficiency and accuracy. Mol Systems Biol. 2011 Apr:
Limor-Waisberg K., Carmi A., Scherz A., Piilpel Y., Furman I. Specialization versus adaptation: two strategies employed by cyanophages to enhance their translation efficiencies. Nucleic Acids Res. 2011 Aug;39(14):6016-28. doi: 10.1093/nar/gkr169. Epub 2011 Apr 5.
Navon S, Pilpel Y. The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. Genome Biol. 2011;12(2):R12. doi: 10.1186/gb-2011-12-2-r12. Epub 2011 Feb 1.
Man O, Pilpel Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 2007 Mar;39(3):415-21. Epub 2007 Feb 4. Erratum in: Nat Genet. 2007 May;39(5):688.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of WGS pan-cancer dataset for cancer specific eQTLs

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Nancy Cox, University of Chicago

Robert L. Grossman, University of Chicago

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Allison P. Heath, University of Chicago **Barbara E. Stranger, University of Chicago**	**Zhenyu, Zhang, University of Chicago**

Background and preliminary data

Expression quantitative trait loci (eQTLs) are genomic loci harboring genetic variation that associates with variation in gene expression levels across a set of individuals. The identification of these loci provides important clues to the underlying genetic mechanism of altered gene expression. Studies have also shown that genetic variants associated with human diseases are enriched for eQTLs, thus the identification of these loci can be informative for generating hypotheses as to causal genes underlying complex disease.As many eQTLS (and characteristics of eQTLs, such as effect size and even direction of the allelic association) are tissue-dependent, we are particularly interested in how gene expression is regulated in cancer tissues, and how this differs from the matching normal tissues.

We are proposing a combined eQTL study with both the Cancer Genome Atlas (TCGA) dataset and ICGC WGS pan-cancer dataset. The results will also be compared to published eQTL studies from the Genotype-Tissue Expression (GTEx) project. We hope our results will help elucidate the genetic component underlying altered gene expression in cancers, and the mechanisms underlying cancer susceptibility identified with numerous cancer GWAS studies.

Timelines & resources dedicated to project

We will use approximately 200 cores of the Bionimbus Protected Data Cloud for this project.

We expect the analysis to take approximately six months (less time if more computing resources are available).

Research proposal
1. We will select ICGC WGS samples with both whole genome sequencing data and RNA sequencing or cDNA microarray data (~1500). Genotyping inferred from ICGC germline whole genomic sequencing data will also be used to identify population structure among the sample donors. Somatic gene expression levels can be obtained from either RNA sequencing or cDNA microarray data from the cancer samples. 2. We will analyze tumor eQTLs in separate groups of the same population and cancer type, starting with breast cancer in the population of European ancestry (the largest population) to establish our methodology. To be specific, germline genotyping data will be filtered with standard QC (call rate, MAF, HWE, donor relation and other QC steps), followed by population stratification together with Hapmap samples. Principal Component Analysis will also be used on tumor gene expression data to possibly create different subgroups of tumor types. Both *cis*- and *trans*- eQTLs will be identified using Matrix eQTL, an ultra fast eQTL analysis package of R. We will also explore using subgroups that control for somatic expression for Copy Number Variations (CNV) and methylation levels, both of which are available in the ICGC dataset. 3. A similar analysis will be performed on the TCGA dataset with Affymetrix SNP 6.0 for genotyping, and Affymetrix H-GU133A data for somatic expression. Different methods of genotyping and expression measurement between TCGA and ICGC will be compared. The results we obtain will also serve as a cross validation analysis of the previous eQTLs identified from ICGC WGS data. 4. We will do a meta-eQTL analysis across TCGA and ICGC data and across different tumor types. 5. We will also explore cross-tissue vs single-tissue eQTLs. Specifically, we will examine whether tumor tissues develop any evidence of a shift toward greater use of cross-tissue eQTLs that might relate to de-differentiation. More generally, we are interested in how broadly the regulatory variation acts, whether it is across multiple tissues or only in single tissues. 6. The analysis in 1) - 4) will identify eQTL targets in tumor tissues, but not necessarily eQTL that are cancer specific. We will analyze the publicly-available Genotype-Tissue Expression project (GTEx) dataset (for which Drs. Cox and Stranger are Consortium members) for eQTLs mapped in normal human tissues that match the tumor types in ICGC/TCGA samples. eQTLs identified specifically in tumor tissues, but not in the matching normal tissue will be of particular interest, and compared with previously identified cancer susceptibility loci from various GWAS studies.
Legacy plans
1. We will package our pipelines and workflows as virtual machines and make them available to any qualified researchers who want to use them in the Bionimbus Protected Data Cloud or other clouds that have access to the required data. 2. We will also make available the results of all eQTL analyses, both as flat files and via a queryable database.

Nancy J. Cox

Education

Yale University, PhD Human Genetics, 1982
University of Notre Dame, BSc in Biology, 1978

Positions

2005 - present, Professor and Chief, Section of Genetic Medicine, The University of Chicago
2004 - present, Professor, Departments of Human Genetics and Medicine, The University of Chicago
1999 - 2004, Associate Professor, Departments of Human Genetics & Medicine, The University of Chicago
1990 - 1990, Research Associate (Assistant Professor), Department of Medicine, The University of Chicago
1987 - 1990, Research Associate, Howard Hughes Medical Institute, The University of Chicago
1985 - 1987, Research Associate, Department of Human Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA
1982 - 1985, Postdoctoral Fellow in Genetic Epidemiology, Department of Psychiatry, Washington University School of Medicine, St. Louis, MO

Honors

2013 Dinstinguished Faculty Award, Biological Sciences Division, University of Chicago
2012 Pritzker Scholar, University of Chicago
2010 Winner of the Leadership Award from the International Genetic Epidemiology Society
2008 Co-winner of Landon Award, American Association for Cancer Research

Publications

1. Zhang W, Duan S, Bleibel WK, Wisel SA, Huang RS, Wu X, He L, Clark TA, Chen TX, Schweitzer AC, Blume JE, Dolan ME, **Cox NJ** (2009) Identification of common genetic variants that account for transcript isoform variation between human populations. Hum Genet 125(1):81-93. PMC2665168.
2. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, **Cox NJ** (2010) Trait associated SNPs are more likely to be eQTLS: Annotation to enhance discovery from GWAS. PLoS Genetics 2010 Apr 1;6(4):e1000888. PMC2848547.
3. Gamazon ER, Huang RS, **Cox NJ**, Dolan ME (2010) Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci USA 107(20):9287-92. PMC2889115.
4. Below JE, Gamazon ER, Morrison JV, Konkashbaev A, Pluzhnikov A, McKeigue PM, Parra EJ, Elbein SC, Hallman DM, Nicolae DL, Bell GI, **Cox NJ**, Hanis CL (2011) Genome-wide association and meta-analysis in populations from Starr County, Texas and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals. Diabetologia 54(8):2047-55. PMC3761075.
5. Gamazon ER, Nicolae DL, **Cox NJ** (2011) A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS Genetics 7(2):e1001292. PMC3033384.
6. Gamazon ER, Ziliak D, Im HK, Lacroix B, Park DS, **Cox NJ**, Huang RS (2012) Genetic architecture of MicroRNA Expression: Implications for the transcriptome and complex traits. Am J Hum Genet 90(6):1046-63. PMC3370272.
7. Im HK, Gamazon ER, Nicolae DL, **Cox NJ** (2012) On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet 90(4):591-8. PMC3322234.
8. Elbein SC, Gamazon ER, Das SK, Rasouli N, Kern PA, **Cox NJ** (2012) Genetic risk factors for type 2 diabetes: a trans-regulatory genetic architecture? Am J Hum Genet 91(3):466-77. PMC3512001.
9. Chen LS, Hsu L, Gamazon ER, **Cox NJ**, Nicolae DL. (2012) An exponential combination procedure for set-based association test in sequencing studies. Am J Hum Genet 91(6):977-86. PMC351661.
10. Davis LK, Gamazon ER, Kistner-Griffin E, Badner JA, Liu C, Cook EH, Sutcliffe JS, **Cox NJ** (2012) Loci nominally associated with autism from genome-wide analysis show enrichment of brain expression quantitative trait loci but not lymphoblastoid cell line expression quantitative trait loci. Molecular Autism 3(1):3. PMC3484025.

Robert Grossman

Education

University of California, Berkeley, Postdoc, 1984-1988
Princeton University, PhD Applied Mathematics, 1985
Harvard College, AB Mathematics, 1980

Positions

2011 – present, Chief Research Informatics Officer, Biological Sciences Division, University of Chicago
2010 – present, Professor of Medicine, Section of Genetic Medicine, University of Chicago
1988 – 2010, Professor of Mathematics, Statistics & Computer Science, University of Illinois at Chicago (Assistant Professor, 1988 – 1991; Associate Professor, 1991 – 1995; Professor 1995 – 2010)

Honors

2013 AAAS Fellow
2013 Federal 100 Award Winner
2011 Pritzker Scholar, University of Chicago

Publications

1. David R. Blair, Christopher S. Lyttle, Jonathan M. Mortensen, Charles F. Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V. Bernstam, Søren Brunak, Lars Juhl Jensen, Dan Nicolae, Nigam H. Shah, Robert L. Grossman, Nancy J. Cox, Kevin P. White, Andrey Rzhetsky, A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk, Cell Volume 155, Issue 1, pages 70-80. PMID: 24074861, PMCID: in progress.
2. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP, CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia, Blood, Volume 121, Number 6, pages 975-983, 2013, PMID: 23212519, PMCID: PMC3567344.
3. Heidi L. Alvarez, Malcolm Atkinson, Robert L. Grossman, Matthew Greenway, Christine Harvey, Allison P. Heath, Iraklis Klampanos, Joe J. Mambretti, Ray Powell, Rafael D. Suarez, Walt Wells and Kevin White, The Design of a Community Science Cloud: The Open Science Data Cloud Perspective, SC Companion: High Performance Computing, Networking Storage and Analysis, ACM Press, 2012.
4. Shi Yu, Robert Grossman and Andrey Rzhetsky, Global and Local Approach of Part-of-Speech Tagging for Large Corpora, AAAI-2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text, 2012.
5. Wenxuan Gao, Robert Grossman, Philip Yu, Christopher Brown, Matthew Slattery, Lijia Ma and Kevin White, Discovering Geometric Patterns in Genomic Data, ACM Conference on Bioinformatics, Computational Biology and Biomedicine, ACM Press, 2012.
6. Robert L. Grossman and Kevin P. White, A vision for a Biomedical Cloud, Journal of Internal Medicine, Volume 271, Number 2, pages 122-130, 2012. PMID: 22142244. PMCID: in process.
7. Xin Feng, Robert L. Grossman and Lincoln Stein, PeakRanger: a Cloud-enabled Peak Caller for ChIP-seq Data, BMC Bioinformatics, Volume 12:139, PMID: 21554709, PMCID: PMC3103446.
8. Nicolas Negre, Christopher D. Brown, Lijia Ma, et. al., Cis-Regulatory Map of the Drosophila Genome, Nature, Volume 471, pages 527–531, 2011, [doi:10.1038/nature09990], PMID: 21430782. PMCID: in process.
9. The modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V. Kharchenko, et. al., Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE, Science, Volume 330 (6012), pages 1787-1797, 2010, [DOI:10.1126/science.1198374]. PMID: 21177974. PMCID: in process.
10. Robert L. Grossman, Yunhong Gu, Joe Mambretti, Michal Sabala, Alex Szalay, and Kevin White, An Overview of the Open Science Data Cloud, Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC '10), ACM, 2010. PMCID: in process.

Allison Park Heath

Education

Rice University, Postdoctoral Computer Science, 2011
Rice University, Ph.D. Computer Science, 2010
Rice University, M.S. Computer Science, 2007
Rice University, B.S. Computer Science, 2004

Positions

2012 - present, Research Professional, Institute for Genomics & Systems Biology, University of Chicago
2011 - 2012, Consultant, Dimensional Insight, Coral Springs, FL
2010 - 2011, Postdoctoral Research Assistant, Dept. of Computer Science, Rice University, Houston, TX
2009 - 2010, Consultant, Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
2004 - 2010, Graduate Research Assistant, Dept. of Computer Science, Rice University, Houston, TX
2007 - 2010, Graduate Student Trainee, Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX
Summer 2007, Research Intern, Text Mining, Search, and Navigation, Microsoft Research, Redmond, WA
2002 - 2004, Undergraduate Research Assistant, Dept. of Computer Science, Rice University, Houston, TX

Honors

2009 NIH Keck Center Fellowship "Biomedical Discovery from Large Scale Data Sets"
2008 Google Anita Borg Scholarship
2004 NSF Graduate Research Fellowship
2003 CRA-W Distributed Mentor Program Award
2002 Brown Undergraduate Research Internship Award

Publications

1. Grossman RL, Greenway M, Heath AP, Powell R, Suarez RD, Wells W, White K, Atkinson M, Klampanos I, Alvarez HL, Harvey C, Mambretti JJ. The Design of a Community Science Cloud: The Open Science Data Cloud Perspective. Proceedings of the 2012 SC Companion: High Performance Computing, Networking, Storage and Analysis, 1051-1057 (2012).
2. Heath AP, Bennett GN, and Kavraki LE. Identifying branched metabolic pathways by merging linear metabolic pathways. Proceedings of the 15th Annual International Conference on Research in Computational Molecular Biology (RECOMB), 6577:70–84 (2011).
3. Heath AP, Bennett GN, and Kavraki LE. An algorithm for efficient identification of branched metabolic pathways. Journal of Computational Biology, 18(11):1575–1597 (2011).
4. Heath AP, Bennett GN, and Kavraki LE. Finding metabolic pathways using atom tracking. Bioinfomatics, 26(12):1548–1555 (2010).
5. Heath AP and Kavraki LE. Computational challenges in systems biology. Computer Science Review, 3(1):1–17 (2009).
6. White RW, Richardson M, Bilenko M, and Heath AP. Enhancing web search by promoting multiple search engine use. SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, p. 43–50 (2008).
7. Heath AP, Kavraki LE, and Balázsi G. Bipolarity of the Saccharomyces cerevisiae genome. 2nd International Conference on Bioinformatics and Biomedical Engineering (iCBBE), p. 330–333 (2008).
8. Balázsi G, Heath AP, Shi L, and Gennaro ML. The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. Molecular Systems Biology, 4 (2008).
9. Heath AP, Kavraki LE, and Clementi C. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. Proteins, 68(3):646–661 (2007).

Barbara E. Stranger

Education

Wellcome Trust Sanger Institute, Postdoc, 2004-2009
University of Barcelona, Postdoc, 2002-2004
University of Montana, PhD Biology, 2002
University of Chicago, BA Biological Sciences, 1994

Positions

2012-present, Assistant Professor of Medicine, Division of Genetics, University of Chicago
2010-2012, Assistant Professor, Harvard Medical School
2009-2012, Associate Geneticist, Department of Medicine, Brigham and Women's Hospital
2009-2012, Associate Member, The Broad Institute of Harvard and MIT
2009-2010, Instructor, Harvard Medical School

Honors

2012	Eleanor and Miles Shore Scholar in Medicine Award, Harvard Medical School and Brigham and Women's Hospital
2009	Stellar Abstract Award, 3rd Annual Meeting of Harvard School of Public Health Program in Quantitative Genomics
2006	Postdoctoral Basic Research Presentation Award, 56th Annual Meeting of The American Society of Human Genetics
2002-2004	Postdoctoral Fellowship, Ministry of Education, Spain.
2002	BA Biological Sciences with Honors, University of Chicago.

Publications

1. Li, Q., A. Stram, C. Chen, C. Haiman, **B.E. Stranger**, P. Kraft, M.L. Freedman. Expression QTL based analyses reveal candidate causal genes and loci across five tumor types, *in revision*
2. Raj, T., M. Kuchroo, J.M. Replogle, S. Raychaudhuri, **B.E. Stranger*,** P.L. De Jager*. *Cis*-regulatory regions influencing inflammatory disease are targets of recent positive selection, American Journal of Human Genetics, 92:1-13.
3. Li, Q., J.-H. Seo, **B.E. Stranger**, A. McKenna, I. Pe'er, T. Laframboise, M. Brown, S. Tyekucheva, M. L. Freedman. Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci, Cell, 152(3):633-41.
4. Dimas, A.S., A.C. Nica, S.B. Montgomery, **B.E. Stranger**, T. Raj, A. Buil, T. Giger, T. Lappalainen, M. Gutierrez-Arcelus, MuTHER Consortium, M.I. McCarthy, E.T. Dermitzakis. Sex-biased genetic effects on gene regulation in humans, Genome Research**,** 22(12): 2368-75.
5. **Stranger, B.E.,** S.B. Montgomery, A.S. Dimas, L. Parts, O. Stegle, C.E. Ingle, M. Sekowska, G. Davey Smith, D. Evans, M. Gutierrez-Arcelus, A.L. Price, T. Raj, J. Nisbett, A. Nica, C. Beazley, R. Durbin, P. Deloukas, E.T. Dermitzakis. 2012. Patterns of *cis* regulatory variation in diverse human populations, PLoS Genetics, 8(4): e1002639.
6. Raj, T., B. Keenan, J. Shulman, **B.E. Stranger**, and P.L. DeJager. 2012. Alzheimer's disease susceptibility loci: evidence for natural selection and altered gene expression, American Journal of Human Genetics**,** 90(4): 720-726.
7. Dimas, A.S.*, S. Deutsch*, **B.E. Stranger***, S.B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez-Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E.T. Dermitzakis, S.E. Antonarakis. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246-1250.
8. **Stranger, B.E.,** M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, R. Redon, C.P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S.W. Scherer, S. Tavaré, P. Deloukas, M.E. Hurles, E.T. Dermitzakis. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.
9. **Stranger, B.E.**, A.C. Nica, M.S. Forrest, A. Dimas, C.P. Bird, C. Beazley, C.E. Ingle, M. Dunning, P. Flicek, S. Montgomery, S. Tavaré, P. Deloukas, E.T. Dermitzakis. 2007. Population genomics of human gene expression. Nature Genetics 39: 1217-1224.
10. **Stranger, B.E**, M.S. Forrest, A.G. Clark, M.J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S.E. Antonarakis, S. Tavaré, P. Deloukas, E.T. Dermitzakis. 2005. Genome-wide associations of gene expression variation in humans. PLoS Genetics 1:e78

Zhenyu Zhang

Education

Fudan University, China, BS Biochemistry, 1998
Fudan University, China, MS Genetics, 2001
University of Texas at Austin, MS Statistics, 2007
University of Texas at Austin, PhD Cell and Molecular Biology, 2008
University of Texas at Austin, Postdoc, 2009
University of Chicago, Postdoc, 2010-2013

Positions

2013 – Bioinformatician, Center for Data Intensive Science, University of Chicago

Achivements

- Statistically predicted Rb1-drug interaction in human cancers and validated result in cancer cell lines.
- Elucidated modes of epigenetic regulation with RNA sequencing study of differential expression of orthologous genes in mouse-rat fusion cells.
- Discovered different functional specificities associated with ten isoforms of sulfotransferase Pipe and its enzymatic target in Drosophila embryonic pattern formation.

Publications

1. An occludome map of mouse fibroblasts reveals prevalence and stability of cis-silenced genes. Looney T, Lee J, Zhang L, Chen C, Zhang Z, and et al. Submitted.
2. Mutation of the Retinoblastoma tumor suppressor gene sensitizes cancers to mitotic inhibitor induced cell death. Zhang Z, Zhao J, Liao Y, Du W. American J or Cancer Research. Accepted.
3. Localization and activation of the Drosophila protease easter require the ER-resident saposin-like protein seele. Stein D, Charatsi I, Cho YS, Zhang Z, Nguyen J, DeLotto R, Luschnig S, Moussian B. Curr Biol. 2010 Nov 9;20(21):1953-8. Epub 2010 Oct 21.
4. Distinct functional specificities are associated with protein isoforms encoded by the Drosophila Dorsal-Ventral patterning gene pipe. Zhang Z, Zhu X, Stevens LM, Stein D. Development. 2009 Aug;136(16):2779-89.
5. Sulfation of Eggshell Components by Pipe Defines Dorsal-Ventral Polarity in the Drosophila Embryo. Zhang Z, Stevens LM, Stein D. Currt Biol. 2009 Jul 28;19(14):1200-5.
6. No requirement for localized Nudel protein expression in Drosophila embryonic axis determination. Stein D, Suk Cho Y, Zhang Z, Stevens LM. Fly. 2008 Jul 15; 2(4).
7. Functional characterization and expression analysis of members of the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase family from Drosophila melanogaster. Ten Hagen KG, Tran DT, Gerken TA, Stein DS, Zhang Z. J Biol Chem. 2003 Sep 12; 278(37): 35039-48.
8. A recombinant fusion protein and DNA vaccines against foot-and-mouth disease virus type Asia 1 infection in guinea pigs. Zhang Q, Zhu M, Yang Y, Shao M, Zhang Z, Lan H, Yan W, Wu J, Zheng Z. Acta Virol. 2003; 47(4): 237-43.
9. Immunogenicity of a recombinant fusion protein of tandem repeat epitopes of foot-and-mouth disease virus type Asia 1 for guinea pigs. Zhang Q, Yang Y, Zhang Z, Li L, Yan W, Jiang W, Xin A, Lei C, Zheng Z. Acta Virol. 2002; 46(1): 1-9.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Molecular correlates of kataegis, including structural variants involving *TERT*

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Chad Creighton, Baylor College of Medicine

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

I currently co-chair the kidney chromophobe (ChRCC) project for TCGA (manuscript close to submission). We have found that a subset of our ChRCC cases contained genomic rearrangements leading to breakpoints in the *TERT* promoter region, which correlated with elevated *TERT* expression and manifestation of kataegis. Whole genome sequencing (WGS) for 50 of our 66 ChRCC cases was performed (60X and 30X coverage for tumor and normal pairs, respectively). By WGS analysis, a subset of ChRCC analyzed by WGS manifested kataegis, a phenomenon involving highly localized substitution mutations (C>T or C>G or both). Consistent with observations in other cancers, we found that regions of kataegis in ChRCC were found in the vicinity of genomic rearrangements. We compared gene expression profiles between ChRCC cases with and without a strong kataegis pattern (n=3 and n=47, respectively), and identified 29 differentially expressed genes (FDR<0.05) including *TERT* (P<1E-10, t-test, FDR<1E-6). Subsequent WGS analysis identified genomic rearrangements involving the *TERT* promoter region, leading to breakpoints within the region in six out of 50 ChRCC cases; these cases also had the highest levels of *TERT* expression (P<1E-20, t-test), and three showed the strongest manifestation of kataegis (P=0.001, one-sided Fisher's exact). In five ChRCC cases, the TERT-associated rearrangements were intrachromosomal (one involving part of *PDCD6*), while the sixth case involved NEK5 on chromosome. Recently, point mutations in the TERT promoter, leading to up-regulation, have been uncovered in cancers such as melanoma, but this represents the first finding of recurrent breakpoints in the *TERT* promoter associated with cancer. (The above findings were contributed by Caleb Davis, Chad Creighton, Peter Park, and others in the ChRCC Analysis Working group. Structural variants identified using a combination of Breakdancer and Meerkat algorithms.)

Timelines & resources dedicated to project

1. Execution of Meerkat/Breakdancer or similar algorithm for detection of structural variants, focusing on the genomic region surrounding *TERT*. For ChRCC, Peter Park and colleagues had carried out Meerkat analysis, with computing resources facilitated by Sang-Cheol Kim and others in South Korea.
2. Examine WGS mutation calls for known activating mutations in *TERT* promoter (C228T and C250T). Coverage may be inadequate for some cases, however. Where possible, consider working with BCM-HGSC to use PCR primers to target the promoter region (currently being carried out for ChRCC as well as for other projects); only cases that show high *TERT* mRNA expression (where data available) would need to be probed.
3. Use the WGS mutation annotation file (MAF) to identify samples with clear manifestation of kataegis. These cases can be identified using rainfall plots, as well as using one-sided Fisher's exact tests (within a given pter/qter region, for enrichment of C>T or C>G mutations involving inter-mutation distances below 10kb).
4. Where mRNA expression data are available (e.g. for TCGA cases), compare WGS cases with or without kataegis patterns, to identify a robust gene signature of kataegis. The hypothesis is that the signature will include *TERT*, but include other genes that would be a basis for discovery.

Research proposal

The proposed research will carry out the proposed steps outlined above, in order to determine whether the novel associations involving *TERT* and kataegis, identified by TCGA for kidney chromophobe may apply to other cancers. Genomic rearrangements within the *TERT* promoter region represent a novel mechanism by which *TERT* may be activated, which possibility may be explored in other cancers. It may be that other genes in other cancers would also be implicated. APOBEC is understood to have a role in kataegis, but other genes may be involved as well. We may also explore with others the possibility of carrying out functional experiments, to determine whether over-expressing *TERT* or APOBEC would lead to induction of kataegis in cells; this would involving partnering with a molecular biologist (the PI knowing several) as well as with the BCM-HGSC (of which the PI is a member). The PI (Chad Creighton) has extensive experience in Team Science and leadership on big genomics projects, including working as part of TCGA on several of their marker papers in Nature.

Legacy plans

Chad J. Creighton

A. Education

University of Idaho, Moscow, ID B.S. 1992-1996 Physics
University of Michigan, Ann Arbor, MI Ph.D. 2001-2006 Bioinformatics

B. Personal Statement

My focus is on bioinformatics analysis of gene expression, microRNA expression, and DNA copy number. Fundamentally, my work seeks to obtain meaningful information from large scale molecular datasets, on questions relevant to improving cancer diagnosis and treatment. This often involves integration of molecular profiling results from different sources, such as linking mRNA expression with DNA copy number changes, or correlating the expression patterns observed in experimental models with the corresponding patterns observed in human tumor specimens. I have a successful track record in collaborating with various investigators who want help in getting the most out of their data. Prior to my graduate training, I first spent two years working as a wet bench laboratory technician in microbial genetics, followed by three years as a software engineer. I carried out my Ph.D. training under Drs. Sam Hanash and Arul Chinnaiyan, investigators well-established in the fields of genomics and proteomics. My diverse experiences have helped me to bridge both fields of molecular biology and applied statistics. Among other things, I participate extensively in The Cancer Genome Atlas (TCGA) consortium, a large-scale effort to systematically characterize the genomic changes that occur in cancer. With TCGA, I have been involved in aspects involving data integration and pathway analysis. For several TCGA papers published or in review in Nature, I have contributed display items and have served on the writing committees. I have acted in the capacity of project-wide Analysis Coordinator for TCGA clear cell kidney project and as Data Coordinator for TCGA bladder project, and I currently co-chair TCGA's kidney chromophobe project.

C. Positions and Honors

1996-1998 Research Associate, Department of Microbiology, Molecular Biology, and Biochemistry, University of Idaho, Moscow, ID
1998-1999 Software Engineer, Pacific Simulation, Inc., Moscow, ID
1999-2001 Software Development Team Leader, Pacific Simulation, Inc., Moscow, ID
2006-2011 Assistant Professor, Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX
Associate Professor (tenured, 2013), Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX

D. Selected Peer-reviewed Publications (Selected from over 130 papers)

1. The Cancer Genome Atlas Network (including **Creighton CJ**). "Comprehensive molecular characterization of clear cell renal cell carcinoma." Nature. 499(7456):43-9, 2013. (PMID:23792563) (Project-wide Analysis Coordinator and project-wide Manuscript Coordinator. Contributed main Figures 5 and 3d.)
2. The Cancer Genome Atlas Research Network (including **Creighton CJ**). "Comprehensive molecular portraits of human breast tumours." Nature. 490(7418):61-70, 2012. (Contributed main figures 3 and 5b) (PMID:23000897, PMCID:PMC3465532)
3. The Cancer Genome Atlas Research Network (including **Creighton CJ**). "Integrated genomic analyses of ovarian carcinoma." Nature. 29;474:609-15, 2011. (Contributed main figures 2b and 2c) (PMID: 21720365, PMCID: PMC3163504)
4. **Creighton CJ***, Fountain MD*, Yu Z*, Nagaraja AK, Zhu H, Khan M, Olokpa E, Zariff A, Gunaratne PH, Matzuk MM, Anderson ML. "Molecular profiling uncovers a p53-associated role for microRNA-31 in inhibiting the proliferation of serous ovarian carcinomas and other cancers." Cancer Res. 70(5):1906-15, 2010. (* = equal contributors) (PMID: 20179198, PMCID: PMC2831102)
5. **Creighton CJ***, Li X*, Landis M*, Dixon JM, Neumeister VM, Sjolund A, Rimm DL, Wong H, Rodriguez A, Herschkowitz JI, Fan C, Zhang X, He X, Pavlick A, Gutierrez MC, Renshaw L, Larionov AA, Faratian D, Hilsenbeck SG, Perou CM, Lewis MT, Rosen JM, Chang JC. "Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features." Proc Natl Acad Sci U S A. 106(33):13820-5, 2009. (* = equal contributors) (PMID: 19666588, PMCID: PMC2720409)
6. **Creighton CJ***, Benham AL*, Zhu H*, Khan MF, Reid JG, Nagaraja AK, Fountain MD, Dziadek O, Han D, Ma L, Kim J, Hawkins SM, Anderson ML, Matzuk MM, Gunaratne PH. "Discovery of novel microRNAs in female reproductive tract using Next Generation Sequencing." PLOS One. 5(3):e9637, 2010. (* = equal contributors) (PMID: 20224791, PMCID: PMC2835764)
7. **Creighton CJ**, Osborne CK, van de Vijver MJ, Foekens JA, Klijn JG, Horlings HM, Nuyten D, Wang Y, Zhang Y, Chamness GC, Hilsenbeck SG, Lee AV, Schiff R. "Molecular profiles of progesterone receptor loss in human breast tumors." Breast Cancer Res Treat. 114(2):287-99, 2009. (PMID: 18425577, PMCID: PMC2635926)
8. Castro P*, **Creighton CJ***, Ozen M, Berel D, Mims M, Ittmann M. "Genomic profiling of prostate cancers from African-American men." Neoplasia. 11(3):305-12, 2009. (* = equal contributors) (PMID: 19242612, PMCID: PMC2647733)
9. Gibbons DL, Lin W*, **Creighton CJ***, Rizvi Z, Gregory PA, Goodall GJ, Thilaganathan N, Du L, Zhang Y, Pertsemlidis A, Kurie JM. "Microenvironmental cues promote tumor cell EMT and metastasis by regulating miR-200 family expression." Genes Dev. 23(18):2140-51, 2009. (* = equal contributors) (PMID: 19759262, PMCID: PMC2751985)
10. Gibbons DL*, Lin W*, **Creighton CJ***, Zheng S, Berel D, Yang Y, Raso MG, Liu DD, Wistuba II, Lozano G, Kurie JM. "Expression Signatures of Metastatic Capacity in a Genetic Mouse Model of Lung Adenocarcinoma." PLoS One. 4(4):e5401, 2009 (* = equal contributors) (PMID: 19404390, PMCID: PMC2671160)
11. **Creighton CJ**, Nagaraja AK, Hanash SM, Matzuk MM, Gunaratne PH. "A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions." RNA. 14(11):2290-6, 2008. (PMID: 18812437, PMCID: PMC2578856)
12. **Creighton CJ**, Casa A, Lazard Z, Huang S, Tsimelzon A, Hilsenbeck SG, Osborne CK, Lee AV. "Insulin-like growth factor I (IGF-I) activates gene transcription programs strongly associated with poor breast cancer prognosis." J Clin Oncol. 26(25):4078-85, 2008. (PMID: 18757322, PMCID: PMC2654368)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mitochondrial DNA mutations and their impact on gene expression

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Chad Creighton, Baylor College of Medicine

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

I currently co-chair the kidney chromophobe (ChRCC) project for TCGA (manuscript close to submission). Contrary to what may be expected in other cancers, ChRCC apparently increase their utilization of mitochondria and associated metabolic pathways. When viewed in the context of mitochondrial function, gene expression in ChRCC, as compared to normal kidney, suggested increased utilization of the Krebs cycle and electron transport chain (ETC) for adenosine triphosphate (ATP) generation. In ChRCC, nearly all enzymes in the Krebs cycle showed increased gene expression, with the entry of pyruvate into the Krebs cycle via Acetyl CoA likely through the pyruvate dehydrogenase complex (PDC). Concordantly, all complexes of the ETC demonstrated increases in at least one gene. These patterns could reflect an increased level of mitochondrial biosynthesis, resulting in greater numbers of mitochondria within each tumor cell; this possibility is supported by both the increased expression of mitochondrial biogenesis regulator PPARGC1A (P<1E-5, t-test), and increased mitochondrial genome copy numbers (by mtDNA analysis). In general, the gene expression landscape appeared very different from that of clear cell kidney cancer (CCRCC), where the genes involved in mitochondria functions are strongly suppressed. These findings suggest that various bioenergetics strategies may support tumor growth, and that not all cancers would necessarily seek to minimize their reliance upon oxidative phosphorylation.

Given the indicated prevalent role of mitochondria in ChRCC and the likelihood of rapid mitochondrial genome replication, we sequenced mtDNA from 61 of our 66 ChRCC cases, using a PCR-based amplification approach. In all, we identified 142 somatic mutation events at various levels of heteroplasmy (i.e. mixture of somatic and germline), 75 of these residing within the commonly altered D-Loop non-coding region. Thirty-five mutation events (involving 27 cases) were present in over 50% of mtDNA copies in the tumor (>50% heteroplasmy). Human mtDNA encodes 13 proteins involved in respiration and oxidative phosphorylation, and we found 15 nonsilent mutations in 12 ChRCC cases involving these genes (>50% heteroplasmy), all of which validated using alternative strategies, including alternative PCR and sequencing and analysis of Whole Genome data (from Lynda Chin's group). ETC Complex I genes were altered in 18% of cases (n=11); the most frequently altered gene was *MT-ND5*, in six cases (all with >70% heteroplasmy), with five of these being histologically classified as eosinophilic ChRCC (P<0.01, one-sided Fisher's exact test). We also found *MT-ND5*-mutated ChRCC cases to have a distinct gene transcription signature (724 genes with P<0.001, False Discovery Rate, or FDR<0.05). Genes high in *MT-ND5*-mutated cases were enriched for those associated with mitochondria (43 with Gene Ontology term "mitochondrion", P<5E-6, one-sided Fisher's exact test), including several with roles in ETC (*SDHB*, *NDUFS1*, *ATP5F1*, *COX10*, *COX11*). Notably, mutations in complex I did not result in expression patterns associated with loss of oxidative phosphorylation, as has been previously assumed, suggesting possible alternative roles for complex I alteration in cancer.

Timelines & resources dedicated to project

1. Using GATK Unityper algorithm (developed by others, PMID:22891333), examine all WGS cases for mitochondrial DNA (mtDNA) mutations. For ChRCC, personnel in Lynda Chin's group carried out the analysis.
2. As has been done for kidney cancers (ChRCC and CCRCC), carry out Pan-Can analysis for differential expression of genes involved in core metabolic pathways, including glycolysis, Krebs cycle, oxidative phosphorylation, etc.
3. For commonly mutated mitochondria-encoded genes, define gene expression correlates. In particular, examine whether loss of complex I by mutation may be associated with alterations in oxidative phosphorylation in some cancer subtypes.

Research proposal
The proposed research will carry out the proposed steps outlined above, in order to determine how mtDNA mutations may be associated with metabolic changes. For quite some time, mtDNA mutations have been extensively studied in human cancers, though little if anything has been done to correlate these mutations with gene expression data. Furthermore, many early studies did little to characterize the levels of heteroplasmy of detected mutations (where mutations may be present in a few as 5% of mtDNA copies in the cancer cell). As evidenced by the literature, interest in characterizing mtDNA mutations apparently waned, after the completion of the human genome and the advent of DNA microarrays (with tens of thousands of genes being presented for potential characterization). However, with the rich, multi-platform genomic datasets provided by TCGA and ICGC, a more comprehensive analysis of mtDNA mutations and their effects on pathways is now possible. The PI (Chad Creighton) has extensive experience in Team Science and leadership on big genomics projects, including working as part of TCGA on several of their marker papers in Nature. Many in the mitochondria research community have been somewhat wary of the use of WGS to characterize mtDNA mutations; however, in the ChRCC projects, samples were analyzed using both WGS and the conventionally accepted long-range PCR methods, and the two sets of results were highly concordant with each other, helping to pave the way for the future use of WGS data. The PI also has extensive experience with pathway analysis of molecular profiling data, as well as having worked with experts in metabolic pathways, including Marston Linehan and Christopher Ricketts (key participants in the TCGA kidney projects).
Legacy plans

Chad J. Creighton

A. Education

University of Idaho, Moscow, ID B.S. 1992-1996 Physics
University of Michigan, Ann Arbor, MI Ph.D. 2001-2006 Bioinformatics

B. Personal Statement

My focus is on bioinformatics analysis of gene expression, microRNA expression, and DNA copy number. Fundamentally, my work seeks to obtain meaningful information from large scale molecular datasets, on questions relevant to improving cancer diagnosis and treatment. This often involves integration of molecular profiling results from different sources, such as linking mRNA expression with DNA copy number changes, or correlating the expression patterns observed in experimental models with the corresponding patterns observed in human tumor specimens. I have a successful track record in collaborating with various investigators who want help in getting the most out of their data. Prior to my graduate training, I first spent two years working as a wet bench laboratory technician in microbial genetics, followed by three years as a software engineer. I carried out my Ph.D. training under Drs. Sam Hanash and Arul Chinnaiyan, investigators well-established in the fields of genomics and proteomics. My diverse experiences have helped me to bridge both fields of molecular biology and applied statistics. Among other things, I participate extensively in The Cancer Genome Atlas (TCGA) consortium, a large-scale effort to systematically characterize the genomic changes that occur in cancer. With TCGA, I have been involved in aspects involving data integration and pathway analysis. For several TCGA papers published or in review in Nature, I have contributed display items and have served on the writing committees. I have acted in the capacity of project-wide Analysis Coordinator for TCGA clear cell kidney project and as Data Coordinator for TCGA bladder project, and I currently co-chair TCGA's kidney chromophobe project.

C. Positions and Honors

1996-1998 Research Associate, Department of Microbiology, Molecular Biology, and Biochemistry, University of Idaho, Moscow, ID
1998-1999 Software Engineer, Pacific Simulation, Inc., Moscow, ID
1999-2001 Software Development Team Leader, Pacific Simulation, Inc., Moscow, ID
2006-2011 Assistant Professor, Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX
Associate Professor (tenured, 2013), Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX

D. Selected Peer-reviewed Publications (Selected from over 130 papers)

1. The Cancer Genome Atlas Network (including **Creighton CJ**). "Comprehensive molecular characterization of clear cell renal cell carcinoma." Nature. 499(7456):43-9, 2013. (PMID:23792563) (Project-wide Analysis Coordinator and project-wide Manuscript Coordinator. Contributed main Figures 5 and 3d.)
2. The Cancer Genome Atlas Research Network (including **Creighton CJ**). "Comprehensive molecular portraits of human breast tumours." Nature. 490(7418):61-70, 2012. (Contributed main figures 3 and 5b) (PMID:23000897, PMCID:PMC3465532)
3. The Cancer Genome Atlas Research Network (including **Creighton CJ**). "Integrated genomic analyses of ovarian carcinoma." Nature. 29;474:609-15, 2011. (Contributed main figures 2b and 2c) (PMID: 21720365, PMCID: PMC3163504)
4. **Creighton CJ***, Fountain MD*, Yu Z*, Nagaraja AK, Zhu H, Khan M, Olokpa E, Zariff A, Gunaratne PH, Matzuk MM, Anderson ML. "Molecular profiling uncovers a p53-associated role for microRNA-31 in inhibiting the proliferation of serous ovarian carcinomas and other cancers." Cancer Res. 70(5):1906-15, 2010. (* = equal contributors) (PMID: 20179198, PMCID: PMC2831102)
5. **Creighton CJ***, Li X*, Landis M*, Dixon JM, Neumeister VM, Sjolund A, Rimm DL, Wong H, Rodriguez A, Herschkowitz JI, Fan C, Zhang X, He X, Pavlick A, Gutierrez MC, Renshaw L, Larionov AA, Faratian D, Hilsenbeck SG, Perou CM, Lewis MT, Rosen JM, Chang JC. "Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features." Proc Natl Acad Sci U S A. 106(33):13820-5, 2009. (* = equal contributors) (PMID: 19666588, PMCID: PMC2720409)
6. **Creighton CJ***, Benham AL*, Zhu H*, Khan MF, Reid JG, Nagaraja AK, Fountain MD, Dziadek O, Han D, Ma L, Kim J, Hawkins SM, Anderson ML, Matzuk MM, Gunaratne PH. "Discovery of novel microRNAs in female reproductive tract using Next Generation Sequencing." PLOS One. 5(3):e9637, 2010. (* = equal contributors) (PMID: 20224791, PMCID: PMC2835764)
7. **Creighton CJ**, Osborne CK, van de Vijver MJ, Foekens JA, Klijn JG, Horlings HM, Nuyten D, Wang Y, Zhang Y, Chamness GC, Hilsenbeck SG, Lee AV, Schiff R. "Molecular profiles of progesterone receptor loss in human breast tumors." Breast Cancer Res Treat. 114(2):287-99, 2009. (PMID: 18425577, PMCID: PMC2635926)
8. Castro P*, **Creighton CJ***, Ozen M, Berel D, Mims M, Ittmann M. "Genomic profiling of prostate cancers from African-American men." Neoplasia. 11(3):305-12, 2009. (* = equal contributors) (PMID: 19242612, PMCID: PMC2647733)
9. Gibbons DL, Lin W*, **Creighton CJ***, Rizvi Z, Gregory PA, Goodall GJ, Thilaganathan N, Du L, Zhang Y, Pertsemlidis A, Kurie JM. "Microenvironmental cues promote tumor cell EMT and metastasis by regulating miR-200 family expression." Genes Dev. 23(18):2140-51, 2009. (* = equal contributors) (PMID: 19759262, PMCID: PMC2751985)
10. Gibbons DL*, Lin W*, **Creighton CJ***, Zheng S, Berel D, Yang Y, Raso MG, Liu DD, Wistuba II, Lozano G, Kurie JM. "Expression Signatures of Metastatic Capacity in a Genetic Mouse Model of Lung Adenocarcinoma." PLoS One. 4(4):e5401, 2009 (* = equal contributors) (PMID: 19404390, PMCID: PMC2671160)
11. **Creighton CJ**, Nagaraja AK, Hanash SM, Matzuk MM, Gunaratne PH. "A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions." RNA. 14(11):2290-6, 2008. (PMID: 18812437, PMCID: PMC2578856)
12. **Creighton CJ**, Casa A, Lazard Z, Huang S, Tsimelzon A, Hilsenbeck SG, Osborne CK, Lee AV. "Insulin-like growth factor I (IGF-I) activates gene transcription programs strongly associated with poor breast cancer prognosis." J Clin Oncol. 26(25):4078-85, 2008. (PMID: 18757322, PMCID: PMC2654368)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Clonal architecture, evolution, and diversity of pan cancer from whole-genome sequencing data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Song Cao, Kai Ye
The Genome Institute, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Clonal architecture, diversity, and evolution play important roles in cancer progression and metastasis (1,2, 3). The currently available next-generation sequencing data provide an unprecedented opportunity to study cancer clonality at the DNA level based on somatic variant allele fraction (VAF).

In a 2012 paper on *Acute myeloid leukemia* (AML) (1), the PI and collaborators reported clonal evolution from the primary tumor to relapse tumor. It was observed that a subclone in the primary tumor gained additional mutations, eventually evolving into the relapse clone. In a more recent pan-cancer paper (3), we investigated clonality across 12 major cancer types using exome sequencing data. Clonality analysis can help identify founding clone/subclone(s) information in different cancer types and can establish the link between clonal structure, tumor progression, and treatment response.

[1] Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, DiPersio JF, Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing, Nature, 2012, 481, 506-510

[2] Walter MJ, Shen D, Ding L, Shao J, Koboldt DC, Chen K, Larson DE, McLellan MD, Dooling D, Abbott R, Fulton R, Magrini V, Schmidt H, Kalicki-Veizer J, O'Laughlin M, Fan X, Grillot M, Witowski S, Heath S, Frater JL, Eades W, Tomasson M, Westervelt P, DiPersio JF, Link DC, Mardis ER, Ley TJ, Wilson RK, Graubert TA, N. Engl. J. Med., 2012, 22, 1090-1098.

[3] Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L, Mutational landscape and significance across 12 major cancer types, Nature, 2013, 502, 333-339.

Timelines & resources dedicated to project

1) Make tools for clonality analysis available and access/retrieve whole-genome sequencing data in a cloud environment by March 2014.

2) Participate genomic variant calling group to obtain variant allele frequency of somatic variants from the whole-genome sequencing data from ICGC by July 2014.

3) Perform clonality analysis for different cancer types and study the association with clinical outcome by December 2014.

4) Prepare and submit manuscripts for the results of clonality analysis by March 2015.

Research proposal

With the available whole genome sequencing (WGS) data of ~2,000 cancer samples across more than 20 cancer types from ICGC, we will systematically analyze the clonal architecture, evolution, and diversity across cancer types. Whole genome data can provide somatic variant information in both coding and non-coding regions, which will greatly extend the current pan-cancer study primarily based on exome sequencing. We will extract variant allele fractions for somatic variants detected in coding/non-coding regions and use the SciClone algorithm for clonality analysis. We expect significant findings in the following three categories:

1) Discovery and characterization of clonal architecture and diversity across cancer types.

2) Correlation of clonal growth and expansion patterns with specific types/subtypes of cancer.

3) Analysis of clonal evolution from primary tumor to metastasis and the association with the clinical outcome.

Undoubtedly, the additional whole genome sequencing data of ~2,000 cancer samples from ICGC will advance our understanding of cancer initiation and progression. The clonality analysis we propose aims to generate a complete picture of clonal architecture, diversity, and evolution across cancer types, which will eventually help clinicians select effective therapeutic targets for cancer interventions.

Legacy plans

We will make the clonality analysis tools available for general use by creating a website or uploading them to the existing public accessible website.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L.** VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L.** Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L. Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L.** MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res.*** 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L.** Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

CURRICULUM VITAE

Song Cao

Staff Scientist, The Genome Institute
Washington University School of Medicine, St. Louis MO

Cell phone: 314-358-4740
Email: songcao@gmail.com

Specialties

Pipeline development
Cancer genomics and cancer proteomics
Virus discovery and metagenomics
RNA secondary/three-dimensional structure prediction
miRNA target site prediction
Algorithms (Recursive, Dynamic, DFS (depth-first search), BFS (breadth-first search), etc.)
Data structure (Tree, Graph, Linked list, Stack, Queue, Hash, etc.)
Linux/Unix (>10 yrs); High performance cluster computing (>10 yrs)
Programming (Perl (>4 yrs), C/C++ (>10 yrs), HTML (>4 yrs), Java (>1 yr), Javascript (>1 yr)
, Shell script (> 2 yrs), R (> 2 yrs), etc.)

Professional experience

10/2013-present: The Genome Institute at Washington University
Staff Scientist
Bioinformatics, Computational Biology, Cancer genomics, Cancer proteomics

09/2011-09/2013: Washington Universtiy School of Medicine
Bioinformaticist
Bioinformatics, Computational Biology, Virus Discovery, Metagenomics
1) Developed an efficient pipeline for detecting the novel viruses for the Illumina data
2) Wrote a Perl script for automatically generating summary reports for virus discovery
3) Analyzed 60 454 runs, 30 MiSeq runs and 10 HiSeq runs
4) Performed the Case-Control study of diarrhea from Gambia and Kenya
5) Detected several novel viruses from the sequencing data
6) Worked in different projects with postdoctoral fellows, faculty and independent PIs

06/2008-08/2011: University of Missouri-Columbia
Research Associate
Bioinformatics, Computational biology, Biophysics, RNA structure prediction
10/2003-06/2008: University of Missouri-Columbia
Postdoctoral fellow
Bioinformatics, Computational biology, Biophysics, RNA structure prediction

Education

1999-2003 Zhejiang University, Ph.D., Computational Physics
1995-1999 Harbin University of Science and Technology, B.E., Electronic material and device

Honors and awards

2003-2005 MU Life Sciences Postdoctoral Fellowship, USA
2002 National Academia Sinica award for Excellent Graduate Students, China
2001 FERROTEC award for Excellent Graduate Students, China

The Genome Institute at Washington University in St. Louis

Kai Ye, PhD

Positions and degrees

- *2012-now*: Research assistant professor, the genome institute at Washington University in St. Louis
- *2009-2012*: Assistant professor, Leiden University Medical Center, the Netherlands
- *2008-2009*: Postdoc European Bioinformatics Institute, UK
- *2004-2008*:**PhD.** *Cum Laude* Biopharmaceutical science at Leiden University, The Netherlands
  Novel algorithms for protein sequence analysis
- *Aug. 2003-Dec. 2003*: Lecturer in the college of Pharmacy at Wuhan University, China.
- *Sept. 1995-Jul. 2003*: **B.Sc.** and **M.Sc.** of Biopharmaceutical science at Wuhan University, China.

Awards and grants

- June 2009: 'best paper' (on Pindel) presented at the Short-SIG on Next-Generation Sequence and Algorithms for Short Read Analysis, ISMB/ECCB 2009 at Stockholm, Sweden. 500 GBP
- 'Researcher of 2008', Faculty of Science, Leiden University.
- 2008 C. J. Kok prize, Leiden University, the Netherlands. 2,500 EURO
- 2008 **PhD Cum Laude**, Leiden, the Netherlands
- NGI/EBI fellowship, The Netherlands. 36,000 euro per year.
- 'Top 300 Outstanding PhD students abroad' Award, China. 5,000 USD
- *2004-2005* Leiden University Scholarship, The Netherlands.
- *2000-2003* Wuhan University fellowship for excellent graduate student, China.
- *1996-1999* Wuhan University fellowship for excellent undergraduate, China.

Programming experience (and related)

- Good programming experience in C/C++ and MatLab;
- Familiar with Perl, Python, PHP, MySQL and linux
- Good experience in using InsightII, AutoDock, PyMol, SPDBV and WHAT IF

Research interests

- Large scale data mining and sequence analysis: protein, DNA
- Pharmacology modeling: mathematical modeling of receptor-receptor, receptor-ligand interaction
- Homology modeling, MD simulation and docking
- Cancer genomics

kye@genome.wustl.edu

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

The landscape of microsatellite instability in pan-cancer genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Beifang Niu, Mingchao Xie
The Genome Institute, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Microsatellites are present at millions of sites in the human genome and have various lengths due to impaired DNA mismatch repair. Microsatellite instability (MSI) is the condition of genetic hypermutability which has been linked to many genetic diseases, including hereditary nonpolyposis colorectal cancer (Lynch syndrome). Tumor MSI status is also a predictor of benefit in other cancer types, such as colorectal more broadly and endometrial cancers.

Levels of MSI are traditionally determined experimentally. The sizes of microsatellite marker sets in tumor DNA are compared via electrophoresis with corresponding DNA isolated from a normal tissue sample of the same patient. The most widely used set of markers was recommended by a National Cancer Institute (NCI) consensus group and consists of either 5 or 7 repeat markers (1). Samples are normally classified as microsatellite instability high (MSI-H), microsatellite instability low (MSI-L) and microsatellite stable (MSS). While the experimental approach is considered the gold standard, its detection procedure is expensive and its scope is limited to only a small subset of microsatellites.

Conversely, paired tumor-normal genome sequencing allows for comprehensive investigation of MSI sites simultaneously and will likely become a routine part of diagnosis and treatment procedures. Several recent research papers show movement in this direction. For example, Lu et al. reported that MSI status can be derived from RNA-seq data (2). Recently, Tae-Min Kim et al. published the landscape of MSI in colorectal and endometrial cancer genomes (3). However, they provide neither a standalone tool required for quantifying MSI in various paired tumor-normal genome sequencing data nor do they report somatic status of corresponding microsatellite sites in the human genome. These aspects are urgently needed in cancer research.

The latest DNA sequencing technologies have led to a paradigm-shift in cancer genomics, from the sequencing and analysis of a single tumor sample to that of thousands of samples from many tumor types. With the pan-cancer whole genome sequencing, exome sequencing, and RNA sequencing data set of more than 2000 ICGC cancer samples across 50 cancer types, we aim to provide a standalone and cloud compatible MSI detecting tool and pto resent a comprehensive genome-wide analysis of the MSI in pan-cancer genomes.

[1] Vasen, H.F., et al. (1999) New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC, Gastroenterology, 116, 1453-1456.
[2] Lu, Y., et al. (2013) Anovel Approach for characterizing microsatellite instability in cancer cells, PLoS ONE, 8, e63056
[3] Tae-Min Kim, Peter W. Laird and Peter J. Park. (2013) The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes, Volume 155, Issue 4, 7 November 2013, Pages 858–868.

Timelines & resources dedicated to project

1) March 2014: Standalone MSI detection tool development.

2) June 2014: Release standalone and cloud compatiable tool.

3) Sept 2014: MSI analysis on cloud using WGS/WXS/RNA-seq data across pan-cancer.

4) Oct 2014: Various correlation analyses between MSI sites and mutation rate, clinical data.

5) Dec 2014: Publish the landscape of MSI in individual cancer and pan-cancer.

Research proposal

The International Cancer Genome Consortium (ICGC) has generated various kinds of sequence data for about 2,000 samples from 50 different cancer types of socio-clinical importance across the globe. These rich data offer the chance to do systematic and comprehensive pan-cancer MSI analysis. First, we will develop an MSI detection tool that can be used for automatically detecting somatic microsatellite changes. We will then use this tool to scan various sequencing data in ICGC. We will likely release a cloud compatible version and process ICGC data on the cloud. Finally, we will integrate various correlations among MSI sites with mutation rate and clinical data to investigate whether MSI sites portend improved survival in multiple cancer types. The significance of this project consists of the following two parts:

a). Provide the community with an open source software tool that can be applied to ICGC data for MSI screening with the anticipation of a wider usage of this tool in cancer clinical sequencing.
b). Generate a comprehensive landscape of MSI in individual cancers and pan-cancer data of ICGC.

Legacy plans

All software developed as part of this project will be made publicly available both during development and subsequent to project completion on our laboratory's GitHub repository.

Li Ding, Ph.D.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L**. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L**, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L**. MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res***. 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L**. Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

BEIFANG NIU

Staff Scientist
The Genome Institute & School of Medicine
Washington University, St. Louis
Box 1910
St. Louis, MO, 63108

Email: bniu@wugsc.genome.edu & beifang.cn@gmail.com

RESEARCH INTERESTS

Cancer genomics and Metagenomics. Bioinformatics software development and algorithm optimization. Large scale bioinformatics data analysis and sequence clustering, alignment and assembly algorithm for next generation sequencing data. High performance parallel computing and cloud computing.

EDUCATION

2002 – 2009 **Ph.D.** in Computer Software and Theory, Super Computing Center, Chinese Academy of Sciences, Beijing, China.
Thesis Title: *Research on Short Oligonucleotide Alignment and Assembly Algorithm.* Advisor: Professor Xuebin Chi.

1998 – 2002 **S.B.** in Computer Science, Shandong Agriculture University, Shandong, China

EXPERIENCE

2012 – present **Staff Scientist**, (Cancer Genomics), The Genome Institute, School of Medicine, Washington University in St. Louis, Missouri, US.
Sponsor: Professor Li Ding.

2009 – 2012 **Postdoctoral Associate**, (Bioinformatics), The Center for Research in Biological Systems, University of California, San Diego, California, US.
Sponsor: Doctor Weizhong Li.

2007 – 2008 **Research Intern**, (Bioinformatics), Beijing Genomics Institute (BGI), Beijing, China.
Sponsor: Professor Jun Wang

2003 – 2007 **Research Assistant**, (Parallel Computing), Super Computing Center, Chinese Academy of Sciences, Beijing, China.
Sponsor: Professor Xuebin Chi

MINGCHAO XIE

Email: mxie@genome.wustl.edu
Tel: 314-737-3288

Education

08/2012-present	**Washington University in St Louis, MO** Graduate Student in Computational and Systems Biology
08/2008-08/2010	**Pennsylvania State University (University Park), PA** Master of Science in Biochemistry and Molecular Biology
09/2004-07/2007	**Tsinghua University, Beijing, China** Master of Science in Molecular Cell Biology
08/2000-07/2004	**Shandong University, Jinan China** Bachelor of Science in Biological Science

Research Experience

01/2013-present	**Research Assistant, the Genome Institute, Washington University in St Louis** Research advisor: Dr. Li Ding Thesis Project: "Cancer susceptibility variants study in large-scale sequence data"
09/2010-08/2012	**Research Analyst, Department of Genetics, Washington University** Supervisor: Dr. Ting Wang Develop sequencing-based DNA methylation analysis and transposable element analysis pipelines, and design epigenome browser (http://vizhub.wustl.edu/).
01/2009-08/2010	**Research Assistant, Department of Biochemistry and Molecular Biology, Pennsylvania State University** Research advisor: Dr. Kathleen Postle Thesis Project: "Study the assembly of ExbD transmembrane domains and identify the proton channel in the TonB/ExbB/ExbD system"
09/2004-07/2007	**Research Assistant, School of Medicine, Tsinghua University** Research advisor: Dr. Zhao Wang Thesis Project: "Study the evolution of SPATA4 gene and mechanism of its anti-apoptosis activity"

Awards

08/2008-07/2010	**Teaching Assistant Scholarship**, Pennsylvania State University
06/2009	**Braucher Scholarship**, Pennsylvania State University
09/2008	**Homer F. Braddock Fellowship**, Pennsylvania State University
05/2005	**Excellent Paper Award**, Annual Conference of Chinese Pharmacological Society

Selected Publications

- **Mingchao Xie,** Chibo Hong, Bo Zhang, *et al*. "DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape". *Nature Genetics*, 45:836–841 (2013)
- **Mingchao Xie**, Chao Ai, ShangFeng Liu, Xiumei Jin, Zhao Wang. "Cloning and characterization of chicken SPATA4 gene and analysis of its specific expression", *Molecular and Cellular Biochemistry*, 306: 79-85 (2007)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Analysis of germline variation across pan-cancer genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Jiayin Wang, Kim Johnson
The Genome Institute, Department of Genetics,
Washington University in St Louis, Missouri 63108

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Cancer is one of the most common causes of death worldwide, with fatalities in the US exceeding 577,000 Americans in 2012, i.e. 1,500 cancer deaths per day [1]. Cancer evolves through a wide array of dynamic genomic changes, including oncogene activation, tumor suppressor gene inactivation, and chromosome gains, losses, and rearrangements. In addition to somatic genomic changes, inherited germline variation also plays an important role in cancer induction and evolution [2]. However, the landscape of germline variation across multiple cancer types is still unknown.

We analyzed available TCGA whole genome and exome sequencing data for germline variants for multiple major cancer types, including ovarian cancer, breast cancer, renal cell carcinoma, and lower grade glioma. Although rare germline variants have been implicated in certain cancers, for example rare variants in ***BRCA1*** and ***BRCA2*** genes in early-onset breast and/or ovarian cancer, a large number of cancer cases with a strong hereditary components are not adequately explained by rare pathogenic variants in known cancer genes.

In order to efficiently discover both known and novel cancer predisposition genes and variants across the whole genome/exome space, we developed a germline variant calling pipeline that incorporates GATK and VarScan for SNV and short indel detections and Pindel for small and large indel detections. To date, we have analyzed four different cancer types (ovarian, breast, renal cell carcinoma (RCC), and low grade glioma (LGG)) for rare cancer predisposition variants in datasets of unrelated individuals. We recently finalized the analysis for the first large scale integrated analysis of germline and somatic variation from the exome data of 429 TCGA ovarian cancer cases (Kanchi et al., In Press ***Nature Communications***). Our analytic strategy was applied to millions of variants and resulted in the identification of a priority set of rare likely pathogenic variants in known ovarian cancer predisposition genes, as well as novel genes for further functional evaluation. In addition, integrated analysis of somatic and germline variation led to the identification of pathways that are enriched for rare likely pathogenic germline variants and somatic mutations. Based on the analysis of over 2,000 TCGA cancer cases, our preliminary results further support our ability to efficiently identify enrichment of rare germline truncation variants (nonsense, splice site, frame shift) using an unbiased analytic approach.

We are currently developing methods for improved analysis of truncation and missense germline variants that will efficiently nominate a set of rare candidate variants from among millions of observed variants for further downstream characterization. The studies described above, as well as analysis of additional cancer types that will be incorporated in the near future, will form the foundation for a pan-cancer analysis that will identify similarities and differences in cancer susceptibility variants and germline somatic interactions among tumor types.

1. American Cancer Society. Cancer Facts and Figures (2012).

2 Anand, P. et al. Cancer is a preventable disease that requires major lifestyle changes. Pharmaceutical Research 25, 2097-2116 (2008).

3. Towler, W. I. et al. Analysis of BRCA1 variants in double-strand break repair by homologous recombination and single-strand annealing. Hum Mutat 34, 439-445, doi:10.1002/humu.22251 (2013)

Timelines & resources dedicated to project

1) June 2014: Analysis germline variants across multiple cancer types using WGS and WXS sequencing data.
2) Sept 2014: Develop association analysis approach for germline/somatic variant association studies.
3) Oct 2014: Various correlation analyses on germline calls with clinical data.
4) Dec 2014: Publish the mutational landscape of germline variation across pan-cancer genomes.

Research proposal

The International Cancer Genome Consortium (ICGC) has generated various sequence data of about 2,000 samples from 50 different cancer types of clinical and societal importance across the globe. These rich data offer us a chance to apply comprehensive analysis on germline mutational events across pan-cancer.

The significance of this project consists of the following two parts:

a). Analysis results on germline causation variants across pan-cancer based on ICGC WGS and WXS sequencing data.

b). Development of new association approaches to identify novel associations between germline variants and clinical phenotypes.

Legacy plans

All software developed as part of this project will be made publicly available both during development and subsequent to project completion on our laboratory's GitHub repository.

Li Ding, Ph.D.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L**. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L**, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L**. MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res***. 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L**. Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

JIAYIN WANG

EDUCATION

- **Ph.D. in Computer Science and Engineering**
  September 2009 – May 2013
  University of Connecticut, Storrs
  Advisor: Yufeng Wu; associate advisors: Ion Mandiou and Jinbo Bi

- **B. S. in Computer Science and Technology**
  September 2004 - June 2008
  Xi'an Jiaotong University, China

WORKING EXPERIENCE

- **Postdoctoral Research Associate**
  June 2013 – Now
  The Genome Institute
  Washington University in St. Louis

- **Research Assistant, Computer Science and Engineering**
  September 2009 – May 2013
  University of Connecticut, Storrs

SELECT PUBLICATIONS

1. Xuanping Zhang, **Jiayin Wang**, Aiyuan Yang, Chunxia Yan, Feng Zhu, Zhongmeng Zhao, Zhi Cao, Identifying interacting genetic variations by fish-swarm logic regression, *BioMed Research International* **2013**, Article ID 574735, 11 pages.
2. **Jiayin Wang**, Zhongmeng Zhao, Zhi Cao, Aiyuan Yang, Jin Zhang, A probabilistic method for identifying rare variants underlying complex traits, *BMC Genomics* **14**, Suppl 1, 2013.
3. Jin Zhang, **Jiayin Wang**, Yufeng Wu, An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data, *BMC Bioinformatics* **13**, Suppl 6: S6, 2012. SCI(941CA), IF = 3.02
4. **Jiayin Wang**, Xuanping Zhang, Yanqin Liu, Jin Zhang, Yufeng Wu, A synchronization detection approach for identifying rare mutations underlying common disease, *5th International Conference on Bioinformatics and Computational Biology (BICoB 2013)*, Honolulu, HI, 2013.

CURRICULUM VITAE

Kimberly J. Johnson, M.P.H. Ph.D.

EDUCATION:

University of Minnesota Minneapolis, Minnesota	Epidemiology (major) Human Genetics (minor)	Ph.D.	2004-2007

CURRENT POSITION (2010-present):

Assistant Professor, Masters of Public Health Program, Brown School, Washington University in St. Louis, St. Louis, MO

GRANT FUNDING:

Alex's Lemonade Stand Foundation for Childhood Cancer 7/1/13-6/30/15
Principal Investigator
Title: Identification of risk factors for pediatric brain tumors in a high risk population

NIH R01-CA180006-01 2/1/2013-1/31/2017
Co-Investigator (PI-Li Ding, PhD)
Title: Cancer Susceptibility Variant Discovery in High Throughput Sequencing Data

NIH CTSA UL1 TR000448 ICTS JIT funding 1/7/2013-1/6/2014
Principal Investigator
Title: Pediatric Cancer Epidemiology in the Privately Insured

Siteman Research Development Award 8/1/12-7/31/13
Principal Investigator
Title: Genetic variation in folate metabolism genes and risk of pediatric brain cancer in the offspring

Alex's Lemonade Stand Foundation for Childhood Cancer 7/1/12-6/30/13
Principal Investigator
Title: Neurofibromatosis Type I as a model for pediatric bran cancer prevention and control research

American Cancer Society Institutional Research Grant 1/1/12-12/31/12
Principal Investigator
Title: Inherited tumor predisposition syndromes as model populations for pediatric brain cancer prevention and control research

SELECTED PEER-REVIEWED PUBLICATIONS:

1. Zhang J, et al. and the St Jude Children's Research Hospital Washington University Pediatric Cancer Genome Project. Discovery of novel recurrent mutations and rearrangements in early T-cell precursor acute lymphoblastic leukaemia by whole genome sequencing. *Nature* 2012; 481(7380):157-63. PMID: 23334668
2. **Johnson KJ**, Fisher MJ, Listernick RL, North KN, Schorry EK, Viskochil D, Weinstein M, Rubin J, Gutmann DH. Parent of origin and sex effects in children with NF1 and optic gliomas. *Fam Cancer*. 2012 Dec;11(4):653-6. doi: 10.1007/s10689-012-9549-z. PMID: 22829012
3. Braganza M, Kitahara CM, Berrington de González A, Inskip P, **Johnson KJ**, Rajaraman P. Ionizing Radiation and the Risk of Brain and Central Nervous System Tumors: A Systematic Review. *Neuro Oncol*. 2012 Nov;14(11):1316-24. doi: 10.1093/neuonc/nos208. PMID: 22952197
4. **Johnson KJ**, Hussain I, Williams K, Santens R, Mueller N, Gutmann DH. Development of an International Internet-Based Neurofibromatosis Type 1 Patient Registry. *Contemp Clin Trials*. 2012 Dec 14. PMID: 23246715
5. Kanchi K*, **Johnson KJ***, Lu C*, McLellan MD*, , Ding L. Exome Sequencing Reveals Somatic and Germline Landscape in Ovarian Cancer (*co-first authors, In Press: *Nature Communications*)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Systematic detection and analysis of mutations in 2,000 Cancer Samples

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Michael C. Wendl, Beifang Niu
The Genome Institute, Department of Genetics, Washington University in St Louis, Missouri 63108

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

We have been active in whole genome tumor-normal sequencing, the 1000G project for germline and common variants, and in Pan-Caner exome sequencing data for somatic substitutions, short indels and gene fusions. With the Pan-Cancer genomic, transcriptomic and epigenomic data set of more than 1000 subjects, we will be able to answer the following questions.

1. What are the important cancer-specific and pan-cancer mutation signatures in these data and what contributors do they implicate?
2. What are the cancer type specific and shared germline and somatic variants from whole genomic data?
3. Which genomic variants affect or are more associated with clonal expansion and which are more associated with genetic stability?
4. Which genes are important drivers and how are they best classified, e.g. tumor suppressors versus oncogenes, mutually exclusive versus co-occurring, and prognoses of better or worse outcome?

Timelines & resources dedicated to project

1. Make data and tools available for analysis: download key and small data set to local storage at the genome institute; modify tools for data processing on the cloud for large data set. By the end of Feb 2014
2. Participate genomic variant calling group to obtain a comprehensive call set with balanced sensitivity and specific over a variety of variant types and complete size spectrum. By July 2014
3. Clonal and signature analysis of mutation data. By August 2014
4. Classification analysis for mutation data and their genes. By December 2014

Research proposal

The ICGC data corpus of ~2,000 samples furnishes the best data set and represents the highest potential statistical power to date for researching germline predisposition and somatic mutation dynamics to ultimately deliver a basic understanding of cancer initiation, progression, and treatment options. Most of the established bioinformatic toolsets, including our own, will require some improvement and extension-of-use to take full advantage of these data, for example in more reliably capturing events having low variant allele fraction and better separating germline vs. somatic events. Somatic calls will be analyzed in a variety of ways that have become somewhat standardized with recent Pan-Cancer work. Using our MuSiC system, we will evaluate mutation significance broadly, including assessment of point mutations, small indels, copy number alterations, and structural variants, with the goal of better identifying driver genes and events. For example, we expect that they will fall into recognizable categories like transcription factors, histone modifiers, and specific signalling pathways. We will furthermore examine distribution and clustering of mutation rates across genomes to help identify factors in tumor development. Music will also be used to classify mutation signatures, which often point to specific contributors such as cigarette smoke. In collaboration with Dr. Ben Raphael's lab, we will use Dendrix and allied algorithms to identify mutually exclusive and co-occuring mutations and will also integrate clinical features and data with the genomic analysis (correlation and survival analysis, e.g. Cox proportional hazard) to assess prognostic implications for specific cancer types and across multiple types. Finally, we will integrate analysis clonality analysis to form a more complete picture in terms of clonal timing and gene-action for somatic effects. Regarding improved germline classification, we will also use MuSiC to search for and evaluate correlations very broadly with cancer type, including for susceptibility, and with phenotypes. The large numbers of ICGC samples also mean that broad cross-correlations of somatic and germline events can be examined in tandem with increasingly sufficient statistical power. We anticipate findings that implicate germline and somatic events *in combination*.

Legacy plans

We will modify MuSiC, Bassovac, SciClone, and allied tools for cloud analysis and make the latest source code available for general use in the research community. Other software proposed for this project is open-source and will very likely remain available and functional long after this project concludes.

Li Ding, Ph.D.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L**. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L**, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L**. MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res***. 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L**. Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

Michael C Wendl

A. Education

BS Mechanical Engineering (1989) Washington University in St. Louis
MS Engineering and Applied Science (1990) Washington University in St. Louis
ScD Engineering and Applied Science (1994) Washington University in St. Louis

B. Positions

2005-present	Research Assistant Professor, The Genome Institute, Washington University School of Medicine, St Louis, MO
2010-present	Research Assistant Professor (courtesy), Department of Mathematics, School of Arts and Sciences, Washington University, St Louis MO
2001-2005	Research Instructor, The Genome Center, Washington University School of Medicine, St Louis, MO
1994-2001	Research Associate, The Genome Center, Washington University School of Medicine, St Louis, MO
1999-present	Research Assistant Professor (courtesy), Department of Mechanical Engineering and Materials Science, School of Engineering, Washington University, St Louis MO

C. Selected Peer-Reviewed Publications (chronological order, selected from 58 journal publications)

1. Ewing B, Hillier L, **Wendl MC**, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Research 1998 Mar:8(3):175-85.
2. **Wendl MC**, Waterston RH. Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. Genome Research 2002 Dec:12(12):1943-9. PMCID: PMC187573.
3. **Wendl MC**. Collision probability between sets of random variables. Statistics and Probability Letters 2003: 64(3): 249-54.
4. **Wendl MC**. A general coverage theory for shotgun DNA sequencing. Journal of Computational Biology 2006 Jul-Aug:13(6):1177-96.
5. **Wendl MC.** Random covering of multiple one-dimensional finite domains with an application to DNA sequencing. SIAM Journal on Applied Mathematics 2008: 68(3): 890-905.
6. **Wendl MC**, Wilson RK. Statistical aspects of discerning indel-type structural variation via DNA sequence alignment. BMC Genomics 2009 Aug 5:10:359. PMCID: PMC2748092.
7. Ding L, **Wendl MC**, Koboldt DC, and Mardis ER. Analysis of next-generation genomic data in cancer: Accomplishments and challenges. Human Molecular Genetics 2010: 19(2), 188-196.
8. **Wendl MC,** Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, and Ding L. PathScan: A tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics 2011: 27(12): 1595-1602.
9. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, **Wendl MC**, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, DiPersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012: 481(7382), 506-510.
10. **Wendl MC**, Kota K, Weinstock GM, Mitreva M. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. Journal of Mathematical Biology 2013: 67(5), 1141-1161 DOI 10.1007/s00285-012-0586-x.
11. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, **Wendl MC**, Ley TJ, Wilson RK, Raphael BJ, Ding L. Mutational landscape and significance across 12 major cancer types. Nature 2013: 502(7471), 333-339.
12. Ding L, **Wendl MC**. Differences that matter in cancer genomics. Nature Biotechnology 2013: 31(10), 892-893.
13. Ding L, Raphael BJ, Chen F, and **Wendl MC**. Advances for studying clonal evolution in cancer. Cancer Letters 2013: in press.

BEIFANG NIU

Staff Scientist
The Genome Institute & School of Medicine
Washington University, St. Louis
Box 1910
St. Louis, MO, 63108

Email: bniu@wugsc.genome.edu & beifang.cn@gmail.com

RESEARCH INTERESTS

Cancer genomics and Metagenomics. Bioinformatics software development and algorithm optimization. Large scale bioinformatics data analysis and sequence clustering, alignment and assembly algorithm for next generation sequencing data. High performance parallel computing and cloud computing.

EDUCATION

2002 – 2009 **Ph.D.** in Computer Software and Theory, Super Computing Center, Chinese Academy of Sciences, Beijing, China.
Thesis Title: *Research on Short Oligonucleotide Alignment and Assembly Algorithm.* Advisor: Professor Xuebin Chi.

1998 – 2002 **S.B.** in Computer Science, Shandong Agriculture University, Shandong, China

EXPERIENCE

2012 – present **Staff Scientist**, (Cancer Genomics), The Genome Institute, School of Medicine, Washington University in St. Louis, Missouri, US.
Sponsor: Professor Li Ding.

2009 – 2012 **Postdoctoral Associate**, (Bioinformatics), The Center for Research in Biological Systems, University of California, San Diego, California, US.
Sponsor: Doctor Weizhong Li.

2007 – 2008 **Research Intern**, (Bioinformatics), Beijing Genomics Institute (BGI), Beijing, China.
Sponsor: Professor Jun Wang

2003 – 2007 **Research Assistant**, (Parallel Computing), Super Computing Center, Chinese Academy of Sciences, Beijing, China.
Sponsor: Professor Xuebin Chi

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Discovery of significant non-coding mutations in whole cancer genomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Matthew A. Wyczalkowski
Michael D. McLellan
The Genome Institute, Department of Genetics, Washington University in St Louis, Missouri 63108

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Large-scale cancer-genomics sequencing efforts to date have focused with few exceptions on the analysis of coding variants. A primary reason for this emphasis is the availability of large sets of data specific to coding regions, primarily next-generation exome capture and RNA-Seq. Such data have enabled the identification of large numbers of significantly mutated oncogenes, tumor suppressors, and their corresponding networks and pathways.

Highly sensitive approaches have identified causative somatic and germline mutations in genes that are significantly mutated, fused, amplified, or deleted, but a significant population of samples lacks non-synonymous mutations in such genes. Expression level analysis and clustering of these samples nevertheless suggests they harbor mutations in such genes and pathways and hints at the presence of unidentified non-coding mutations and genomic alterations which may account for tumorigenesis.

The availability of the ENCODE database of regulatory elements in conjunction with large-scale whole-genome sequence data for two thousand samples presents an opportunity to discover the unknown mechanisms modulating cancer related pathways and gene networks. Utilizing 2000 ICGC samples, we aim to identify non-coding and apparently “silent” mutations that lead to aberrant splicing and gene regulation due to changes to miRNA binding affinity, the disruption of RNA secondary/three dimensional structure, alteration of promoter function, and the dysregulation of gene expression in cancer related genes and pathways.

Timelines & resources dedicated to project

1) Feb 2014: Import, filter, normalize, and curate data
2) June 2014: Analyze data and develop automated software tools
3) Sept 2014: Refine analysis for publications
4) Oct 2014: Release software tools to public repository
5) 2015: Publish list of functional cancer-related non-coding variants

Research proposal

The remarkable breadth of whole genome data available for thousands of patients across a range of cancers presents research opportunities on several fronts. Some non-coding and synonymous mutations play a key role in regulating protein expression, but sequencing technologies which are biased toward coding regions make the interpretation of their roles difficult. Also, background mutation rates (BMR) display significant heterogeneity along the length of the genome, as well as across cancer types. Obtaining accurate estimates of such mutation rates is critical for identifying significantly mutated genes or genomic regions, which drive cancer progression (Lawrence et al. 2014). In both cases, progress requires large numbers of samples, each having high coverage whole genome data, for each cancer type and subtype.

We aim to evaluate the effects of known copy number variation and coding mutations on individual genes and pathways and then to identify samples having similar patterns of dysregulation for noncoding and synonymous mutations that are candidate causal variants. We will analyze variants in 5' promoter and 3' UTR/miRNA regulatory regions, as well as other ENCODE-annotated regulatory elements using various sliding window statistical analyses. These tools will identify candidate regions enriched with potential causative mutations. Refined per-cancer heterogenous background mutation rate estimates will be used to extend the MuSiC SMR (significantly mutated region) tool (Dees 2012) and to evaluate each potential significantly mutated region. Finally, we will investigate novel RNA isoforms corresponding to splice altering mutations in exons, introns, and the spliceosome.

Legacy plans

All software developed as part of this project will be made publicly available both during development and subsequent to project completion on our laboratory's GitHub repository.

An integrated, precomputed, genome-wide scoring matrix will be released to weight the functional effects of all possible non-coding and synonymous mutations that are predicted to modulate gene expressions and pathway regulation.

Li Ding, Ph.D.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L**. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L**, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L**. MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res.*** 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L**. Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

Matthew A. Wyczalkowski, Ph.D.

Postdoctoral Research Associate
The Genome Institute
Washington University School of Medicine

4106 Wyoming St.
St. Louis, MO 63116
wyczalkowskim@wustl.edu

EDUCATION

Washington University in St. Louis, MO **2009**
Ph.D. in Biomedical Engineering.

University of California, Berkeley, CA **2000**
M.S. in Mechanical Engineering.

Pennsylvania State University, State College, PA **1996**
B.S. in Engineering Science, Engineering Mechanics minor.

RESEARCH

Postdoctoral Research Associate, Genome Institute, Washington University **2013 - present**
Advisor: Li Ding, Ph.D.
Pathogen discovery in whole genome sequences and associated tool development.

NIH NRSA Postdoctoral Fellow, Biomedical Engineering, Washington University **2010 - 2013**
Advisor: Larry A. Taber, Ph.D.
Experimental and computational investigation wound healing in chick embryonic epithelia with application to morphogenesis.

Graduate Research Associate, Department of Biomedical Engineering and Center for Computational Biology, Washington University **2004 - 2009**
Advisor: Rohit V. Pappu, Ph.D.
Methods development for free energy and entropy of solvation calculations from molecular dynamics and Monte Carlo simulations

Graduate Research Assistant, Mechanical Engineering, Univ. California, Berkeley **1997 - 2000**
Advisor: Andrew J. Szeri, Ph.D.
*N*onlinear control schemes for acoustically driven microbubbles as contrast agents in medical ultrasound,

SELECTED PUBLICATIONS

Kandoth, C, MD McLellan, F Vandin, K Ye, B Niu, C Lu, M Xie, Q Zhang, JF McMichael, **MA Wyczalkowski**, MDM Leiserson, CA Miller, JS Welch, MJ Walter, MC Wendl, TJ Ley, RK Wilson, BJ Raphael, L Ding. Mutational landscape and significance across 12 major cancer types. Nature 502, 333-339 (2013).

Wyczalkowski, MA, VD Varner and LA Taber. Computational and Experimental Study of the Mechanics of Embryonic Wound Healing. J Mech Behav Biomed Mater 28, 125-146 (2013).

Wyczalkowski*, MA, Z Chen*, BA Filas, VD Varner and LA Taber. Computational Models for Mechanics of Morphogenesis. Birth Defects Research Part C: Embryo Today: Reviews. 96(2), 132–152 (2012)
** Denotes equal contributions*

Wyczalkowski, MA, A Vitalis and RV Pappu. New Estimators for Calculating Solvation Entropy and Enthalpy and Comparative Assessments of Their Accuracy and Precision. J. Phys. Chem. B 114, 8166-8180 (2010).

Michael D. McLellan II
Business & Technology Applications Analyst III
The Genome Institute - Campus Box 8501
Washington University School of Medicine
4444 Forest Park Blvd, St. Louis, Missouri 63108
E-mail: mmclella@wustl.edu
Phone: 314.856.5557

Research Interests

Cancer genomics and human genetics. Bioinformatics pipeline development and data integration, with a focus on high-throughput genomic and proteomic sequence data from large cohorts. Analysis of gene, pathway, and network interaction and dysregulation.

Education

1994-1998	Bachelor of Science, Biological Sciences, University of Missouri – Columbia
1994-1998	Bachelor of Arts, German Literature, University of Missouri – Columbia
1996-1997	University of Stuttgart, Stuttgart, Germany

Experience

2013 – Present	Business & Technology Analyst III, Cancer Genomics
2008 – 2012	Senior Programmer Analyst, Medical Genomics
2005-2008	Research Lab Supervisor, Medical Genomics
2004-2005	Assistant Lab Coordinator, Mutational Profiling
1999- 2004	Senior Research Technician, Genome Finishing Group

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Building a comprehensive catalogue of somatic substitutions, indels and structure variants, as well as the characteristics of transcriptome and epigenome in ICGC samples

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Kai Ye, R. Jay Mashl
The Genome Institute, Department of Genetics, Washington University in St Louis, Missouri 63108

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

We have been working with 1000 Genomes data to identify germline and common variants (1, 2) and TCGA sequencing data to discover significant mutations and genes in ovarian, breast, endometrial, lung, and other cancer types. Recently, we analyzed Pan-Cancer exome sequencing data for somatic substitutions, short indels, and gene fusions from 12 major cancer types (3).

To support these efforts, we have also been extending our processing pipeline for variant calling to operate in cloud environments. The prototype runs on Amazon's EC2 Elastic Cloud, where we have successfully tested our analysis tool suite and its interaction with the 1000 Genomes data set located in Amazon's S3 storage resource.

[1] Durbin et al., A map of human genome variation from population-scale sequencing, Nature, 467, 1061-1073
[2] Mills et al., Mapping copy number variation by population-scale genome sequencing, Nature, 470, 59-65

[3] Kandoth et al., Mutational landscape and significance across 12 major cancer types, Nature, 2013, 502, 333-339.

Timelines & resources dedicated to project

1. Collect variant calls from major pipelines run in ICGC and compare/merge them. We will evaluate calls from different callers and stratify sensitivity and specificity by variant size, variant type, and sequence composition for a performance matrix. We will then decide which calls to pick up for the final merged set. July 2014
2. Process ICGC data with additional tools using cloud computing. We will modify tools such as BreakTrans, BreakfFusion and PASSion for cloud execution for RNA-seq data analysis. July 2014
3. Release the curated somatic variant list. Oct 2014

Research proposal

The International Cancer Genome Consortium (ICGC) has generated whole genome sequence data of about 2,000 samples from 50 different cancer types of clinical and societal importance across the globe. This whole genomic sequence data allows, for the first time, the comprehensive discovery of somatic genomic alterations of any variant types and of full size spectrum among major cancer types. In the genomic variant discovery phase, we will first survey the landscape of somatic variants in all cancer types. We will utilize VarScan2, BreakDancer, Pindel and CNVnator to detect somatic variants, such as substitutions, short indels, CNVs, and complex structural variants, including interchromosomal translocations. We will continue to modify various analysis tools for use in cloud environments. Our cloud pipeline is sufficiently general to be ported to other cloud systems if necessary. We will process the ICGC data using cloud computing, if analysis using alternative tools is required. We will investigate caller performance for different variant size, variant types, and sequence complexity. We will then design a merging strategy to select the right tools for a given size, type and sequence property to maximize sensitivity and specificity.

Legacy plans

We will modify VarScan2, BreakDancer, Pindel, BreakTrans, BreakFusion, PASSion and Bassovac for cloud analysis and make source code available for general research community.

Li Ding, Ph.D.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L**. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L**, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L**. MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res***. 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L**. Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

The Genome Institute at Washington University in St. Louis

Kai Ye, PhD

Positions and degrees

- *2012-now*: Research assistant professor, the genome institute at Washington University in St. Louis
- *2009-2012*: Assistant professor, Leiden University Medical Center, the Netherlands
- *2008-2009*: Postdoc European Bioinformatics Institute, UK
- *2004-2008*:**PhD.** *Cum Laude* Biopharmaceutical science at Leiden University, The Netherlands
  Novel algorithms for protein sequence analysis
- *Aug. 2003-Dec. 2003*: Lecturer in the college of Pharmacy at Wuhan University, China.
- *Sept. 1995-Jul. 2003*: **B.Sc.** and **M.Sc.** of Biopharmaceutical science at Wuhan University, China.

Awards and grants

- June 2009: 'best paper' (on Pindel) presented at the Short-SIG on Next-Generation Sequence and Algorithms for Short Read Analysis, ISMB/ECCB 2009 at Stockholm, Sweden. 500 GBP
- 'Researcher of 2008', Faculty of Science, Leiden University.
- 2008 C. J. Kok prize, Leiden University, the Netherlands. 2,500 EURO
- 2008 **PhD Cum Laude**, Leiden, the Netherlands
- NGI/EBI fellowship, The Netherlands. 36,000 euro per year.
- 'Top 300 Outstanding PhD students abroad' Award, China. 5,000 USD
- *2004-2005* Leiden University Scholarship, The Netherlands.
- *2000-2003* Wuhan University fellowship for excellent graduate student, China.
- *1996-1999* Wuhan University fellowship for excellent undergraduate, China.

Programming experience (and related)

- Good programming experience in C/C++ and MatLab;
- Familiar with Perl, Python, PHP, MySQL and linux
- Good experience in using InsightII, AutoDock, PyMol, SPDBV and WHAT IF

Research interests

- Large scale data mining and sequence analysis: protein, DNA
- Pharmacology modeling: mathematical modeling of receptor-receptor, receptor-ligand interaction
- Homology modeling, MD simulation and docking
- Cancer genomics

kye@genome.wustl.edu

ROBERT JAY MASHL

EDUCATION

University of California, Los Angeles	Chemistry	Ph.D., 1998
The University of Chicago	Chemistry	M.S., 1995
University of Wisconsin, Madison	Chemistry; Mathematics	B.S., with Distinction, 1992

RESEARCH EXPERIENCE

The Genome Institute, Washington University — 2013–present
Staff Scientist, Ding Lab

National Center for Supercomputing Applications, University of Illinois — 2006–2013
Research Scientist

Department of Molecular and Integrative Physiology and The Beckman Institute, University of Illinois — 1998–2006
Postdoctoral Research Associate

PUBLICATIONS (SELECTED)

Mashl, R. J., B. Acs, J. R. Schmidt, W. F. Polik, and E. N. Wiziecki. (2013). Enhancing chemistry teaching and learning through computational tools: A computational chemistry cloud prototype using WebMO. Proceedings of the *7th International Multi-conference on Society, Cybernetics, and Informatics: IMSCI 2013*.

Natarajan, S., R. J. Mashl, and E. Jakobsson. (2010). Evolutionary coupling in the Kv1.2-beta2 complex. Channels. 4(5). DOI: 10.4161/ chan.4.5.12813.

Toghraee, R., R. J. Mashl, K.I. Lee, E. Jakobsson, and U. Ravaioli. (2009). Simulation of charge transport in ion channels and nanopores with anisotropic permittivity. *J. Comput. Electron*. 8: 98–109.

Chen, D., D. Kearney, J. Mashl, N. Sobh, E. Jakobsson. (2008). "NanoGromacs" DOI: 10254/ nanohub-r4123.4. (https:// nanohub.org/ resources/ nanogromacsii).

Mashl, R. J. and E. Jakobsson. (2008). End-point targeted molecular dynamics: Large-scale conformational changes in potassium channels. *Biophys. J*. 94: 4307–4319.

Mashl, R. J., S. Joseph, N. R. Aluru, and E. Jakobsson. (2003). Anomalously immobilized water: a new water phase induced by confinement in nanotubes. *Nano Lett*. 3: 589–592.

Mashl, R. J., Y. Tang, J. Schnitzer, and E. Jakobsson. (2001). Hierarchical approach to predicting permeation in ion channels. *Biophys. J.* 81: 2473–2483.

Mashl, R. J., N. Gronbech-Jensen, M. Fitzsimmons, M. Lutt, and D. Li. (1999). Theoretical and experimental adsorption studies of polyelectrolytes on an oppositely charged surface. *J. Chem. Phys.* 110: 2219–2225.

Mashl, R. J. and R. F. Bruinsma. (1998). Spontaneous-curvature theory of clathrin-coated membranes. *Biophys. J.* 74: 2862–2875.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

The impact of somatic structure variants on transcriptome and epigenome

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Li Ding
The Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St Louis, Missouri 63108, USA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Kai Ye, Matthew A. Wyczalkowski
The Genome Institute, Department of Genetics, Washington University in St Louis, Missouri 63108

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

We have participated in the 1000 Genomes Project for germline and common structural variants (1, 2), as well as Pan-Cancer exome sequencing data for somatic gene fusions (3). With whole genome, transcriptomic, and epigenomic data sets for more than 2000 subjects across a variety of cancer types, we will investigate the functional impact of large structural variants on the transcription of nearby genes and on the local methylation patterns.

[1] Durbin et al., A map of human genome variation from population-scale sequencing, Nature, 467, 1061-1073
[2] Mills et al., Mapping copy number variation by population-scale genome sequencing, Nature, 470, 59-65

[3] Kandoth et al., Mutational landscape and significance across 12 major cancer types, Nature, 2013, 502, 333-339

Timelines & resources dedicated to project

4. Download and merge somatic structural variants calls from ICGC. March 2014
5. Download and merge transcriptome as well as epigenome result. March 2014
6. If necessary, process ICGC data on the cloud with additional cloud compatible tools for additional structural variants. June 2014
7. If necessary, process ICGC transcriptome data on the cloud for splicing variants and transcription levels. June 2014
8. If necessary, process ICGC methylation data on the cloud for methylation pattern per gene and per genomic region. June 2014
9. Integrated analysis of genome, transcriptome and methylation for association mining. September 2014
10. Publish findings in 2015

Research proposal

The International Cancer Genome Consortium (ICGC) has generated whole genome sequence data of about 2,000 samples from 50 different cancer types of clinical and societal importance across the globe. This whole genomic sequence data allows, for the first time, the comprehensive discovery of somatic genomic alterations of any variant type and across the full size spectrum among major cancer types. For a subset of the samples with transcriptomic and/or methylation data available, we will investigate the functional impact of many common somatic structural genomic variants on the expression level and epigenomic status of nearby genes. We will group samples into two subsets based on the presence of a given common somatic variant. Then we will compare the transcription and methylation profiles between the two groups. We will rank the common somatic structural variants based on the functional impact and further investigate the regulatory mechanism. We will use BreakTrans and Passion to investigate gene fusions and splice junctions in RNA-Seq data and will identify any links between the genomic variants and significant features in the transcriptomic and epigenome data. Finally we will search for genomic, transcriptomic, and epigenomic signatures which are common across -- as well as specific to -- particular cancer types. Taken together, our integrated analysis of ICGC data will shed light on new diagnostic biomarkers, personalized cancer treatment, and ultimately prevention.

Legacy plans

We will modify existing transcriptome and epigenome tools for cloud analysis and make source code available for general research community.

Li Ding, Ph.D.

A. Positions

Assistant Director, The Genome Institute, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

B. Selected Peer-Reviewed Publications

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. ***Nature*** 2008 Oct 23:455(7216):1069-75. PMCID: PMC2694412.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L**. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. ***Bioinformatics*** 2009 Sep 1:25(17):2283-5. PMCID: PMC2734323.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, **Ding L**, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods*** 2009 Sep:6(9):677-81.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. ***Nature*** 2010 Apr 15:464(7291):999-1005. PMCID: PMC2872544.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, Dipersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. ***Nature***. 481, 506–510 (26 January 2012) PubMed PMID: 22237025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L**. MuSiC: Identifying mutational significance in cancer genomes. ***Genome Res***. 2012 Aug;22(8):1589-98

International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, **Ding L**, Sander C, Stuart JM, Stein LD, Lopez-Bigas N. Computational approaches to identify functional genetic variants in cancer genomes. ***Nat Methods.*** 2013 Jul 30;10(8):723-9. doi: 10.1038/nmeth.2562.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, Mclellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK and **Ding L**. BreakTrans: uncovering the genomic architecture of gene fusions. ***Genome Biol.*** 2013 Aug 23;14(8):R87.

Kandoth C., McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson M, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, and **Ding L.** Mutational Landscape and Significance across 12 Major Cancer Types. ***Nature.*** 502, 333-339

The Genome Institute at Washington University in St. Louis

Kai Ye, PhD

Positions and degrees

- *2012-now*: Research assistant professor, the genome institute at Washington University in St. Louis
- *2009-2012*: Assistant professor, Leiden University Medical Center, the Netherlands
- *2008-2009*: Postdoc European Bioinformatics Institute, UK
- *2004-2008*:**PhD.** *Cum Laude* Biopharmaceutical science at Leiden University, The Netherlands
  Novel algorithms for protein sequence analysis
- *Aug. 2003-Dec. 2003*: Lecturer in the college of Pharmacy at Wuhan University, China.
- *Sept. 1995-Jul. 2003*: **B.Sc.** and **M.Sc.** of Biopharmaceutical science at Wuhan University, China.

Awards and grants

- June 2009: 'best paper' (on Pindel) presented at the Short-SIG on Next-Generation Sequence and Algorithms for Short Read Analysis, ISMB/ECCB 2009 at Stockholm, Sweden. 500 GBP
- 'Researcher of 2008', Faculty of Science, Leiden University.
- 2008 C. J. Kok prize, Leiden University, the Netherlands. 2,500 EURO
- 2008 **PhD Cum Laude**, Leiden, the Netherlands
- NGI/EBI fellowship, The Netherlands. 36,000 euro per year.
- 'Top 300 Outstanding PhD students abroad' Award, China. 5,000 USD
- *2004-2005* Leiden University Scholarship, The Netherlands.
- *2000-2003* Wuhan University fellowship for excellent graduate student, China.
- *1996-1999* Wuhan University fellowship for excellent undergraduate, China.

Programming experience (and related)

- Good programming experience in C/C++ and MatLab;
- Familiar with Perl, Python, PHP, MySQL and linux
- Good experience in using InsightII, AutoDock, PyMol, SPDBV and WHAT IF

Research interests

- Large scale data mining and sequence analysis: protein, DNA
- Pharmacology modeling: mathematical modeling of receptor-receptor, receptor-ligand interaction
- Homology modeling, MD simulation and docking
- Cancer genomics

kye@genome.wustl.edu

Matthew A. Wyczalkowski, Ph.D.

Postdoctoral Research Associate
The Genome Institute
Washington University School of Medicine

4106 Wyoming St.
St. Louis, MO 63116
wyczalkowskim@wustl.edu

EDUCATION

Washington University in St. Louis, MO **2009**
Ph.D. in Biomedical Engineering.

University of California, Berkeley, CA **2000**
M.S. in Mechanical Engineering.

Pennsylvania State University, State College, PA **1996**
B.S. in Engineering Science, Engineering Mechanics minor.

RESEARCH

Postdoctoral Research Associate, Genome Institute, Washington University **2013 - present**
Advisor: Li Ding, Ph.D.
Pathogen discovery in whole genome sequences and associated tool development.

NIH NRSA Postdoctoral Fellow, Biomedical Engineering, Washington University **2010 - 2013**
Advisor: Larry A. Taber, Ph.D.
Experimental and computational investigation wound healing in chick embryonic epithelia with application to morphogenesis.

Graduate Research Associate, Department of Biomedical Engineering and Center for Computational Biology, Washington University **2004 - 2009**
Advisor: Rohit V. Pappu, Ph.D.
Methods development for free energy and entropy of solvation calculations from molecular dynamics and Monte Carlo simulations

Graduate Research Assistant, Mechanical Engineering, Univ. California, Berkeley **1997 - 2000**
Advisor: Andrew J. Szeri, Ph.D.
*N*onlinear control schemes for acoustically driven microbubbles as contrast agents in medical ultrasound,

SELECTED PUBLICATIONS

Kandoth, C, MD McLellan, F Vandin, K Ye, B Niu, C Lu, M Xie, Q Zhang, JF McMichael, **MA Wyczalkowski**, MDM Leiserson, CA Miller, JS Welch, MJ Walter, MC Wendl, TJ Ley, RK Wilson, BJ Raphael, L Ding. Mutational landscape and significance across 12 major cancer types. Nature 502, 333-339 (2013).

Wyczalkowski, MA, VD Varner and LA Taber. Computational and Experimental Study of the Mechanics of Embryonic Wound Healing. J Mech Behav Biomed Mater 28, 125-146 (2013).

Wyczalkowski*, MA, Z Chen*, BA Filas, VD Varner and LA Taber. Computational Models for Mechanics of Morphogenesis. Birth Defects Research Part C: Embryo Today: Reviews. 96(2), 132–152 (2012)
** Denotes equal contributions*

Wyczalkowski, MA, A Vitalis and RV Pappu. New Estimators for Calculating Solvation Entropy and Enthalpy and Comparative Assessments of Their Accuracy and Precision. J. Phys. Chem. B 114, 8166-8180 (2010).

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identifying clinically relevant oncogenic gene clusters on Chr1q

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

P. Andrew Futreal, MD Anderson Cancer Center, ICGC
Lynda Chin, MD Anderson Cancer Center, ICGC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Hannah Cheung, MD Anderson Cancer Center **Sahil Seth, MD Anderson Cancer Center**	**Jianhua Zhang, MD Anderson Cancer Center**

Background and preliminary data

Gains involving the long arm of chromosome 1 are the highest recurrent somatic event across multiple primary tumor types.[1] Among them, gains of the entire 1q arm are the most common, while other types of alterations include: isochromosome formation i(1q), translocations (t(1;16), t(1;19), t(1;3)), polysomy, monosomy, and focal gains and deletions.[2-13] Multiple groups have narrowed down consensus regions of gain to: 1q21, 1q32.1, 1q23-q25 and 1q44 that include notable candidate genes *MCL1*, *MDM4*, *KIF14*, *NORE1*, *KISS1*, *ARF1*, and *MUC1*, many of which show comparable increases in expression and two (*KIF14*, *NORE1*) that enhance tumorigenesis when overexpressed *in vitro*.[14-30] Some of these consensus regions occur more frequently in melanoma[19,32] and renal cell carcinoima patients[33] with metastasis and correlate with poor prognosis for ependymoma patients,[21] decreased survival for those with retinoblastoma,[33] and higher stage of cervical carcinomas.[27] However, using multiple tumor sets on Oncomine, none of these candidate genes alone associates strongly with clinical status (metastasis, overall survival, recurrence, and survival after five years). As well, patients with 1q21.1 microduplication syndrome do not appear to have early onset cancer.[34] Therefore, it is more likely that the oncogenic selective pressure of 1q gain involves multiple consensus regions or acts concurrently with other genomic aberrations, which would explain the recurrent translocations and associations with 6p gain in melanoma,[35-36] 8q gain in lung adenocarcinomas,[23] and 16q loss in breast cancer.[17] Historically, allelic imbalances resulting from these gains were measured using microsatellite markers that underwent losses of heterozygosity. With the advent of whole genome sequencing, these allelic fractions can now be characterized at the basepair level and breakpoints from translocations and focal copy number variations can be mapped. The ability to calculate allelic imbalance using both copy number and haplotyping tools will enable a more comprehensive analysis of whether clinical observations resulting from somatic gains in 1q are allele-specific, ploidy-specific or dependent on other concurrent genetic aberrations that may be tumor-specific.

Timelines & resources dedicated to project

Aim 1: Characterize the levels of increased gene expression in tumor samples with 1q gain
Time expected: 2-3 months
This aim relies on matched copy number WGS and RNASeq data results.
Aim 2: Determine whether tumors without 1q gain have gain-of-function point mutations in 1q genes
Time expected: 1 month.
Aim 3: Conduct multiple comparisons of gain or activating point mutations on 1q and clinical outcome
Time expected: 8-9 months
Resources: We shall use the computing infrastructure provided by the consortium, and if needed could use our nodes co-localized with CGHub or MD Anderson HPCC.

International Cancer Genome Consortium

Research proposal

There are multiple mechanisms for increasing oncogenic activity, including: copy number gain, focal amplifications, enhanced promoter activation, decreased decay of the transcript or protein, activating point mutations, and rearrangements that place genes under more active promoters. We will conduct a systematic study to identify pan-cancer, oncogenic drivers from Chr1q, focusing first on copy number gain, focal amplifications, activating point mutations, and activating rearrangements. Once other ICGC working groups complete their analyses on noncoding regions, we will incorporate their findings into our study.

Aim 1: Characterize the levels of increased gene expression in tumor samples with 1q gain
Using matched whole genome sequencing and RNAseq data from each normal-tumor paired sample that has gain in 1q, the length and degree of amplification will be assessed for subsequent increases in expression. We anticipate that the extent to which expression is elevated for each gene may also be influenced by coding and noncoding sequence variants that arose before or after duplication[37] as well as tumor type. Allele-specific copy number state will be derived with tools such as HATS[38] and adapted hapLOH.[39] We will also map and annotate any structural variations with tools such as Breakpointer[40] and CREST[41] to find rearrangement-mediated increases in expression. Those genes with tissue-specific expression increases will be filtered out in order to focus on those that with comparable changes across multiple tissue types.

Aim 2: Determine whether tumors without 1q gain have gain-of-function point mutations in 1q genes
Activating point mutations are infrequent in oncogenes that have undergone gains or focal amplifications. Therefore, tumors without 1q gain may utilize alternate mechanisms, such as activating point mutations, to increase levels of 1q oncogenes. Tools such as Paradigm Shift[42] and MutSig[43] will be employed to identify gain-of-function mutations for such cases.

Aim 3: Conduct multiple comparisons of 1q gain with other tumor-related features
We will separate samples according to classes of recurrent 1q gain identified in Aims 1 and 2 for supervised clustering analyses. There, we will identify associations between 1q gain and other genetic alterations (concordant and exclusive) as well as clinical outcome. Then, we will also conduct multiple principal component and machine-learning tests of single and combinations of activated 1q genes to identify genes or regions that have the most impact on clinical outcome. If proven valuable in enhancing the outputs of these tests, we will incorporate predictive scores from Paradigm Shift, TUSON[44], MutSig, and MuSIC[45] to increase signals from the genes and mutations. This will be done in a pan-cancer analysis, within individual cancer types and within individual samples that have longitudinal sequencing data. We will also note any concordant or non-correlating chromosomal events that, in combination with increased expression of 1q oncogenes, associate with clinical outcome.

Note:
Learning from this analysis/approach can be applied to other regions of interest, as suggested by the group.

Legacy plans

The analysis protocols may be automated to study other common chromosomal aberrations. We will make publicly available any tools that we develop.

International Cancer Genome Consortium

Sample References

1. Beroukim, R. et al., Nature (2010) 463:899.
2. Tsarouha et al., Cancer Genet. Cytogenet. (1999) 113:156.
3. Dutrillaux et al., Cancer Genet. Cytogenet. (1990) 49:203.
4. Pandis et al., Genes Chr. Cancer (1992) 5:235.
5. Tirkkonen, M. et al., Genes Chr. Cancer (1998) 21:177.
6. Parmitier, A.H. et al., Can. Res. (1986) 46:1526.
7. Kanayama, H. et al., J. Med. Genet. (2001) 38:165.
8. Chen, J. et al., Can. Cell (2003) 4:405.
9. Girard, L. et al., Can. Res. (2000) 60:4894.
10. Chitale, D. et al., Oncogene (2009) 28:2773.
11. Atkin, NB et al., Cancer (1979) 44:604.
12. Orsetti, B. et al., BJC (2006) 95:1439.
13. Hing, S. et al. Am. J. Path. (2001) 158:393.
14. Lu, et al., Genes. Chr. Cancer (1998) 20:275.
15. Fabris et al., Leukemia (2007) 21:1113.
16. Chen, L.C. et al. PNAS (1989) 86:7204.
17. Stange, D.E. et al. Clin. Can. Res. (2006) 12:345.
18. Zack, T.I. et al., Nat. Gen. (2013) 45:1134.
19. Bastian, B.C. et al., Can.Res. (1998) 58:2170.
20. Prat, E. et al., Urology (2001) 57:986.
21. Mendryzk, F. et al., Clin. Can. Res. (2006) 12:2070.
22. Baudis, M. and M.L. Cleary Bioinformatics (2001) 17:1228.
23. Chitale, D. et al., Oncogene (2009) 28:2773.
24. Chen, Y.-J. et al., Can. Res. (2003) 63:817.
25. Kitamura, Y. et al., Genes Chr. Cancer (2000) 27:244.
26. Kitamura, Y. et al., Clin. Can. Res. (2000) 6:1819.
27. Cheung, T.H. et al., Cancer (1999) 86:1294.
28. Corson, T.W. et al., Oncogene (2005) 24:4741.
29. Simon, R. et al., J. Path. (1998) 185:345.
30. Wreesman, V.B. et al. Can. Res. (2004) 64:3780.
31. Borg, A., et al. Genes Chr. Cancer (1992) 5:311.
32. Aalto et al., Invest. Ophthlmol. (2001) 42:313.
33. Gronwald, J. et al. Can. Res. (1997) 57:481.
34. http://omim.org/entry/612475
35. Namiki, T. et al., Can. Gen. Cytogen. (2005) 157:1.
36. Van Dijk, M. et al., Genes Chr. Cancer (2002) 36:151.
37. Nik-Zainal, S. et al., Cell (2012) 149:994.
38. Dewal, N. et al., Genome Res. (2012) 22:362.
39. Vattathil, S. and P. Scheet. Genome Res. (2013) 23:152.
40. Drier, Y. et al., Genome Res. (2013) 23:228.
41. Wang, J. et al., Nat. Met. (2011) 8:652.
42. Ng, S. et al., Bioinformatics (2012) 28:i640.
43. Lawrence, M. et al. Nature (2013) 499:214.
44. Davoli, T. et al. Cell (2013) 155:948.
45. Dees, N.D. et al., Gen. Res. (2012) 22:1589.

International Cancer Genome Consortium

CURRICULUM VITAE

Andrew Futreal, PhD
Phone: (713) 794-4764 Email: afutreal@mdanderson.org

PRESENT TITLE AND AFFILIATION

Primary Appointment
Professor, Department of Genomic Medicine, Division of Cancer Medicine
The University Of Texas MD Anderson Cancer Center

Honorary Faculty Member
Director, Cancer Genetics and Genomics
Co-Director, Cancer Genome Project
Wellcome Trust Sanger Institute

EDUCATION

Degree-Granting Education
B.S. in Biology, UNC-Charlotte, 1987
Ph.D. in Pathology, UNC-Chapel Hill, 1993

Postgraduate Training
Postdoctoral Fellowship, 1993-1995
Post-Doctoral IRTA Fellow Award
National Institute of Environmental Health Sciences,
National Institutes of Health

EXPERIENCE/SERVICE

Academic Appointments
Honorary Faculty Member
Director, Cancer Genetics and Genomics
Co-Director, Cancer Genome Project
Wellcome Trust Sanger Institute

ONGOING RESEARCH SUPPORT

N/A STARS Award Futreal (PI)
06/06/2012 – 06/05/2015
R1205 Futreal (PI) Cancer Prevention & Research Institute of Texas (CPRIT)
11/02/2011 – 11/01/2016

PUBLICATIONS (last 5)

Peer-Reviewed Original Research Articles
Stephens PJ, et al. Nature. 2012 May 16;486(7403):400-4.
Jonasch E, et al. Mol Cancer Res. 2012 Jul;10(7):859-80.
Nik-Zainal S, et al. Cell. 2012 May 25;149(5):979-93.
Nik-Zainal S, et al. Cell. 2012 May 25;149(5):994-1007
Ong CK, et al. Nat Genet. 2012 May 6;44(6):690-3.

International Cancer Genome Consortium

CURRICULUM VITAE

Lynda Chin, MD
Email: LChin@mdanderson.org

PRESENT TITLE AND AFFILIATION

Professor and Chair, Dept of Genomic Medicine
Scientific Director, Institute for Applied Cancer Science

EDUCATION/TRAINING

Degree-Institution and Location

09/84-06/88 BS in Neuroscience, Brown University, Providence, RI
09/89-06/93 MD in Medicine, Albert Einstein College of Medicine, Bronx, NY
07/93-06/94 Internship, Columbia Presbyterian Medical Center, NY, NY
07/94-06/97 Residency, Albert Einstein College of Medicine, Bronx, NY
07/93-06/97 Postdoctoral, Albert Einstein College of Medicine, Bronx, NY

POSITIONS AND HONORS

1996 – 1997 Chief Resident, Dermatology, Albert Einstein College of Medicine (AECOM), NY
1998 – 2004 Assistant Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute (DFCI), Boston, MA
1999 – 2004 Scientific Director, Arthur & Rochelle Belfer Cancer Genomics Center, DFCI, Boston, MA
2005 – 2009 Associate Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2008 – Member, scientific steering committee, International Cancer Genome Consortium (ICGC).
2009 – 2011 Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2009 – Associate Member, the Broad Institute of MIT and Harvard, Boston, MA
2009 – 2011 Co-director, Melanoma Program, Dana-Farber/Harvard Cancer Center, Boston, MA
2009 – 2011 Scientific Director, the Belfer Institute for Applied Cancer Science, DFCI, Boston, MA
2009 – Member, Executive Subcommittee, The Cancer Genome Atlas (TCGA), USA
2011 – Professor and Chair, Department of Genomic Medicine, UTMDACC, Houston, TX
2011 – Scientific Director, Institute for Applied Cancer Science, UTMDACC, Houston, TX

ONGOING RESEARCH SUPPORT

1P01 CA163222 NIH/NCI Fisher (PI) 12/01/11-11/30/16 Role: Project PI
1U01 CA168394 NIH/NCI Mills (PI) 05/01/12-04/31/17 Role: Project PI
R1204 CPRIT Chin (PI) 12/16/11-12/31/16 Role: PI
7P01 CA117969 NIH/NCI DePinho (PI) 04/15/06-12/31/15
7U01 CA141508 NIH/NCI Chin (PI) 08/01/09-07/31/14
5U24 CA143845 NIH/NCI Getz (PI) 08/01/09-07/31/14
U24 CA144025 NIH/NCI Kucherlapati (PI) 08/01/09-07/31/14
5U54 CA163125 NIH/NCI Chin (PI) 08/01/09-07/31/14

PUBLICATIONS (last 5)

Peer-Reviewed Original Research Articles

Cheng, C.S. et al., Nat Commun. 2013 4:2672.
Yen, J. et al. Genome Biol. 2013 14:R113.
Brennan, C.W. et al., Cell 2013 155:462.
Cancer Genome Atlas Research Network, et al. Nat Genet. 2013 45:1113.
Watson, I.R. et al., Nat Rev Genet. 2013 14:703.

International Cancer Genome Consortium

CURRICULUM VITAE

Hannah Cheung, PhD, PMP
HCCheung@mdanderson.org
(713) 301-4311

Education

2001 B.Sc. (Hon) Molecular Genetics, University of Alberta, Canada
2008 Ph.D. Genes and Development, UT-HSC, Houston, Texas
2013 PMP certified July 15, 2013, PMP Number: 1642228

Positions

2013-present	Research Scientist, MD Anderson Cancer Center, Futreal Lab
2013	Freelancer, Cactus Communications, Inc.
2008-2013	Postdoctoral Associate, Texas Children's Hospital, Plon Lab
2002-2008	Graduate Student, UT-HSC, Cote Lab
2001-2002	Research Technologist, von Borstel Lab
1999-2000	Research Intern, Hao/Roa Lab
1998	Research Assistant, Reha-Krantz Lab

Publications (last 5)

Cheung, H.C., San Lucas, F.A., Hicks, S., Bertuch, A.A., Ribes-Zamora, A. An S/T-Q cluster domain census unveils new putative targets under Tel1/Mec1 control *BMC Genomics* (2012) 13: 664 provisional PDF.

Izaguirre, D.I., Zhu, W., Hai, T., **Cheung, H.C.**, Krahe, R. and Cote, G.J. PTBP1-dependent Regulation of USP5 Alternative RNA Splicing Plays a Role in Glioblastoma Tumorigenesis *Molecular Carcinogenesis* (2012) 51: 895-906.

Cheung, H.C., Yatsenko, S.A., Kadapakkam, M., Legay, H., Su, J., Lupski, J.R., and Plon, S.E. Constitutional tandem duplication of 9q34 that truncates *EHMT1* in a child with ganglioglioma *Pediatric Blood & Cancer* (2012) 58: 801-5.

Plon, S.E., Wheeler, D.A., Strong, L.C., Tomlinson, G.E., Pirics, M., Meng, Q., **Cheung, H.C**., Begin, P.R., Muzny, D.M., Lewis, L., Biegel, J.A. and Gibbs, R.A. Identification of Genetic Susceptibility to Childhood Cancer through Analysis of Genes in Parallel. *Cancer Genetics* (2011) 204: 19-25.

Cheung, H.C., Hai, T., Baggerly, K.A., Tsavachidis, S., Krahe, R., and Cote, G.J. Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cells. *Brain* (2009) 132: 2277-88.

Research and Training Support

2009 - 2011 Early Career Award, Thrasher Research Fund
2006 - 2008 Rosalie B. Hite Fellowship

International Cancer Genome Consortium

Sahil Seth

Houston, TX 77054
+1 551 556 8052
sseth@mdanderson.org

Education

Aug 2009-May 2011 | Master of Health Sciences in BIOSTATISTICS, **Johns Hopkins School of Public Health**, Baltimore, MD
MAJOR: Bioinformatics
THESIS: *"Estimation of functional data using functional principal component analysis (FPCA) with applications to diffusion tensor imaging (DTI) tractography data"* | Advisor: Dr. Ciprian Crainiceanu

Aug 2004-Jun 2008 | Bachelor of Technology in BIOTECHNOLOGY, **Amity University**, Noida, UP INDIA
HONORS: Shri Baljeet Shastri Shield (2004-08), a prestigious award of the university awarded for All Round Excellence, encompassing Academic, Leadership and Interpersonal Skills
THESIS: *"Computational studies and molecular modeling of STAT 3 inhibitors"* | Advisor: Dr. Shakti Sahi

Work / Research Experience

Institute Associate Scientist III Inst. of Applied Cancer Sci., MD Anderson Cancer Center, Houston, TX

Jul 2011-*current*
- Collaborate with research scientists to provide computational and statistical solutions to cancer research issues
- Mine through cancer genomics (TCGA) data to look for potential gene targets for cancer therapeutics.
- Optimized processing of genomic data by parallelizing, bringing down computation time (3 days to <24 hours)
- Develop and improve the pipeline to analyze mutations & structural variations in the cancer genome.

Bioinformatics Analyst II, Medical Oncology, Dana-Faber Cancer Institute, Boston, MA

Jul 2011-*current*
- Worked on building analysis tools for next generation sequencing data.
- Developed a database and toolkit to search and browse cell line fingerprint data.

Graduate Research Assistant, Sidney Kimmel Cancer Center, Johns Hopkins School of Medicine

Jan 2010-Dec 2010 | *"Analysis of methylation patterns in cancer cell lines and tumors" "Drug response expression analysis of breast cancer cell line"*

Student Consultant, Biostatistics Center, Johns Hopkins School of Public Health

Mar 2010-May 2010
- Duties included conducting weekly school wide biostatistics consulting clinics.
- Research and analysis for the 'Lead exposure among factory workers study'.

Research Assistant, Dept. of Biostatistics, Johns Hopkins School of Public Health

Oct 2009-May 2010 | *"Development of an R package for initial screening and cleaning of SNP data"*
"Family Based Association Analysis"

Publications / Presentations - recent

(S) A. Viale,...,S. Seth... *"Pancreatic tumor initiating cells resistant to inhibition of oncogenic signaling are dependent on mitochondrial function"*, Nature
(S) The Cancer Genome Atlas Research Network. *"Diversity of Lung Adenocarcinoma Revealed by Integrative Molecular Profiling"*, Nature
(S) The Cancer Genome Atlas Research Network. *"Comprehensive molecular characterization of urothelial bladder carcinoma"*, Nature
The Cancer Genome Atlas Research Network *The Cancer Genome Atlas Pan-Cancer analysis project* Nat Genet. 2013 Oct;45(10):1113-20.
The Cancer Genome Atlas Research Network. *Integrated genomic characterization of endometrial carcinoma*.Nature. 2013 Aug; 497(7447):67-73.
(Poster) S.Seth, C.A. Bristow et. al. *Pan-Cancer Analysis of Mitochondrial DNA Mutations and Aberrations*, TCGA symposium, 2012
(S): submitted

Professional Membership

ASA: American Statistical Association | IMS: Institute of Mathematical Statistics | ISCB: Intl. Society for Comp. Biology

Certificates and Achievements

Jun 2008 | Merit Award for Excellence in organization of Cultural Activities, Amity University
Mar 2007 | Award of Excellence, Indian Institute of Tourism And Future Management Trends, *"Conserve Energy; Save Ecology"*, a self initiated effort of cycling a stretch of 200mi from Delhi to Chandigarh to spreading this message.

CURRICULUM VITAE

JIANHUA (JOHN) ZHANG, PhD
Email: jzhang22@mdanderson.org

PRESENT TITLE AND AFFILIATION

Associate Director, Genomics, Institute for Applied Cancer Science

EDUCATION/TRAINING

Degree-Institution and Location

Yunnan University, PRC, B. Sc. (biology). High distinction.
University of St. Thomas, St. Paul, Minnesota, M. Sc. (software engineering).
University of Western Ontario, London, Ontario, Ph. D. (biology).

POSITIONS AND HONORS

1993 – 1998 Research Scientist**,**
Research Centre, Agriculture and Agri-Food Canada, Harrow, Ontario.
1998 – 2001 Senior Research Associate**,**
Department of Agronomy and Plant Genetics, University of Minnesota
2010 – Present Partner, Bioconductor project
2001 – 2009 Core developer, Bioconductor project
2001 – 2005 Senior Application Developer
Dept. of Biostatistics, Dana-Farber Cancer Institute/Harvard School of Public Health
2005 –2008 Research Scientist, Belfer Institute for Applied Cancer Science
2008 – 2011 Senior Research Scientist, Belfer Institute for Applied Cancer Science
2010 – 2011 Group leader, Belfer Institute for Applied Cancer Science

PUBLICATIONS

The Cancer Genome Atlas Research Network. 2013, *The somatic genomic landscape of glioblastoma*. Cell. 2013 Oct 10;155(2):462-77. doi: 10.1016/j.cell.2013.09.034.

The Cancer Genome Atlas Research Network. 2013. *The Cancer Genome Atlas Pan-Cancer analysis project*. Nat Genet. 2013 Oct;45(10):1113-20. doi: 10.1038/ng.2764.

Hu J, Ho AL, Yuan L, Hu B, Hua S, Hwang SS, **Zhang J**, Hu T, Zheng H, Gan B, Wu G, Wang YA, Chin L, DePinho RA.
From the Cover: Neutralization of terminal differentiation in gliomagenesis.
Proc Natl Acad Sci U S A. 2013 Sep 3;110(36):14520-7. doi: 10.1073/pnas.1308610110. Epub 2013 Aug 5.

The Cancer Genome Atlas Research Network. 2013 *Integrated genomic characterization of endometrial carcinoma.*
Nature. 2013 May 2;497(7447):67-73. doi: 10.1038/nature12113. Erratum in: Nature. 2013 Aug 8;500(7461):242.

The Cancer Genome Atlas Research Network. 2013 *Comprehensive genomic characterization of squamous cell lung cancers.*
Nature. 2012 Sep 27;489(7417):519-25. doi: 10.1038/nature11404. Epub 2012 Sep 9. Erratum in: Nature. 2012

Larman TC, DePalma SR, Hadjipanayis AG; Cancer Genome Atlas Research Network, Protopopov A, **Zhang J**, Gabriel SB, Chin L, Seidman CE, Kucherlapati R, Seidman JG.
Spectrum of somatic mitochondrial mutations in five cancers.
Proc Natl Acad Sci U S A. 2012 Aug 28;109(35):14087-91. Epub 2012 Aug 13.

The Cancer Genome Atlas Research Network. 2011. Integrated genomic analyses of ovarian carcinoma**.** Nature 474**:** 609-615.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mapping patients' data to cell lines.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Levi Garraway, Broad Institute

Gad Getz, Broad Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Mahmoud Ghandi, Broad Institute Gregory Kryukov, Broad Institute	

Background and preliminary dataGC

Human cancer cell lines are invaluable models for cancer research allowing experimental manipulation, mechanistic studies and various high-throughput applications. Recently cell line collections have been used for large-scale pharmacogenomics screens. Such studies raise two important questions. The first question is how well cell lines represent the genetic landscape of primary tumors and metastases for various cancer types. Second, how well the genetic determinants of sensitivity/resistance observed in cell line based systems can be mapped to clinical data on patients response to various therapies.

We propose to investigate these questions though joined analysis of ICGC/TCGA and Cancer Cell Line Encyclopedia (CCLE) data, that we generated for over a thousand cell lines representing 40 different tumor types. In this study, we aim to

i) **map the genetic alterations in ICGC patients to the cancer cell lines** and find how well individual alteration and genetic sub-types are represented by current cell line models. This will allow us to detect "blind spots" of current collections and prioritize genetic alterations for which new cell line models need to be established.

ii) **correlate the ICGC clinical data with drug/shRNA sensitivity in cell lines**. A major advantage of using cell lines is that we can measure the cell line response to different drugs and gene knockdowns. We will analyze whether gene/sensitivity correlations present in clinical data can be detected in cell lines and vice versa.

iii) **identify molecular mechanism of action of some driver mutations and potential ways to target them.** Cancer sequencing studies generate lists of significantly mutated genes, but the biology behind them often remains obscure at the beginning. Mapping recurrent genetic alterations to cell lines allows a potential shortcut though analysis of pathway activation signatures and set of genetic dependencies uncovered though shRNA screens in cell lines harboring mutations of interest

Timelines & resources dedicated to project

We would be only dependent on the initial analysis of ICGC data (the VCF files containing the variant calls: substitutions, indels, rearrangements, copy number, retrotranspositions)

In addition, during the project if other subgroups develop new methods for variant calling, it would be beneficial to run the exact same methods on the cell lines data for consistency of the comparisons).

Research proposal

1-Mapping the genetic alterations in ICGC patients' data to the cancer cell lines data:

We will generate a list of recurrent alterations (including point mutations, indels, gene fusions, copy number alterations, rearrangements, differentially methylated regions, etc.) and intersect that with the Cancer Cell Line Encyclopedia (CCLE) data to find out how well each recurrent alteration is represented in different cancer subtypes.

2-Correlate the ICGC clinical data with drug/shRNA sensitivity in cell lines:

For samples for which therapy response clinical data exist, we will compare genetic/response correlations present in ICGC patients data with the available cell line pharmacological data (from CCLE, CTDD and CGP projects)..

3-Search for pathway activation signatures / phenotypic correlates of recurrent alterations in cell lines (such as shift in metabolic profiles, phospho-proteome, epigenetic data, etc.):

We will use the rich repertoire of cancer cell line data in CCLE, to search for any signature of abnormal metabolite level or DNA methylation pattern or certain signaling protein phosphorylations in the cell line models mapped to any of the genetic alterations in ICGC. We expect to find novel biomarkers that can be used for efficient indication of certain subtypes of cancers or as predictors of drug treatment outcome in the clinic. We also expect that these new findings can create new research opportunities and generate testable hypotheses for new drug targets in the context of specific genetic aberrations.

In summary, we believe that the cancer cell line data is a very rich dataset that can greatly increase our power to interpret a significant subset of the events found in ICGC study and can improve our understanding of the data. This would not be possible without careful mapping of the ICGC data to cancer cell lines data considering the right context for each cancer type and the alterations, and using rigorous statistical analytics which is proposed by this abstract.

Legacy plans

By the end of this project, we will annotate all the recurrent variants in ICGC by the available cell line models in CCLE.

We will also generate a list of cancer subtypes and specific recurrent alterations for which no cell line model currently exists. This can be beneficial for the new cell line generation methods.

We can also develop webpages (or help with implementing those on the ICGC homepage), that allow the researchers to easily access the analysis results and connect the recurrent alterations in the ICGC data to the available cell line data.

CURRICULUM VITAE

Date Prepared: October 8, 2013

Name: LEVI ALEXANDER GARRAWAY

Office Address: Dana Building, Room 1542,
Dana-Farber Cancer Institute
44 Binney Street
Boston, MA 02115

Home Address: 363 Walnut Street
Newton, MA 02460

Work Phone: 617-632-6689

Work E-Mail: levi_garraway@dfci.harvard.edu

Work FAX: 617-582-7880

Place of Birth: Oakland, California

Education

1990	A.B.	Biochemical Sciences	Harvard College, Cambridge, MA
1999	M.D.	Medicine	Harvard Medical School, Boston, MA
1999	Ph.D.	Biological Chemistry and Molecular Pharmacology (Ph.D. Adviser: Dr. Stephen M. Beverley)	Harvard Graduate School of Arts & Sci., Cambridge, MA

Postdoctoral Training

06/93-06/98	Research Assistant	Biological Chemistry and Molecular Pharmacology	Harvard Medical School
06/99-06/01	Resident	Internal Medicine	Massachusetts General Hospital, Boston, MA
07/99-11/02	Clinical Fellow	Medicine	Harvard Medical School
07/01-06/05	Clinical Fellow	Medical Oncology	Dana-Farber Cancer Institute, Boston, MA
07/01-06/05	Clinical Fellow	Medicine	Brigham and Women's Hospital Boston, MA
01/03-12/03	Chief Resident	Medicine	Massachusetts General Hospital

Faculty Academic Appointments

07/05-05/07	Instructor	Medicine	Harvard Medical School
06/07-	Assistant Professor	Medicine	Harvard Medical School

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

A. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997	Military Service - Captain
1997-1998	Tel Aviv. Univ. MSc student
1998-2000	Maximal Innovative Intelligence (part time)
1998-2003	Weizmann Institute of Science. PhD student
2004-2007	Broad Institute of MIT and Harvard. Postdoc
2007-2012	Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013-	Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

5. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
6. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
7. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
8. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
9. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
10. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
11. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
12. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
13. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal

somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
14. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
15. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
16. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
17. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
18. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
19. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

Mahmoud Ghandi
Computational Biologist
Broad Institute of MIT and Harvard
301 Binney St, 5043C, Cambridge, MA 02142
Phone: (571) 230-4394, Email: mghandi@broadinstitute.org

EDUCATION

Ph. D. in Biomedical Engineering, **Johns Hopkins University**, Baltimore, Maryland 2012
Research focus: Chromatin Structure and Gene Regulation in Yeast
M. Sc. in Electrical Engineering, **Sharif University of Technology**, Tehran, Iran 2006
Research focus: Automatic Detection of Emboli using Transcranial Doppler Ultrasound
B. Sc. in Electrical Engineering, **Sharif University of Technology**, Tehran, Iran 2003
Research focus: Coding of Motion Vectors in Video

PROFESSIONAL EXPERIENCES

Broad Institute of MIT and Harvard, Cambridge, Massachusetts
Computational Biologist, Professor Levi Garraway's Laboratory Jun 2012 - present

- Developing state-of-the-art computational methods for integrative analysis of large-scale biological data including next generation sequencing (DNA, RNAseq), proteomics (RPPA, TK-Luminex, Mass Spec), Metabolomics and Epigenetic data. Using these methods to predict sensitivities to drug treatment and shRNA gene knockdown, find new therapeutic targets, and investigate the mechanisms of drug resistance.
- Leading the computational biology effort in the Cancer Cell Line Encyclopedia (CCLE) project-- the world largest public collection of cancer cell line data consisting of over 1000 cancer cell lines)-- for QC, normalization and analysis of Reduced Representation Bisulfite Sequencing (RRBS), NanoString microRNA expression, and Reverse Phase Protein Array (RPPA) data. Also contributed to the experimental design, data generation and analysis of metabolomic profiling and DNA and RNAseq datasets in the CCLE project, collaborating closely with biologists and engineers from both academia and industry.

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland
Research Assistant, advisor: Professor Michael Beer Sep 2006 – May 2012
Developed novel computational analysis and experimental validation:

- To normalize high-throughput genomic assays and statistically remove sequence specific biases
- To predict genome scale nucleosome positioning using a context-based sequence model
- Developed experimental system to validate predictions of mutation impact on gene regulation

Bristol-Myers Squibb Company, East Syracuse, New York
Summer Intern, supervisor: Dr. Ying Jing Jun 2010 – Aug 2010

- Developed a novel method to process RNA-Seq short reads and reconstruct CHO CDNA sequences
- Analyzed Microarray data, designed and performed quantitative PCR and Western blot experiments to identify and verify gene expression changes in dexamethasone treated CHO cells.

NexBio Company, Cupertino, California
Summer Intern, supervisor: Dr. Satnam Alag Jun 2008 – Sep 2008

- Developed context-based document indexing tools for the NextBio.com search engine using JAVA

Advanced Communications Research Institute, Tehran, Iran
Senior Researcher, advisor: Professor Farokh Marvasti Sep 2004 – Aug 2006

- Developed a real-time method to detect micro-emboli in transcranial Doppler ultrasound signal using autoregressive modeling and wavelet transforms
- Developed accelerated adaptive methods for reconstruction of signals from non-uniform samples

Mir Gostar Farda Company, Tehran, Iran
Co-Founder and Software Team Manager Oct 2001 – Sep 2004

- Supervised a group of four computer programmers
- Developed dental patient management software that won Hedayat National Innovation Award
- Designed, prototyped and tested a third generation electronic apex locater

AWARDS AND HONORS

- Ranked 2nd among all students of biomedical engineering at Sharif University of Technology 2005
- Hedayat Innovation Award for software development 2003
- **Silver Medal worldwide** (and 1st in Asia), among more than 1000 universities from 70 countries in annual Association for Computing Machinery International Collegiate Programming Contest 2001
- Programming Excellence Award, Upsilon Pi Epsilon, honor society for computing science 2001
- **Silver Medal**, 11th International Olympiad in Informatics 1999
- **Gold Medal**, 8th National Informatics Olympiad 1998

RELEVANT COURSE WORK

Molecules and Cells, Principles of Immunology, Transcription Mechanisms, Epigenetics, Chromatin/Gene Expression, Computational Functional Genomics, Leaning Theory, Foundations of Optimization, Medical Imaging Systems, Foundations of Computational Biology and Bioinformatics, Digital Signal Processing.

RESEARCH INTERESTS

Analyzing large biological datasets, including Next Gen sequencing and Microarray data, Developing mathematical models for biological processes, Developing tools for automatic data analysis and model inference, Biostatistics, Machine Learning, Computational Biology, Numerical Analysis, Algorithm Design, Optimization.

COMPUTER SKILLS

Advanced programming using C++, JAVA, R, Matlab, Linux, Perl, Python, Bioconductor and SPSS.

SELECTED PUBLICATIONS

- M. Ghandi, D. Lee, M. Mohammad-Noori and M. Beer, 'Enhanced Regulatory Sequence Prediction using Gapped k-mer,' submitted to ***PLoS Computational Biology***.
- O. Botvinnik, W. Kim, C. Birger, J. Rosenbluh, Y. Shrestha, M. Abazeed, et. al, 'Mapping Genomic Alterations to Functional Profiles of Pathway Activation, Gene Dependency and Drug,' submitted to ***Nature Biotechnology***.
- M. Ghandi, M. Mohammad-Noori, and M.A. Beer, 'Robust k-mer frequency estimation using gapped k-mers,' ***Journal of Mathematical Biology***, July 2013.
- S.C. Baca, D. Prandi, M.S. Lawrence, J.M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T.Y. MacDonald, M. Ghandi, M., et al., 'Punctuated evolution of prostate cancer genomes,' ***Cell***, April 2013.
- M. Ghandi and M. Beer, 'Group Normalization for Genomic Data,' ***PLoS ONE***, July 2011.
- Y. Jing, Y. Qian, M. Ghandi, A. He, S. Pan, Z. Jian, 'A Mechanistic Study on the Effect of Dexamethasone in Moderating Cell Death in Chinese Hamster Ovary Cell Cultures,' ***Biotechnology Progress***, Nov 2011.
- M. Ghandi, M. Beer, 'A Novel Sequence Based Model for Nucleosome Positioning in Yeast,' 12th Annual International Conference on Research in Computational Molecular Biology, Dec 2008, Boston, MA.
- M. Ghandi, M. Yekta, F. Marvasti, 'Some Nonlinear/Adaptive Methods for Fast Recovery of the Missing Samples of Signals,' *Elsevier Journal of* ***Signal Processing***, vol. 88, issue 3, Mar 2008.
- S. Marvasti, M. Ghandi, A. Marvasti, A. Deb and H. Markus, 'Improved Detection of Embolic Signals using Multi Scale Wavelet Filtering, AR and ANN, for TCD Ultrasound,' IEEE International Seminar of Medical Applications of Signal Processing, Nov 2005, London, UK.
- M. Ghandi, F. Marvasti, 'Recovery of Missing Samples using Novel and Adaptive Iterative Methods,' International Workshop on Sampling Theory and Applications, July 10-15, 2005, Samsun, Turkey.
- M. Ghandi, M. M. Ghandi, M. B. Shamsollahi, 'A Novel Context Modeling Scheme for Motion Vectors Context-Based Arithmetic Coding,' IEEE Canadian Conference on Electrical and Computer Engineering, May 2004, Ontario, Canada.
- O. Fatemi, M. M. Ghandi, E. Modirzadeh, M. Ghandi, 'Bit-rate Reduction of MPEG Compressed Video,' IEEE Canadian Conference on Electrical & Computer Engineering, May 12-15, 2002, Manitoba, Canada.

Curriculum Vitae for Gregory V. Kryukov, Ph.D.

EDUCATION

Ph.D. in Biochemistry	December 2002, University of Nebraska-Lincoln
M.S. in Physics (with highest honors)	January 1998, Moscow State University, Russia

RESEACH EXPERIENCE

8/2010 – present	Senior Computational Biologist Broad Institute of MIT and Harvard, Cancer Program
1/2010 – present	Instructor in Medicine Harvard Medical School
7/2003 – 12/2009	Postdoctoral Research Fellow Brigham & Women's Hospital, Division of Genetics / Harvard Medical School
1/2003 – 6/2003	Postdoctoral Research Associate University of Nebraska-Lincoln, Department of Biochemistry
8/1998 – 12/2002	Graduate Research Assistant University of Nebraska-Lincoln, Department of Biochemistry
9/1995 – 5/1998	Engelhardt Institute of Molecular Biology, Moscow, Russia Joint Human Genome Program with Argonne National Laboratory (USA) Research Assistant

LIST OF PUBLICATIONS (**10 selected of 48 peer-reviewed original research publications**)

1. Lawrence MS*, Stojanov P*, Polak P*, **Kryukov GV**, …, Getz G. "Mutational heterogeneity in cancer and the search for new cancer-associated genes." Nature (2013) 499:214-218
   **) authors contributed equally to work*
2. Huang FW*, Hodis E*, Xu MJ, **Kryukov GV**, Chin L, Garraway LA. "Highly recurrent TERT promoter mutations in human melanoma" *Science* (2013) 339:957-959.
   **) authors contributed equally to work*
3. McFarland CD, Korolev KS, **Kryukov GV**, Sunyaev SR, Mirny LA "Impact of deleterious passenger mutations on cancer progression" *Proc Natl Acad Sci U S A* (2013) 110:2910-2915.
4. Hodis E*, Watson IR*, **Kryukov GV,** …, Getz G, Garraway LA, Chin L. "A landscape of driver mutations in melanoma" *Cell* (2012) 150, 251-263.
   **) authors contributed equally to work*
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, **Kryukov GV**, …, Garraway LA "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity" *Nature* (2012) 483, 603-607.
6. Price AL*, **Kryukov GV***, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR "Pooled association tests for rare variants in exon-resequencing studies" *Am J Hum Genet* (2010) 86, 832-838
   **) authors contributed equally to work*
7. **Kryukov G.V.**, Shpunt A.A., Stamatoyannopoulos J.A., Sunyaev S.R. "Power of deep all-exon resequencing for discovery of human trait genes" *Proc Natl Acad Sci U S A* (2009) 106, 3871-3876.
8. **Kryukov G.V.**, Pennacchio L.A., Sunyaev S.R. "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies". *Am J Hum Genet* (2007) 80, 727-739.
9. **Kryukov G.V.**, Schmidt S., Sunyaev S. "Small fitness effect of mutations in highly conserved non-coding regions". *Hum Mol Genet* (2005) 14, 2221-2229.
10. **Kryukov G.V.**, Castellano S., Novoselov S.V., Lobanov A.V., Zehtab O., Guigo R., Gladyshev V.N. "Characterization of mammalian selenoproteomes" *Science* (2003) 300, 1439-1443.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Allele-Specific Expression Analysis

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Levi Garraway (DFCI, Broad Institute)
Gad Getz (MGH, Broad Institute)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Gregory Kryukov (Broad Institute, BWH)	Virginia Savova (DFCI) Alexander Gimelbrant (DFCI)

Background and preliminary data

Epigenetic loss of heterozygosity (eLOH) is a common process in normal development, affecting a substantial fraction of human genes. One well-known example is X chromosome inactivation, which is established early in the development of female embryos and maintained in the nuclei of all descendant cells. While DNA for both copies is still present in all the nuclei, mitotically stable silencing of one of the two parental copies results in a functional equivalent of loss of heterozygosity. In an appropriate genetic background, such as hemizygous deletion of a tumor suppressor gene, this can lead to tumor initiation and progression. Only cells with the epigenetic silencing of the "good" allele give rise to tumors, while the cells with the opposite allelic choice remain normal.

We and others have recently discovered that more than 10% of autosomal genes in human and mouse are subject to monoallelic silencing in a way that resembles X-inactivation. In fact, eLOH is much more common than previously appreciated and affects multiple cancer-related autosomal genes. Monoallelic silencing of the affected genes is mitotically stable and leads to cells sharing a particular genome-wide eLOH pattern: a combination of all specific allelic choices in each of multiple loci. Our central hypothesis is that some of these eLOH patterns predispose cells to tumor initiation by causing functional LOH in critical genes.

We propose to systematically explore the role of epigenetic LOH in cancer by applying our expertise in allele-specific expression analysis to the integrated genome sequence data and RNA-seq. Identification of tumor-specific epigenetically-controlled patterns of monoallelic expression would be highly significant: since epigenetic changes are in principle reversible, these new data should lead to chemoprevention, diagnostic and treatment targets.

Timelines & resources dedicated to project

Expected input data:

1) aligned WGS and RNASeq .bam files
2) VCF/MAFs files containing called germline and somatic variant

Research proposal

1) Quantification of allele-specific expression through integrative analysis of WGS and RNASeq data

We will estimate ratio of transcripts generated from two haplotypes for every gene and in all samples, where powered by sufficient expression level and presence of heterozygous germline SNPs to be used as haplotype markers. For that we will use enhanced and modified pipelines previously developed by the investigators for this purpose. To enhance sensitivity of allele-specific expression detection, haplotypes will be statistically phased. Tumor samples purity estimates will be incorporated into analysis.

2) Identification of cis-events that might explain allele-specific expression

We will intersect detected instances of expression allelic imbalance (after excluding those that can be explained by known germline eQTLs) with somatic alterations detected in corresponding samples. It will not only highlight putatively functional promoter/enhancer mutations or fusion events, but also will allow us to estimate what fraction of allele-specific expression can not be explained by somatic cis- alterations.

3) Integrative analysis of allele-specific expression and methylation data
We will analyze the relationship between allele-specific expression and methylation status both on individual genes and whole-genome level.

4) Analysis of global patterns of allele-specific expression

Our preliminary analysis of WGS and RNASeq data for 320 cancel cell lines indicated that global patterns of allele-specific expression (such as general prevalence, number and distribution of lengths of affected gene clusters) vary significantly both within and across different tumor types. We will investigated this phenomena in primary tumors data.

5) Search for factors affecting global patterns of allele-specific expression

We will search for genes and pathways which mutational or expression/activation status correlates significantly with global alteration in allele-specific expression patterns.

Legacy plans

By the end of this project, we will make publically available

1) [Gene]x[Sample] matrix with expression allelic imbalance coefficient
2) A set of summary statistics describing global properties of allele-specific expression for every sample

CURRICULUM VITAE

Date Prepared: October 8, 2013

Name: LEVI ALEXANDER GARRAWAY

Office Address: Dana Building, Room 1542,
Dana-Farber Cancer Institute
44 Binney Street
Boston, MA 02115

Home Address: 363 Walnut Street
Newton, MA 02460

Work Phone: 617-632-6689

Work E-Mail: levi_garraway@dfci.harvard.edu

Work FAX: 617-582-7880

Place of Birth: Oakland, California

Education

1990	A.B.	Biochemical Sciences	Harvard College, Cambridge, MA
1999	M.D.	Medicine	Harvard Medical School, Boston, MA
1999	Ph.D.	Biological Chemistry and Molecular Pharmacology (Ph.D. Adviser: Dr. Stephen M. Beverley)	Harvard Graduate School of Arts & Sci., Cambridge, MA

Postdoctoral Training

06/93-06/98	Research Assistant	Biological Chemistry and Molecular Pharmacology	Harvard Medical School
06/99-06/01	Resident	Internal Medicine	Massachusetts General Hospital, Boston, MA
07/99-11/02	Clinical Fellow	Medicine	Harvard Medical School
07/01-06/05	Clinical Fellow	Medical Oncology	Dana-Farber Cancer Institute, Boston, MA
07/01-06/05	Clinical Fellow	Medicine	Brigham and Women's Hospital Boston, MA
01/03-12/03	Chief Resident	Medicine	Massachusetts General Hospital

Faculty Academic Appointments

07/05-05/07	Instructor	Medicine	Harvard Medical School
06/07-	Assistant Professor	Medicine	Harvard Medical School

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
20. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
21. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
22. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
23. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
24. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
25. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
26. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
27. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
28. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
29. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
30. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
31. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
32. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
33. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

Curriculum Vitae for Gregory V. Kryukov, Ph.D.

366 Broadway USA: Garraway, Getz
Cambridge, MA
02139, USA
Email: savova@gmail.com
Cell: 857-234-4334

EDUCATION

Ph.D. in Biochemistry — December 2002,
University of Nebraska-Lincoln
M.S. in Physics (with highest honors) — January 1998, Moscow State University, Russia

RESEACH EXPERIENCE

8/2010 – present	Senior Computational Biologist Broad Institute of MIT and Harvard, Cancer Program
1/2010 – present	Instructor in Medicine Harvard Medical School
7/2003 – 12/2009	Postdoctoral Research Fellow Brigham & Women's Hospital, Division of Genetics / Harvard Medical School
1/2003 – 6/2003	Postdoctoral Research Associate University of Nebraska-Lincoln, Department of Biochemistry
8/1998 – 12/2002	Graduate Research Assistant University of Nebraska-Lincoln, Department of Biochemistry
9/1995 – 5/1998	Engelhardt Institute of Molecular Biology, Moscow, Russia Joint Human Genome Program with Argonne National Laboratory (USA) Research Assistant

LIST OF PUBLICATIONS (**10 selected of 48 peer-reviewed original research publications**)

A. Lawrence MS*, Stojanov P*, Polak P*, **Kryukov GV**, …, Getz G. "Mutational heterogeneity in cancer and the search for new cancer-associated genes." Nature (2013) 499:214-218
**) authors contributed equally to work*

B. Huang FW*, Hodis E*, Xu MJ, **Kryukov GV**, Chin L, Garraway LA. "Highly recurrent TERT promoter mutations in human melanoma" *Science* (2013) 339:957-959.
**) authors contributed equally to work*

C. McFarland CD, Korolev KS, **Kryukov GV**, Sunyaev SR, Mirny LA "Impact of deleterious passenger mutations on cancer progression" *Proc Natl Acad Sci U S A* (2013) 110:2910-2915.

D. Hodis E*, Watson IR*, **Kryukov GV,** …, Getz G, Garraway LA, Chin L. "A landscape of driver mutations in melanoma" *Cell* (2012) 150, 251-263.
**) authors contributed equally to work*

E. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, **Kryukov GV**, …, Garraway LA "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity" *Nature* (2012) 483, 603-607.

F. Price AL*, **Kryukov GV***, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR "Pooled association tests for rare variants in exon-resequencing studies" *Am J Hum Genet* (2010) 86, 832-838
**) authors contributed equally to work*

G. **Kryukov G.V.**, Shpunt A.A., Stamatoyannopoulos J.A., Sunyaev S.R. "Power of deep all-exon resequencing for discovery of human trait genes" *Proc Natl Acad Sci U S A* (2009) 106, 3871-3876.

H. **Kryukov G.V.**, Pennacchio L.A., Sunyaev S.R. "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies". *Am J Hum Genet* (2007) 80, 727-739.

I. **Kryukov G.V.**, Schmidt S., Sunyaev S. "Small fitness effect of mutations in highly conserved non-coding regions". *Hum Mol Genet* (2005) 14, 2221-2229.

J. **Kryukov G.V.**, Castellano S., Novoselov S.V., Lobanov A.V., Zehtab O., Guigo R., Gladyshev V.N. "Characterization of mammalian selenoproteomes" *Science* (2003) 300, 1439-1443.

Virginia Savova, Ph.D.

Research

2010 – present Dana-Farber Cancer Institute, Boston, MA
Postdoctoral Fellow. High-throuput Next-Gen Sequencing Analysis. Allele-Specific Expression.

2010 – 2013 Dana-Farber Cancer Institute and Broad Institute, Cambridge, MA
Joint Postdoctoral Appointee. High-throuput Sequencing Analysis.

2007 – 2010 Computational Cognitive Science Group,
Department of Brain and Cognitive Science,
Massachusetts Institute of Technology, Cambridge, MA
Postdoctoral researcher. Hierarchical bayesian approaches to cognition.

Degrees and certificates

2000 – 2007 Johns Hopkins University, Baltimore, MD
M.A. and Ph.D., Department of Cognitive Science.

1995 – 1999 Harvard University, Cambridge, MA
A.B. cum laude, Linguistics.

Recent publications

Anwesha Nag*, **Virginia Savova***, Ho-Lim Fung, Alexander Miron, Guo-Cheng Yuan, Kun Zhang**, Alexander A. Gimelbrant**. "Chromatin signature of widespread monoallelic expression". **Elife,** 2013; vol. 2.

Virginia Savova, Sebastien Vigneau, **Alexander A. Gimelbrant**. Autosomal monoallelic expression: genetics of epigenetic diversity? ***Curr Opin Genet Dev*** 2013; v. 23, pp. 642–648

Virginia Savova and **Alexander A. Gimelbrant**. Autosomal monoallelic expression. In **Epigenetics and Complex Traits**: Anna K. Naumova (editor). Springer 2013. ISBN: 9781461480778

Klochendler, Agnes; Weinberg-Corem, Noa; Moran, Maya; Swisa, Avital; Pochet, Nathalie; **Savova, Virginia**; Vikes, Jonas; Van de Peer, Yves; Brandeis, Michael; **Regev, Aviv**; Nielsen, Finn Cilius; Dor, Yuval; Eden, Amir. A Transgenic Mouse Marking Live Replicating Cells Reveals In-Vivo Transcriptional Program of Proliferation." **Developmental Cell**, 2012; vol. 23(4), pp.681 - 690.

Curriculum Vitae Alexander Gimelbrant, PhD

Contact

Dana-Farber Cancer Institute
450 Brookline Ave. Smith SM922B
Boston MA 02215

email: gimelbrant@genetics.med.harvard.edu
office: 617-582-7326

Education

1996	PhD, Biochemistry	Moscow State University, Russia
1992	BS, Biophysics	Tashkent University, Uzbekistan

Professional experience

2011 - present	Assistant Professor, Department of Genetics, Harvard Medical School, Boston, MA
2008 - present	Assistant Professor, Department of Cancer Biology, Dana-Farber Cancer Institute, and Harvard Medical School, Boston, MA
2005 - 2008	Postdoctoral Fellow, Harvard Medical School and Massachusetts General Hospital, Boston, MA
2000 - 2005	Postdoctoral Associate, Whitehead Institute of Biomedical Research, Cambridge, MA
1996 - 2000	Postdoctoral Scholar, Department of Physiology, University of Kentucky, Lexington, KY
1992 - 1996	Graduate student, A.N. Belozersky Institute, Moscow University, Russia

Honors

2010 – 2014	Pew Scholar in the Biomedical Sciences
2008 – 2010	Claudia Adams Barr Investigator, Dana-Farber Cancer Institute, Boston, MA
1998	Grass Fellow, Marine Biological Laboratory, Woods Hole

Selected publications

*Nag, A., *Savova, V., Fung, H., Miron, A., Yuan, G.-C., Zhang, K., Gimelbrant, A.A. (2013). Chromatin signature of widespread monoallelic expression. ***eLife.*** 2: e01256. **equal contribution*.

Savova, V., Vigneau, S., and Gimelbrant, A.A. (2013). Autosomal monoallelic expression: genetics of epigenetic diversity? ***Current Opinion Genetics Devel.*** 23(6), 642-648.

Zwemer, L.M., Zak, A., Thompson, B.R., Kirby, A., Daly, M.J., Chess, A., Gimelbrant, A.A. (2012). Autosomal monoallelic expression in the mouse. ***Genome Biol.*** 13(2):R10.

*Lengner, C.J., *Gimelbrant, A.A., Erwin, J.A., Cheng, A.W., Guenther, M.G., Welstead, G.G., Alagappan, R., Frampton, G.M., Xu, P., Muffat, J., Santagata S., Powers D., Barrett C.B., Young R.A., Lee J.T., Jaenisch R., Mitalipova M. (2010). Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. ***Cell.*** 141: 872-883. **equal contribution*.

Gimelbrant, A., Hutchinson, J.N., Thompson B.R., and Chess, A. (2007) Widespread monoallelic expression on human autosomes. ***Science.*** 318(5853):1136-40.

Gimelbrant, A.A., and Chess, A. (2006) An epigenetic state associated with areas of gene duplication. ***Genome Res.*** 16(6):723-9.

Gimelbrant, A.A., Ensminger, A.W., Qi, P., Zucker, J., and Chess, A. (2005) Monoallelic expression and asynchronous replication of p120 catenin in mouse and human cells. ***J. Biol. Chem.*** 280(2):1354-9.

Gimelbrant, A.A., Skaletsky, H., and Chess, A. (2004) Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence. ***Proc. Nat. Acad. Sci. USA*** 101(24):9010-22.

*Singh N., *Ebrahimi F.A., Gimelbrant A.A., Ensminger, A.W., Tackett, M.R., Qi, P., Gribnau, J., and Chess, A. (2003) Coordination of the random asynchronous replication of autosomal loci. ***Nat. Genet.*** 33(3):339-41. **equal contribution*.

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

Analysis of structural variation breakpoints & relating them to fusion genes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (No more than 2; 1 page CV for each)

Mark Gerstein, Yale (TCGA Prostate AWG); Kevin White, UChicago (TCGA Breast and Ovarian RNA-seq/gene fusion AWGs)

Name(s) & institute(s) of junior investigators	**Name(s) & institute(s) of non-ICGC collaborators**
Alexej Abyzov, Yale University; Andrea Sboner, Cornell U	Robert Grossman, U Chicago, TCGA Protected Data Cloud Mark Rubin, Cornell U, TCGA Prostate AWG

Background and preliminary data

SVs. Genome structural variations (SVs) such as deletions, duplications, translocations, inversions, retrotranspositions, and more complex rearrangements are present in all types of cancer. While SVs can arise as a result of replicative errors as single events, they can also be caused by catastrophic genome rearrangement events (chromothripsis) (Stephens, Cell, 2011) and mediated by viral integration (Akagi, Genome Research, 2013). SVs have been observed to cluster with certain SNVs (Roberts, Mol. Cell, 2012; Nik-Zainal, Cell, 2012), and it has been observed that SV breakpoints are associated with high methylation in some cancer types (Lin, Neoplasia, 2013). However, mechanisms of SV occurrence, their mutation signatures relevant to the mechanisms and/or cancer types, association with other variant types (SNVs and indels), and functional consequence have not been completely understood.

The Gerstein group has extensive experience in finding SVs, resolving them to single-nucleotide resolution, characterizing their breakpoints, and relating them to functional elements (Korbel, Science, 2007; Lam, Nature Biotech, 2010; Mills, Nature, 2011; Abyzov, Genome Res, 2011; Abyzov, Bioinformatics, 2011; Mu, NAR, 2011; Abyzov, Nature, 2012; Abyzov, Genome Res, 2013; Khurana, Science, 2013). We found that classification of SVs into different classes of likely origin mechanisms provides insight into SV association with chromatin states and other mutation types. For germline variants (Fig.1, left), SVs generated by Non-Allelic Homologous Recombination (NAHR) were enriched with enhancers and associated with active chromatin marks (Khurana, Science, 2012). Whether similar is true for cancer SVs is unknown.

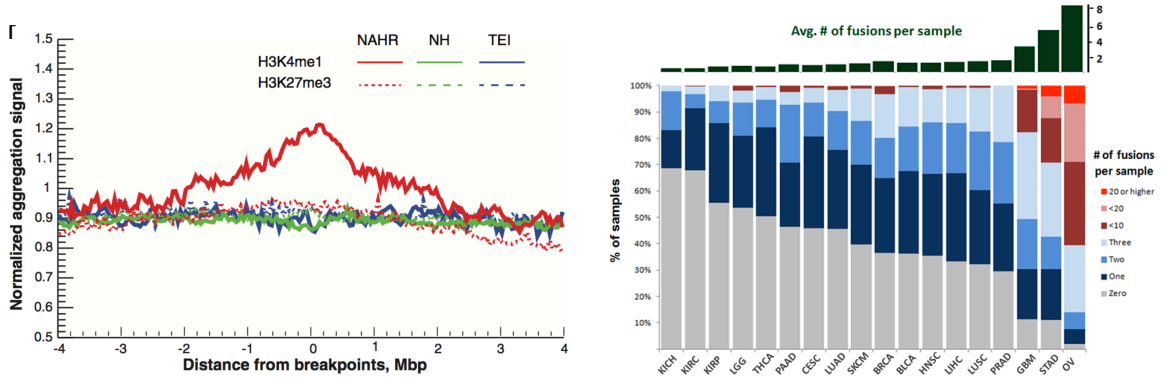


Fig. Left: Aggregation signals for an activating histone mark (H3K4me1) and a repressive mark (H3K27me3) shown.
Right: Frequency distribution of gene fusions by UC-Fusion-Finder from 4,900 primary tumor RNAseq (unpublished).

Fusions. When they intersect genes, SVs can give rise to gene fusions. We have also previously performed analyses for characterizing gene fusion, i.e. chimeric transcripts composed by two or more genes in relation to cancer (Sboner, Genome Biology, 2011; Pflueger, Genome Res., 2011; Demichelis, Genes Chrom. Cancer, 2009; Tanas, Sci Transl Med, 2011). Such fusions are typically a result of chromosomal rearrangements and represent examples of functional consequences of alterations in non-coding regions, where an oncogene is over-expressed because of the strong promoter activity of its 5' partner or a novel oncogenic protein is generated (e.g., BCR-ABL1 fusion protein in CML). In fact, the most clinically relevant SVs are those that generate fusion genes. Many fusions have been identified but their distribution across cancers and their association with SVs has not been studied in detail. The Gerstein and Rubin groups have considerable experience developing the FusionSeq tool for fusion finding and applying it to cancer/TCGA data sets (Sboner, Genome Biology, 2011; Pflueger, Genome Res., 2011; Demichelis, Genes Chrom. Cancer, 2009; Tanas, Sci Transl Med, 2011). In prostate cancer we identified recurrent ETS-TMPRSS2 and other gene fusions with a few base-pair resolution (Pflueger, Genome Res., 2011; Demichelis, Genes Chrom. Cancer, 2009). In parallel studies, the White group's fusion analysis of 4,900 TCGA tumors using "UC-Fusion-Finder" (in preparation) revealed a broad distribution of fusion frequencies across tumor types (Fig. 1, right). In particular, we find a number of tumor-type pairs showing an inverse relationship between the number of somatic copy number alterations (SCNAs) and fusion events, e.g., for glioblastoma (GBM) and bladder cancers (BLCA) (Fig. 1, right; Zack, Nature Gen, 2013; unpub.). It is yet to be determined whether this is due to CN neutral events not detectable by SNP arrays, trans-splicing events or an enrichment of a certain type of SVs.

International Cancer Genome Consortium

Timelines & resources dedicated to project
SV and Fusion calling Jan-July 2014; Scientific analysis Jan-Dec 2014; Manuscript preparation Jan-Feb 2015. Equivalent of 2 FTE from Gerstein, Rubin, Sboner and White labs will work on the project. We will largely rely on generated call sets but will generate some variant calls by ourselves. Resources: 1 TB of disk space (5 TB is better), 10 years of CPU time.
Research proposal
SVs. We propose calling somatic CNVs with CNVnator (Abyzov, Genome Res, 2011) and somatic retroduplications using our new software (Abyzov, Genome Res, 2013). Using TIGRA-SV (http://gmt.genome.wustl.edu/tigra-sv) software and AGE aligner (Abyzov, Bioinformatics, 2011) we will resolve breakpoints with single nucleotide resolution for the most SVs possible. These would include SVs sites discovered as a result of core calling and discovered by us. Next, we will classify SVs into the likely mechanism of their origin by BreakSeq pipeline (Lam, Nature Biotech, 2010). Then we suggest performing the analyses **in different cancer types** and comparing results **between different cancer types**: • Analyze complexity of breakpoints, e.g., presence, content and origin of additional sequence at breakpoints; • Investigate clustering and recurrence across genome, samples and different cancer types of SVs generated by different mechanism; • Analyze co-association of breakpoints with other mutations (SNVs and indels) including clustered mutation by, presumably, APOBEC enzyme. Here nucleotide resolution of breakpoints will allow us to see whether any association on a small scale (100-1000 bps) is apparent, as it was, for instance, observed recently for de novo mutations (Carvalho, Nature Gen, 2013); • Correlation location of SV breakpoints with: replication timing, fragile sites for DNA double stranded breaks (Crosetto, Nature Meth, 2013), G-quadruplex motifs (Paeschke, Nature, 2013) that can cause replication stalling, recombination hotspots, methylation, histone mark from ENCODE project, high resolution Hi-C data (Jin, Nature, 2013), evolutionary conservation scores (GERP and PhastCons); • Analyze enrichment/depletion of SVs intersecting various functional elements: genes, enhancers, promoters, lncRNAs, miRNAs, piRNAs, TF binding sites/motif, and Ultra Conserved Elements (UCE). Here, we will elaborate on analysis SV/CNV intersection with dosage sensitive genes, single copy genes, ohnologs (Makino, Nature Comm, 2013), and UCE, as the latter were suggested as regions for copy-number checks (Derti, Nature Genetics, 2006). • The results for the above analyses will be compared to the similar analysis for germline mutations, as discovered by the 1000 Genomes Project. **Fusions.** We will discover gene fusions from transcriptomes with FusionSeq (Sboner, Genome Biology, 2011) and UC-Fusion-Finder. We will search for genomic evidence of the fusions in WGS in three increasing steps of sensitivity. First, we will correlate identified fusions with chromosomal translocations. Second, we will use the breakpoints identified by TIGRA-SV and AGE aligner to identify fusions from WGS. Third, to account for fusions originating from complex genomic rearrangements (Wu, Genes Chromo Cancer 2012), we will construct a breakpoint graph for each WGS sample and traverse it to identify genes fused with intervening unknown sequences. Besides validating RNAseq based fusion gene discovery in a number of samples, whole genome sequence analysis as part of ICGC PanCancer will help to determine how the genome-wide distributions of SVs relate to fusion gene distributions, SCNAs, indels, SNVs, methylation patterns and chromatin marks. For this purpose we will do the following analyses: - Investigate association between fusion events and characteristics of genomic instability: somatic mutations and chromosomal rearrangements. - Correlate fusions data with DNA sequence characteristics (fragile sites, etc.), replication timing, transcription state and ENCODE marks. Investigate enrichment of these features with recurrent fusions both within and across tumor types. - Identify enrichment of DNA repair genes/pathways associated with enrichment of fusion events - Infer the type of SV that generated each fusion event and determine enrichment of SV types with fusion events and how it varies by tumor type - Explore association between fusions and chromothripsis within and across tumor types - Investigate the variation in genomic rearrangements and fusion events between primary and metastatic tumor samples. Identify correlations with clinical data (chemotherapy, etc) - Attempt to identify trans-splicing events by using RNAseq-only fusion calls. We will remove fusions with insufficient coverage in WGS, control for library preparation artifacts by considering highly expressed fusions (Houseley, PLoS One, 2010), and remove false positives associated with FusionSeq or UC-Fusion-Finder by generating a blacklist from >400 normal TCGA transcriptomes. **Relation to ENCODE.** Finally, results will be compared to evolving ENCODE maps of the human genome annotation. (Both Gerstein's and White's groups are actively engaged in ENCODE data analyses, and Gerstein and White co-chair the ENCODE & Cancer subgroup of the AWG.) As discussed above SVs and fusions will be compared to ENCODE annotations and other types of ENCODE data across multiple cell types – including cell types that may be directly relevant to certain cancer datasets (e.g, MCF-7 cell data and breast cancer, or K562 cells and leukemias).
Legacy plans
All generated results of the suggested studies will be published in peer-reviewed journals. Tools and software created during the course of the suggest studies will be publicly available under Creative Commons License.

Mark Gerstein

Education

Harvard College, AB Physics '89
Cambridge University, PhD Chemistry '93
Stanford University, postdoc '93-'96, Bioinformatics (advisor M Levitt)

Positions

2006- **AL Williams Prof. Biomedical Informatics, Yale**
2002- co-director Yale Computational Biology and Bioinformatics Program
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

Honors

'89-'93 Herchel-Smith Scholarship for PhD at Cambridge
'93-'96 Damon Runyon-Walter Winchell post-doctoral Fellowship
'09 AAAS Fellow

Consortia

Analysis co-chair: NHGRI modENCODE Project AWG ('07-), Brainspan Project ('09-), 1000 Genomes Functional Interpretation Group ('12-), ENCODE & Cancer Group ('13-) exRNA consortium ('13-)

Publications (senior author on all papers listed below, which are selected from a total of >460; H-index=116)

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, M Gerstein (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342:1235587

E Khurana, Y Fu, J Chen, M Gerstein (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comp Bio* 9:e1002886.

M Gerstein, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." *Nature* 489:91

A Abyzov, J Mariani... (16 authors)... M Gerstein, FM Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells." *Nature* 492:438

B Pei, C Sisu... (10 authors)... J Harrow, M Gerstein (2012). "The GENCODE pseudogene resource." *Genome Biol* 13:R51.

C Cheng, R Alexander... (16 authors)... M Gerstein (2012). " Understanding transcriptional regulation by integrative analysis of transcription factor binding data." *Genome Res* 22:1658.

DG MacArthur, S Balasubramanian... (50 authors)... M Gerstein, C Tyler-Smith (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." *Science* 335:823.

A Abyzov, AE Urban, M Snyder, M Gerstein (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21:974

A Sboner, L Habegger... (9 authors)... MA Rubin, M Gerstein (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." *Genome Biol* 11:R104.

HY Lam, XJ Mu, AM Stütz, A Tanzer, PD Cayting, M Snyder, PM Kim, JO Korbel, M Gerstein (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." *Nat Biotech* 28:47.

RP Alexander, G Fang, J Rozowsky, M Snyder, M Gerstein (2010). "Annotating non-coding regions of the genome." *Nat Rev Genet* 11:559.

KK Yan, G Fang, N Bhardwaj, RP Alexander, M Gerstein (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107:9186.

N Bhardwaj, KK Yan, M Gerstein (2010). "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels." *PNAS* 107:6841

M Gerstein, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330:1775.

Kevin P. White

Education

Yale University, New Haven, B.S./M.S., Biology 1993
Stanford University, Stanford, CA, Ph.D., Developmental Biology 1998
Stanford Genome Technology Ctr, Palto Alto, CA, Postdoc, Biochemistry & Genomics, 1998-2000

Professional Positions

2006-present	Director, Joint Institute for Genomics & Systems Biology, The University of Chicago and Argonne National Laboratory
2006-present	James and Karen Frank Family Professor, Human Genetics and Ecology & Evolution, The University of Chicago
2004-2006	Associate Prof. of Ecology & Evolutionary Biology (joint appointment), Yale University
2004-2006	Associate Professor of Genetics, Yale University School of Medicine
2001-2004	Assistant Professor of Genetics, Yale University School of Medicine

Publications Selected from 97 peer-reviewed publications

1. Michelle N. Arbeitman, Eileen E. M. Furlong, Farhad Imam, Eric Johnson, Brian H. Null, Bruce S. Baker, Mark A. Krasnow, Matthew P. Scott, Ronald W. Davis and Kevin P. White. Gene Expression During the Life Cycle of Drosophila melanogaster. **Science,** 297: 2270-2275, **2002.**
2. Giot L, Bader JS, Brouwer C, Chaudhuri, et al. A genome-scale protein interaction map of Drosophila melanogaster. **Science**, 302: 1727-36, **2003.**
3. Viktor Stolc*, Zareen Gauhar*, Christopher Mason*, Gabor Halasz, Marinus F. van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Barbano, Harmen J. Bussemaker, and Kevin P White. A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster. **Science**, 306:655-60, **2004.**
4. Scott Rifkin, David Houle, Junhyong Kim and Kevin P. White. A mutation accumulation assay reveals extensive capacity for rapid gene expression evolution. **Nature,** 438:220-3, **2005.**
5. Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed and Kevin P. White. "Expression profiling in primates reveals a rapid evolution of human transcription factors." **Nature,** 440:242-5, **2006**.
6. Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Nègre N, Ludwig M, Stricker T, Al-Ahmadie HA, Tretiakova M, Camp RL, Perera-Alberto M, Rimm DL, Xu T, Rzhetsky A, White KP. Analysis of Drosophila Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer. **Science,** 323:1218-22, **2009**.
7. Hua SJ, Kittler R, and White KP. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. **Cell.** 137:1259-71, **2009**.
8. modENCODE Consortium, et, al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. **Science**. 330:1787-97. **2010**
9. Nègre N*, Brown CD*, Ma L*, Bristow CA*, Miller S*, Kheradpour P, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony R, Chen J, Hwang L, Wagner U, Auburn R, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Bild NA, Grundstad AJ, Hanley D, Mannervik M, Venken K, Bellen H, White R, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. A cis-regulatory map for the Drosophila genome. **Nature**. 471:527-31. **2011**.
10. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. **Nature.**489:57-74. 2012.:
11. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumors, **Nature.** 490:61-70. 2012
12. Xiaochun Ni, Yong E. Zhang, Nicolas Negre, Sidi Chen, Manyuan Long and Kevin P. White. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. **PLoS. Biology**. 10(11):e1001420. 2012.
13. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. **Blood**. 121: 975-83. 2013.
14. Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP. A comprehensive nuclear receptor network for breast cancer cells. **Cell Rep.** 3:538-51. 2013.
15. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. **Cell**. Sep 26;155:70-80. 2013.

Alexej Abyzov, Ph.D.

Education:

2008	Ph.D. in computational biology, Northeastern University, Boston, MA
2002	M.S. in physics, Moscow Institute of Physics and Technology, Moscow, Russia
2000	B.S. in physics, Moscow Institute of Physics and Technology, Moscow, Russia

Positions:

2012	Associate Research Scientist in computational biology at Yale University
2008-2012	Postdoctoral associate in computational biology at Yale University
2002-2008	Research assistant, Northeastern University
2002	Member of the CMS collaboration at CERN
2000-2002	Research assistant in JINR; member of the HERA-B collaboration at DESY

Professional Honors and Recognition:

2007	Member of The Honor Society of Phi Kappa Phi
2001	Scholarship from European Physical Society for spring school at University of Pavia
2000	Scholarship from DAAD for practical training at DESY

Publications relevant to the proposal:

1. Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, The 1000 Genomes Project Consortium, Lee C, Gerstein MB. **Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division.** *Genome Res. (in press)*
2. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, et al. **Integrative annotation of variants from 1,092 humans: application to cancer genomics.** *Science. 2013 Oct 4;342(6154):1235587.*
3. Abyzov A, Mariani J, Palejev D, Zhang Y, Haney MS, Tomasini L, Ferrandino A, Belmaker LR, Szekely A, Wilson M, Kocabas A, Calixto NE, Grigorenko EL, Huttner A, Chawarska K, Weissman S, Urban AE, Gerstein MB, Vaccarino FM. **Somatic copy-number mosaicism in human skin revealed by induced pluripotent stem cells.** *Nature. 2012 Dec 20;492(7429):438-42.*
4. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, et al. **Architecture of the human regulatory network derived from ENCODE data.** *Nature. 2012 Sep 6;489(7414):91-100.*
5. The ENCODE Project Consortium. **An Integrated Encyclopedia of DNA Elements in the Human Genome.** *Nature. 2012 Sep 6;489(7414):57-74.*
6. Iskow RC, Gokcumen O, Abyzov A, Malukiewicz J, Zhu Q, Sukumar AT, Pai AA, Mills RE, Habegger L, Cusanovich DA, Rubel MA, Perry GH, Gerstein M, Stone AC, Gilad Y, Lee C. **Regulatory element copy number differences shape primate expression profiles.** *Proc Natl Acad Sci U S A. 2012 Jul 31;109(31):12656-61.*
7. Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M**. Genome-Wide Mapping of Copy Number Variation in Humans: Comparative Analysis of High Resolution Array Platforms.** *PLoS One. 2011;6(11):e27859.*
8. Rozowsky J*, Abyzov A*, et al. **AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework.** *Mol Syst Biol. 2011 Aug 2;7:522.*Equal contribution authors*
9. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, Gerstein M. **Identification of genomic indels and structural variations using split reads.** *BMC Genomics. 2011 Jul 25;12(1):375.*
10. Abyzov A, Urban AE, Snyder M, Gerstein M. **CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.** *Genome Res. 2011 Jun;21(6):974-84.*
11. Abyzov A*, Gerstein MB. **AGE: Defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision.** *Bioinformatics. 2011 Mar 1;27(5):595-603. *Corresponding author*
12. Mills RE*, Walter K*, Stewart C*, Handsaker RE*, Chen K*, Alkan C*, Abyzov A*, Yoon S*, Ye K*, et al., The 1000 Genomes Project Consortium. **Mapping structural variation at fine scale by population scale genome sequencing.** *Nature. 2011 Feb 3;470(7332):59-65.*Equal contribution authors*
13. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. **PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.** *Genome Biol. 2009 Feb 23;10(2):R23.*
14. Wang LY, Abyzov A, Korbel JO, Snyder M, Gerstein M. **MSB: A mean-shift-based approach for the analysis of structural variation in the genome.** *Genome Res. 2009 Jan;19(1):106-17.*

Andrea Sboner

Education

University of Trento, Italy MSc Physics 1998
University of Trento, Italy, PhD, Information and Communication technololgy 2005
Yale University, New Haven, CT, Postdoc, Computational Biology and bioinformatics, 2006-2011

Professional Positions

2013 – present Assistant Professor of Pathology and Laboratory Medicine and Assistant Professor of Computational Genomics in Computational Biomedicine
2011 – 2013 Instructor at the Department of Pathology and Laboratory Medicine, Junior Fellow of the Institute of Computational Medicine, Weill Medical College of Cornell University.
2004 – 2006 Researcher, Bioinformatics – Automated Reasoning System Division, ITC-irst, Trento, Italy
1999 – 2003 Researcher, Applied Unit of Medical Informatics and Telemedicine, ITC-irst, Trento, Italy

Publications Selected from 58 peer-reviewed publications

1. Mosquera JM, **Sboner A**, et al. "Novel MIR143-NOTCH Fusions in Benign and Malignant Glomus Tumors" *Genes, Chromosome and Cancer, 2013; 52 (11), 1075-1087*
2. Khurana E, Fu, Y, Colonna V, Mu XJ, Kang XM, Lappalainen T, **Sboner A**, et al. "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics" *Science* 342 (6154), 1235587
3. Antonescu CR, Loarer F, Mosquera JM, **Sboner A**, et al. "Novel YAP1□TFE3 fusion defines a distinct subset of epithelioid hemangioendothelioma" *Genes, Chromosome and Cancer*, 2013 (in press).
4. Baca SA, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, Van Allen E, Kryukov GV, **Sboner A**, et al. "Punctuated Evolution of Prostate Cancer Genomes" Cell 2013;153(3):666-677
5. Mosquera JM, **Sboner A**, et al. "Recurrent NCOA2 Gene Rearrangements in Congenital/Infantile Spindle Cell Rhabdomyosarcoma." *Genes, Chromosome and Cancer* 2013;52(6):538-550.
6. Consortium, The ENCODE Project. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414) (September 6): 57–74. doi:10.1038/nature11247.
7. Habegger, L, Balasubramanian S, Chen DZ, Khurana E, **Sboner A**, et al. "VAT: A Computational Framework to Functionally Annotate Variants in Personal Genomes Within a Cloud-computing Environment." *Bioinformatics* (June 28, 2012:28(17):2267-2269). doi:10.1093/bioinformatics/bts368.
8. Du J, Leng J, Habegger L, **Sboner A**, et al. "IQSeq: Integrated Isoform Quantification Analysis Based on Next-Generation Sequencing", PloS one 2012;7(1):e29175
9. Beltran* H, Rickman* DS, Chae S, **Sboner A**, et al. "Molecular Characterization of Neuroendocrine Prostate Cancer and Identification of New Drug Targets", Cancer Discovery, Nov 2011;1(6):487-495
10. Tanas MR, **Sboner A**, et al. "Identification of a Disease-Defining Gene Fusion in Epithelioid Hemangioendothelioma".Sci Transl Med 2011;3(98):98ra82
11. **Sboner A**, et al. "The real cost of sequencing: higher than you think!" Genome Biology 2011;12(8):125
12. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, **Sboner A**, et al. "The genomic complexity of primary human prostate cancer". Nature, 2011; 470:214–220
13. Habegger L*, **Sboner A***, et al.. "RSEQtools: A modular framework to analyze RNA-Seq data with a concise and confidential format". *Bioinformatics*, 2010; 2011;27:281-283 (* equal contribution)
14. Pflueger D*, Terry S*, **Sboner A***, et al. "Discovery of Non-ETS Gene Fusions in Prostate Cancer using Next Generation RNA Sequencing". Genome Research, 2011;21:56-67 (* equal contribution)
15. The ENCODE Project Consortium, "A User's Guide to the Encyclopedia of DNA Elements (ENCODE)". PLoS Biol 2011;9(4): e1001046. doi:10.1371/journal.pbio.1001046
16. **Sboner A***, Habegger* L, et al. "FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data". *Genome Biology*, 2010; Oct 21;11:R104; (*equal contribution)
17. Gerstein* MB, Lu* ZJ, Van Nostrand* EL, Cheng* C, Arshinoff* BI, Liu* T, Yip* K, Robilotto* R, (+ 87 authors), **Sboner A**, et al. "Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE Project", Science, 2010: 330(6012):1775-1787
18. **Sboner A**, et al. "Molecular sampling of prostate cancer: a dilemma for predicting disease progression". *BMC Med Genomics* 2010;3:8
19. **Sboner A**, et al. "Robust-Linear-Model Normalization To Reduce Technical Variability in Functional Protein Microarrays" *J Proteome Res,* 2009;8:5451
20. Pfluger D, Rickman D, **Sboner A**, et al. "N-myc downstream regulated gene 1 (NDRG1) is fused to ERG in prostate cancer" *Neoplasia* 2009;11:804
21. Demichelis F, Setlur S, Beroukhim R, Perner S, Korbel J, LaFargue C, Pfluger D, Pina C, Hofer M, **Sboner A**, et al. "Distinct genomic aberrations associated with ERG rearranged prostate cancer" *Genes Chromosomes Cancer* 2009;48:366
22. Setlur S, Mertz K, Hoshida Y, Demichelis F, Lupien M, Perner S, **Sboner A**, et al. "Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer" *J Natl Cancer Inst* 2008;100:815
23. Kim* PM, **Sboner* A**, Xia Y, Gerstein MB. "The Role of Disorder in Interaction Networks: A Structural Analysis" *Mol Sys Biol* 2008;4:179 (* equal contribution)
24. Setlur SR, Royce TE, **Sboner A**, Mosquera J, Demichelis F, Hofer MD, Mertz KD, Gerstein MB, Rubin MA "Integrative Microarray Analysis of Pathways Dysregulated in Metastatic Prostate Cancer" *Cancer Res*, 2007;67:10296

Robert Grossman

Education

Harvard College, AB Mathematics, 1980
Princeton University, PhD Applied Mathematics, 1985
University of California, Berkeley, Postdoc, 1984-1988

Positions

2011 – present, Chief Research Informatics Officer, Biological Sciences Division, University of Chicago
2010 – present, Professor of Medicine, Section of Genetic Medicine, University of Chicago
1988 – 2010, Professor of Mathematics, Statistics & Computer Science, University of Illinois at Chicago
(Assistant Professor, 1988 – 1991; Associate Professor, 1991 – 1995; Professor 1995 – 2010)

Honors

2013 AAAS Fellow
2013 Federal 100 Award Winner

Publications

1. David R. Blair, Christopher S. Lyttle, Jonathan M. Mortensen, Charles F. Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V. Bernstam, Søren Brunak, Lars Juhl Jensen, Dan Nicolae, Nigam H. Shah, Robert L. Grossman, Nancy J. Cox, Kevin P. White, Andrey Rzhetsky, A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk, Cell Volume 155, Issue 1, pages 70-80. PMID: 24074861, PMCID: in progress.
2. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP, CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia, Blood, Volume 121, Number 6, pages 975-983, 2013, PMID: 23212519, PMCID: PMC3567344.
3. Heidi L. Alvarez, Malcolm Atkinson, Robert L. Grossman, Matthew Greenway, Christine Harvey, Allison P. Heath, Iraklis Klampanos, Joe J. Mambretti, Ray Powell, Rafael D. Suarez, Walt Wells and Kevin White, The Design of a Community Science Cloud: The Open Science Data Cloud Perspective, SC Companion: High Performance Computing, Networking Storage and Analysis, ACM Press, 2012.
4. Shi Yu, Robert Grossman and Andrey Rzhetsky, Global and Local Approach of Part-of-Speech Tagging for Large Corpora, AAAI-2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text, 2012.
5. Wenxuan Gao, Robert Grossman, Philip Yu, Christopher Brown, Matthew Slattery, Lijia Ma and Kevin White, Discovering Geometric Patterns in Genomic Data, ACM Conference on Bioinformatics, Computational Biology and Biomedicine, ACM Press, 2012.
6. Robert L. Grossman and Kevin P. White, A vision for a Biomedical Cloud, Journal of Internal Medicine, Volume 271, Number 2, pages 122-130, 2012. PMID: 22142244. PMCID: in process.
7. Xin Feng, Robert L. Grossman and Lincoln Stein, PeakRanger: a Cloud-enabled Peak Caller for ChIP-seq Data, BMC Bioinformatics, Volume 12:139, PMID: 21554709, PMCID: PMC3103446.
8. Nicolas Negre, Christopher D. Brown, Lijia Ma, et. al., Cis-Regulatory Map of the Drosophila Genome, Nature, Volume 471, pages 527–531, 2011, [doi:10.1038/nature09990], PMID: 21430782. PMCID: in process.
9. The modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V. Kharchenko, et. al., Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE, Science, Volume 330 (6012), pages 1787-1797, 2010, [DOI:10.1126/science.1198374]. PMID: 21177974. PMCID: in process.
10. Robert L. Grossman, Yunhong Gu, Joe Mambretti, Michal Sabala, Alex Szalay, and Kevin White, An Overview of the Open Science Data Cloud, Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC '10), ACM, 2010. PMCID: in process.

Mark A. Rubin

Education
B.S. University of Wisconsin, Madison, WI, 1984
M.D. Mount Sinai School of Medicine, NY, 1988

Positions

2007-	Professor of Pathology and Laboratory Medicine, Vice Chair for Experimental Pathology, Weill Cornell Medical College
2009-	Homer T. Hirst Professor of Oncology in Pathology, Weill Cornell Medical College
2013-	Director, Institute for Precision Medicine, Weill Cornell Medical College and New York-Presbyterian Hospital
2006-2009	Associate Member, Broad Institute of Harvard and MIT
2006-2007	Staff Physician, Dana Farber Cancer Institute
2002-2007	Associate Professor of Pathology and Chief of Genitourinary Pathology, Brigham and Women's Hospital, Harvard Medical School

Selected Honors
2007 Team Science Award (Co-Leader with Arul Chinnaiyan), American Association for Cancer Research
2012 GU ASCO Keynote Lecture, "Insights from Genomic Approaches to Oncology Discovery"
2012 Huggins Award, Society of Urologic Oncology
2013 Damon Runyon Cancer Research Foundation Clinical Investigator Award (Mentor for H. Beltran)
2013 Prostate Cancer Foundation Mentor of Excellence Award

Committees
Chair, EDRN Prostate Group, NCI (2010-); Chair, PCRP EAB, Department of Defense (2011-); Executive Director, New York Genome Center Executive Committee (2012-)

Publications (selected from over 285)
Dhanasekaran SM…**Rubin MA***, Chinnalyan AM*. Delineation of prognostic biomarkers in prostate cancer. Nature 2001;412:822-826. (1489 Citations) *Co-senior author
Rubin MA…Chinnaiyan AM. alpha-Methylacyl-CoA racemase as a tissue biomarker for prostate cancer. JAMA 2002;287:1662-1670. (567 Citations)
Varambally S…**Rubin MA**, Chinnaiyan AM. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature. 2002 Oct 10;419(6907):624-9. (1342 Citations)
Shah RB…**Rubin MA***, Pienta KJ. Androgen-independent prostate cancer is a heterogeneous group of diseases: lessons from a rapid autopsy program. Cancer Res. 2004 Dec 15;64(24):9209-16. (356 Citations) *Co-senior suthor
Tomlins SA…**Rubin MA**, Chinnaiyan AM. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644-8. (1748 Citations)
Perner S…**Rubin MA***. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res. 2006;66(17):8337-41. (342 Citations) *Senior and corresponding author
Demichelis F…**Rubin MA**. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a Watchful Waiting cohort. Oncogene 2007. (360 Citations) *Co-senior and corresponding author
Berger MF…**Rubin MA***, Garraway LA*. The genomic complexity of primary human prostate cancer. Nature. 2011 Feb 10;470(7333):214-20. (317 Citations) *Co-senior and corresponding author
Rickman DS…Rubin MA. Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci U S A. 2012 Jun 5;109(23):9083-8. (22 Citations)
Demichelis F…**Rubin MA**. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A. 2012 Apr 24;109(17):6686-91. (5 Citations)
Barbieri CE…**Rubin MA***, Garraway LA*. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012 May 20;44(6):685-9. (118 Citations) *Co-senior and corresponding author
Beltran H…**Rubin MA***. Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. Cancer Discov. 2011 Nov;1(6):487-95. (34 Citations) *Senior and corresponding author
Mosquera JM…**Rubin MA**. Concurrent AURKA and MYCN gene amplifications are harbingers of lethal treatment-related neuroendocrine prostate cancer. Neoplasia. 2013 Jan;15(1):1-10.
Baca SC…**Rubin MA***, Garraway LA*. Punctuated evolution of prostate cancer genomes. Cell. 2013 Apr 25;153(3):666-77. *Co-senior author
Khurana E…**Rubin MA**, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1,092 humans: application to cancer genomics. Science. 2013 Oct 4;342(6154):1235587.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrative analysis of cancer evolution

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Gad Getz, MGH / **Broad Institute**

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Scott L. Carter / Broad Institute	

Background and preliminary data

Starting from a normal cell, cancers evolve via multiple rounds of mutation, selection, and expansion. Continued application of this process to the growing cancer-cell population results in branched genetic variegation, whereby multiple cancer subclones relate to each other in a tree-like fashion. Consequently cancer tissues are substantially heterogeneous both across different anatomical regions and within single cancer biopsies.

Clinical events associated with poor prognosis such as relapse, metastasis, and acquired drug resistance can likely be understood in terms of cancer evolution. Recent studies have utilized massively parallel sequencing to characterize genetic evolution in various cancers. For example, our longitudinal analysis of chronic lymphocytic leukemia (CLL) using whole exome sequencing has revealed that the presence of subclonal driver mutations in CLL samples is associated with clonal evolution following treatment (whereby the subclonal drivers are enriched) and shorter duration until clinical relapse.

While these studies have illuminated DNA mutations driving cancer evolution, the interaction between such genetic evolution and the cancer transcriptome and epigenome has not been widely examined. Such integrative analyses will be useful to understand the functional consequence of driver DNA mutations, and to further understand the cellular contexts in which particular DNA mutations are selected for.

Timelines & resources dedicated to project

Timeline:

Evolutionary analysis of DNA alterations (ABSOLUTE, Phylogic), September 2014

Association between genetically-defined subclonal evolution and transcriptional and epigenetic programs, January 2015

Manuscript preparation / submission, May 2015

Resources:

- Variant calls for somatic DNA mutations, copy-number alterations, and rearrangements in 2,000 ICGC/TCGA samples, including 200 cancers sampled at multiple locations or time points.

- Read counts supporting alternate and reference alleles at variant sites (both germline and somatic) in each cancer sample.

- Transcript abundance data from RNA-seq or microarray platforms in each cancer sample.

- Quantitative estimates of net epigenetic alterations in cancer tissues (e.g. from whole-genome bisulfite sequencing, methylation array data).

- Whole exome sequencing (WES) and transcriptome / methylation profiling data from ~8,000 additional cancer samples (TCGA).

- High quality germline SNP genotype calls obtained by joint-calling in 2,000 WGS ICGC/TCGA samples.

- Integration of 8,000 above samples with intersecting reverse phase protein array (RPPA) data.

- Affymetrix 6.0 data for ~10,000 cancer samples (with complete overlap of the 8,000 WES samples) processed with HAPSEG and ABSOLUTE to yield absolute allelic copy-number profiles. (SLC currently has a working version of this dataset).

Research proposal

For all aims below, results from the analysis described will be made available to the TCGA/ICGC community. For aims 1-6, only minor modifications to existing software tools will be required. Aims 7-9 will require novel methodological development, which will likely be done in collaboration with other ICGC groups.

1) We will integrate germline genotype data with variant read counts to produce high-quality allelic copy-ratio profiles for the 2,000 TCGA/ICGC WGS samples.

2) We will estimate sample purity, ploidy, genome-wide absolute allelic copy-numbers, and subclonal structure using ABSOLUTE.

3) For all identified somatic copy-number alterations, point mutations, and structural variants, we will estimate the fraction of sampled cancer cell harboring the alteration (their *cancer cell fraction*, or **CCF**).

4) We will extend our models to produce CCF estimates for compound genomic alterations such as chromothripsis, chromoplexy, and chromanasynthesis. We will work with other groups (e.g. Imielinski, Wala) to model each event identified by these groups using WGS read-depth and paired-end data.

5) We will perform these analyses on cell-line datasets (e.g. the CCLE), for ease of integration with TCGA/ICGC samples.

6) For all families of related cancer samples (>1 cancer samples sequenced), we will infer the evolutionary relationships between all cancer subclones identified by jointly modeling the CCF estimates of all variants identified, using the bespoke software package "Phylogic" (SLC, unpublished work).

7) We will integrate the genetic evolution observed in each sample family with transcriptome and epigenome data to identify modules of coherent genetic, epigenetic, and transcriptional alterations in cancer cells.

8) We will investigate the incidence of these modules in ~8,000 TCGA samples with WES, transcriptome, and epigenome profiling data available. We will investigate correlations with clinical sample annotations, when available.

9) We will investigate the incidence of these modules in cell-line datasets, and attempt to associate them with data describing drug/shRNA sensitivity, metabolite levels, DNA methylation patterns, and signaling protein phosphorylation status.

Legacy plans

Modifications required to the ABSOLUTE algorithm to model the CCF of compound structural variants will be made available as a software package update, possibly accompanied by a short published description.

The software "Phylogic" will be made available for initial public release during the course of this project. This release would be strengthened by publication in one of the journals accepting methods papers from the ICGC effort.

We will make all analysis results of the ICGC/TCGA datasets available to the community, and work with groups developing web portals to release curated results.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

A. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997	Military Service - Captain
1997-1998	Tel Aviv. Univ. MSc student
1998-2000	Maximal Innovative Intelligence (part time)
1998-2003	Weizmann Institute of Science. PhD student
2004-2007	Broad Institute of MIT and Harvard. Postdoc
2007-2012	Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013-	Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University

1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

Curriculum vitae for Scott L. Carter, Ph.D.

Education and Training

2001 B.S., Computer Science and Mathematics, University of Maryland at College Park
2011 Ph.D., Bioinformatics and Integrative Genomics, MIT and Harvard Medical School (HST).
Thesis advisors: Matthew Meyerson and Gad Getz

Research and Professional Experience

2001-2003 *Bioinformatics software developer*, Xpogen, Inc., Cambridge, MA
2004-2006 *Research associate*, Boston Children's Hospital Informatics Program (CHIP), Boston, MA

Publications (8 selected of 48 original research publications)

1 David G. McFadden*, Thales Papagiannakopoulos*, Amaro Taylor-Weiner*, Chip Stewart*, **Scott L. Carter***, *et al.* Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* (*In press*).
2 Jens G. Lohr*, Petar Stojanov*, **Scott L. Carter***, *et al*. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell* (*In press*).
3 Landau, D. A.*, **Carter, S.L.***, *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714-726 (2013).
4 **Carter, S. L.** *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421 (2012).
5 **Carter, S. L.**, Meyerson, M. L. & Getz, G. Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping. *Nature precedings*, 59 (2011).
6 **Carter, S. L.**, Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature genetics* **38**, 1043-1048 (2006).
7 **Carter, S. L.**, Eklund, A. C., Mecham, B. H., Kohane, I. S. & Szallasi, Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC bioinformatics* **6**, 107 (2005).
8 **Carter, S. L.**, Brechbühler, C. M., Griffin, M. & Bond, A. T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**, 2242-2250 (2004).

* - equal contribution

Scientific Software

ABSOLUTE, Software for joint estimation of tumor purity, ploidy, and absolute copy-number inference in human cancer genomes.

HAPSEG, Software for segmentation of allelic copy-ratios across a cancer genome. This allows inference of purity/ploidy using ABSOLUTE. In addition, this software produces long-range haplotypes at heterozygous loci in regions of homologous copy-imbalance.

CAPSEG, Software for producing accurate copy profiles from whole exome sequencing data. This software can remove technical noise in FFPE samples by comparison to many normal samples from similar conditions. *In development for public release*

Allelic CAPSEG, Software for producing accurate allele-specific copy profiles from whole exome sequencing data. This software uses a statistical model for the read depth at heterozygous SNP sites in the exome to infer contribution of both homologous chromosomes to the copy ratio of each genomic segment. *In development for public release*

Phylogic, Software for automatically inferring the phylogenetic relationship between all subclones detected in a set of related cancer samples. *In development for public release.*

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Optimization and benchmarking of somatic mutation detection in whole genome sequencing data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gad Getz (Broad / MGH)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Kristian Cibulskis (Broad) Adam Kiezun (Broad)	

Background and preliminary data

Small (1-50bp) somatic mutations have long been known to play a significant role in tumorigenisis, and have been at the core of numerous next generation sequencing studies of cancer. However, the accurate and sensitive detection of these events is not a solved problem, in large part because both the raw data and the downstream analysis are rapidly evolving. The primary focus of many somatic variant detection methods has been detecting clonal point mutations in deep (>100x) coverage whole exome data. However, the production of vast amounts of lower coverage (~60x) whole genome sequence data is now underway and will require more sophisticated methods to address the lower complexity of non-coding sequence, differential coverage between tumor and normal samples and other novel challenges which will emerge from this unprecedented data set. Moreover, downstream analysis has evolved to include subclonal events and small indels, which are more challenging to detect and have received less focus to date.

The initial publication of our method MuTect (*Cibulskis et al.*, Nature Biotechnology 2013), attempted to address several of these issues. We developed a comprehensive benchmarking approach and showed that our method is more sensitive at a given specificity than other leading methods. We aimed to detect events present in only a small fraction (<5%) of the tumor DNA, enabling the study of subclonal events and highly impure tumor samples. However, we did not focus on short indel detection nor did we deeply address issues specific to non-coding sequence. Additionally, MuTect has now been applied to 1000s of tumors, including TCGA whole genomes and longitudinal studies, which have increased our understanding of rare artifacts and identified opportunities for improvement of both sensitivity and specificity.

Timelines & resources dedicated to project

Timeline: Method development and analysis in a subset of representative whole genomes by July 2014. Application of the method to entire ICGC/TCGA data set and calls published by Dec 2014. Manuscript preparation / submission in May 2015.

Resources: consistently aligned (harmonized) tumor and normal whole genome read data

Research proposal

Towards the goal of delivering the most accurate and sensitive mutation detection algorithm we propose to extend MuTect in several aspects, and deliver these mutation calls along with the method to the community. We anticipate that our methods will be of interest to several other groups submitting research proposals for the ICGC, such as point mutation and indel analysis in both coding and non-coding regions, and will work closely with those groups to deliver our analysis prioritized by those needs. However, we aim to improve the quality of mutation and indel calls in the following way:

1.) Extend MuTect in-silico mixing benchmark approach for indels and apply to leading methods. This will allow us to measure indel detection performance in the same rigorous manner as we have for point mutations, set a baseline for the state of the art, and be able to quantify our improvements.

2.) Extend MuTect to perform indel detection using a haplotype-based approach. Working closely with the developers of the GATK UnifiedGenotyper and HaplotypeCaller, we will recast the core detection strategy of MuTect to use a haplotype-based approach, which combined with the effective false positive reduction strategies in MuTect should yield a high performance method for both point mutations and indels.

3.) Using an extended panels-of-normals for eliminating false positives. In the MuTect paper, this was performed to great effect but in a static, non-extensible manner. We aim to extend this approach to take advantage of the large amount of data made available in the ICGC/TCGA Pan-Can analysis.

4.) Incorporate knowledge of tumor-specific previously detected events. Once a somatic event has been detected in an appropriate number of samples, an increased probability of the same site harboring a somatic event should incorporated into the core detection statistic to allow for more sensitive detection of known events.

5.) Focused measurement and improvement of method in non-coding regions. Explicitly focusing on non-coding regions of the genome, further subdivided by function we aim to specifically quantify and improve the state-of-the-art in somatic variant detection in whole genome data.

6.) Computational performance optimization. In order to analyze large amounts of whole genome data in a cost-effective way, we aim to improve the compute efficiency of our method, investigating both code optimization as well as GPU-based acceleration

Legacy plans

We will provide our mutation detection detection software, and results of that method on the ICGC/TCGA data set to the community.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

16. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
17. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
18. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
19. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
20. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
21. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
22. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
23. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
24. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal

somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
25. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
26. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
27. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
28. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
29. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
30. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

NAME Kristian Cibulskis	POSITION TITLE **Assistant Director, Informatics** **Cancer Genome Analysis** **The Broad Institute, Cambridge, MA**
CONTACT INFORMATION kcibul@broadinstitute.org	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Cornell University, Ithaca, NY	B.Sc	1992-1996	Computer Science

A. Personal Statement

My research interests lie in the understanding of cancer through the application of computational methods at large scale. My training and early career were focused on software engineering in high volume transaction processing systems in the financial industry. Later, I shifted that focus to the genome sequencing domain where I applied my engineering skills to computational problems in both capillary and then next generation sequencing methods development and large scale analysis. I am particularly interested in the goal of completely and accurately characterizing individual cancer genomes, and the potential impact that could have upon clinical medicine.

B. Positions and Honors

1996-2000 Software Architect, Sapient Corporation, Cambridge MA

2000-2001 Principal Architect, Vertica Systems, Medford MA

2001-2003 Technical Lead, Sun Microsystems, Marlborough, MA

2003-2004 Technical Lead, DeNovis, Lexington, MA

2004-2008 Senior Software Engineer, Broad Institute, Cambridge MA

2008-2013 Computational Biologist, Broad Institute, Cambridge MA

2013-Present Assistant Director, Broad Institute, Cambridge MA

Honors: Dean's List, John McMullen Dean's Scholar, Tau Beta Pi Honor Society

C. Recent relevant publications (Selected from 43 peer-reviewed publications, * denotes equal contributions)

1. **Cibulskis K**, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. **Nat Biotechnol**. 2013 Mar;31(3):213-9.
2. **Cibulskis K**, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross- contamination of human samples in next-generation sequencing data. **Bioinformatics**. 2011 Sep 15;27(18):2601-2
3. Banerji S*, **Cibulskis K***, Rangel-Escareno C*, Brown KK* et al., Sequence analysis of mutations and translocations across breast cancer subtypes. **Nature**. 2012 Jun 20;486(7403):405-9
4. Carter SL, **Cibulskis K**, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. **Nat Biotechnol**. 2012 May;30(5):413-21
5. M. Lawrence, P.Stojanov, P.Polak, et al, Mutational heterogeneity in cancer and the search for new cancer-associated genes, **Nature** 499, 2013

Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. ***CELL* 152,** 714–726 (2013).

NAME Adam Kiezun, PhD	POSITION TITLE **Group Leader** **Computational Methods Development** **Cancer Genome Analysis** **The Broad Institute, Cambridge, MA**
CONTACT INFORMATION akiezun@broadinstitute.org	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Warsaw University, Warsaw, Poland	B.Sc	1995-1998	Computer Science
Warsaw University, Warsaw, Poland	M.Sc	1999-2001	Computer Science
Massachusetts Institute of Technology, Cambridge	Ph.D.	2003-2009	Computer Science
Brigham and Women's Hospital / Harvard Medical School	Postdoc	2009-2011	Medical and Population Genetics, Computational Biology

A. Personal Statement

B. Positions and Honors

2000-2003 Software Engineer, IBM Zurich

2004, 2005, Summer Research Intern, IBM Research, Hawthorne, NY

2007 Summer Research Intern, Microsoft, Redmond, WA

2003-2009 PhD Student, Computer Science, MIT

2009-2011 Postdoctoral Fellow, Sunyaev Lab, Brigham and Women's Hospital / Harvard Medical School, Boston

2011-2013 Computational Biologist, Cancer Genome Analysis, Broad Institute of MIT and Harvard
2013-present Group Leader, Computational Methods Development, Broad Institute, Cambridge, MA

Honors: • IBM PhD scholarship in recognition of academic excellence, 2007 ($20'000 awarded individually without restrictions), IBM Research Division Awards in 2005 and 2007, IBM Research Innovation Award for patent application, 2005, ACM SIGSOFT Distinguished Paper Award, International Symposium on Software Testing and Analysis (ISSTA), 2009, Paper selected for expedited journal publication ISSTA 2008, ACM SIGSOFT Distinguished Paper Award – International Conference on Software Engineering (ICSE) 2007, Best Paper Selection – International Conference on Automated Software Analysis (ASE) 2007, ACM SIGSOFT Distinguished Paper Award – European Software Engineering Conference/Conference on Foundations of Software Engineering (ESEC/FSE) 2007, MSc Thesis awarded 3rd price in national annual contest for best Theses in Computer Science in Poland (awarded by the Polish Information Processing Society) 2001

C. Recent relevant publications (Selected from 32 peer-reviewed publications, * denotes equal contributions)

1. J. M. Francis*, **A. Kiezun*,** A. H. Ramos*, et al Somatic mutation of CDKN1B in small intestine neuroendocrine tumors. **Nature Genetics** 45: 1483–1486, 2013
2. **A. Kiezun**, et al. Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency, **PLoS Genetics**, 2013
3. M. Lawrence, P.Stojanov, P.Polak, et al, Mutational heterogeneity in cancer and the search for new cancer-associated genes, **Nature** 499, 2013
4. T. Pugh et al. The genetic landscape of high-risk neuroblastoma, **Nature Genetics**, 2013
5. **A. Kiezun** et al, Exome sequencing and the genetic basis of complex traits, **Nature Genetics**, 44:623-630, 2012
6. E Kim et al, Genome sequencing reveals insights into physiology and longevity of the naked mole rat, **Nature** 2012
7. N. Stitziel*, **A. Kiezun***, S. Sunyaev, Computational and statistical approaches to analyzing variants identified by exome sequencing, **Genome Biology**, 2011
   D. Jordan*, **A. Kiezun*,** S. Baxter* et al, Development and Validation of a Computational Method for Assessment of Missense Variants in Hypertrophic Cardiomyopathy, **American Journal of Human Genetics**, 2011

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Landscape of somatic indels and indel processes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gad Getz (Mass General Hospital, Broad Institute)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Adam Kiezun (Broad Institute) **Kristian Cibulskis (Broad Institute)**	

Background and preliminary data

The unprecedented amount of sequencing data available across multiple cancer types allows for the first time to systematically search both for phenomena common in cancer and for those contrasting different types of the disease. While the 'landscape' of somatic point mutations has been extensively studied and multiple determinates have been described (such as the tri-nucleotide context, increased mutation rates at CpG sites, increased rate of C>A mutations caused by tobacco smoke, or C>T mutations caused by ultraviolet exposure) the forming processes and causes are less understood for small insertions and deletions (indels). With the aim of such pan-cancer analysis of indels, we propose to use ICGC/TCGA data (WES and WGS) and study their distributions, across different cancer types, indel sizes, genomic position, types, sequence contexts, and locations ("hot spots").

Timelines & resources dedicated to project

Timeline:

By September 2014: indel callset for all WGS and WES data in ICGC/TCGA.
By end of December 2014: analysis
By May 2015: manuscript preparation

Research proposal

We propose the following plan for identification and analysis of somatic short indels:

1) **Identify and set up a set of somatic indel calling tools.** We will identify and attempt to set up multiple somatic indel callers. We will make scripts to convert the each caller's output to common format (VCF). One of the callers will be the improved version of Broad's caller MuTect (see abstract by Cibulskis and Kiezun) which will be continually improved in the course of this analysis.
2) **Build a knowledge database (KDB) to quantify caller performance.** To quantify callers' performance it is necessary to establish 'the ground truth'. To that end, we will create a collection of datasets and curated callsets for benchmarking cancer data. For this, we will use established pairs of tumor/normal cell line pairs (eg HCC1143, HCC1954, COLO829) and curate high-confidence callsets by applying multiple tools to replicates of the data and manually reviewing and/or lab-validating detected indels. This is similar and related to NIST Genome in a Bottle for cancer.
3) **Quantifying callers' performance.** We will use real data (from KDB cell lines) and synthetic data (from mixing experiments, either in vitro or in silico) to quantify each caller's sensitivity and false positive rate**.** We will select a core set of best-performing callers and, if necessary, establish their best combination, i.e., make an ensemble caller that combines strengths of multiple callers.
4) **Panel of Normals.** Two sources of false positives, germline variants and regional noise, can be diminished by using a Panel of Normals – a database of variation detected in a set of normal data. We will create a Panel of Normals for WES and WGS data by first identifying a set of normal data not contaminated with tumor cells and then collecting counts of insertions and deletions at each position across samples.
5) **Create callset.** We will use the core set of best performing callers to identify somatic indels in all ICGC/TCGA exome and WGS data**.**
6) **Annotate and Analyze.** We will richly annotate the detected indels as to the transcript, sequence context, etc. We will then examine the indel distributions, across different cancer types, indel sizes, genomic regions, types, sequence contexts, and locations ("hot spots").

-

Legacy plans

The following will be made available:
Knowledge Database – curated callsets for benchmark cancer data
Indel callset for WGS and WES
Panel of Normals for WES and WGS
Software codes

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces

1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

NAME Adam Kiezun, PhD	POSITION TITLE **Group Leader** **Computational Methods Development** **Cancer Genome Analysis** **The Broad Institute, Cambridge, MA**
CONTACT INFORMATION akiezun@broadinstitute.org	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Warsaw University, Warsaw, Poland	B.Sc	1995-1998	Computer Science
Warsaw University, Warsaw, Poland	M.Sc	1999-2001	Computer Science
Massachusetts Institute of Technology, Cambridge	Ph.D.	2003-2009	Computer Science
Brigham and Women's Hospital / Harvard Medical School	Postdoc	2009-2011	Medical and Population Genetics, Computational Biology

A. Personal Statement

B. Positions and Honors

2000-2003 Software Engineer, IBM Zurich

2004, 2005, Summer Research Intern, IBM Research, Hawthorne, NY

2007 Summer Research Intern, Microsoft, Redmond, WA

2003-2009 PhD Student, Computer Science, MIT

2009-2011 Postdoctoral Fellow, Sunyaev Lab, Brigham and Women's Hospital / Harvard Medical School, Boston

2011-2013 Computational Biologist, Cancer Genome Analysis, Broad Institute of MIT and Harvard
2013-present Group Leader, Computational Methods Development, Broad Institute, Cambridge, MA

Honors: • IBM PhD scholarship in recognition of academic excellence, 2007 ($20'000 awarded individually without restrictions), IBM Research Division Awards in 2005 and 2007, IBM Research Innovation Award for patent application, 2005, ACM SIGSOFT Distinguished Paper Award, International Symposium on Software Testing and Analysis (ISSTA), 2009, Paper selected for expedited journal publication ISSTA 2008, ACM SIGSOFT Distinguished Paper Award – International Conference on Software Engineering (ICSE) 2007, Best Paper Selection – International Conference on Automated Software Analysis (ASE) 2007, ACM SIGSOFT Distinguished Paper Award – European Software Engineering Conference/Conference on Foundations of Software Engineering (ESEC/FSE) 2007, MSc Thesis awarded 3rd price in national annual contest for best Theses in Computer Science in Poland (awarded by the Polish Information Processing Society) 2001

C. Recent relevant publications (Selected from 32 peer-reviewed publications, * denotes equal contributions)

9. J. M. Francis*, **A. Kiezun*,** A. H. Ramos*, et al Somatic mutation of CDKN1B in small intestine neuroendocrine tumors. **Nature Genetics** 45: 1483–1486, 2013
10. **A. Kiezun**, et al. Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency, **PLoS Genetics**, 2013
11. M. Lawrence, P.Stojanov, P.Polak, et al, Mutational heterogeneity in cancer and the search for new cancer-associated genes, **Nature** 499, 2013
12. T. Pugh et al. The genetic landscape of high-risk neuroblastoma, **Nature Genetics**, 2013
13. **A. Kiezun** et al, Exome sequencing and the genetic basis of complex traits, **Nature Genetics**, 44:623-630, 2012
14. E Kim et al, Genome sequencing reveals insights into physiology and longevity of the naked mole rat, **Nature** 2012
15. N. Stitziel*, **A. Kiezun***, S. Sunyaev, Computational and statistical approaches to analyzing variants identified by exome sequencing, **Genome Biology**, 2011

D. Jordan*, **A. Kiezun*,** S. Baxter* et al, Development and Validation of a Computational Method for Assessment of Missense Variants in Hypertrophic Cardiomyopathy, **American Journal of Human Genetics**, 2011

NAME Kristian Cibulskis	POSITION TITLE **Assistant Director, Informatics** **Cancer Genome Analysis** **The Broad Institute, Cambridge, MA**
CONTACT INFORMATION kcibul@broadinstitute.org	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Cornell University, Ithaca, NY	B.Sc	1992-1996	Computer Science

A. Personal Statement

My research interests lie in the understanding of cancer through the application of computational methods at large scale. My training and early career were focused on software engineering in high volume transaction processing systems in the financial industry. Later, I shifted that focus to the genome sequencing domain where I applied my engineering skills to computational problems in both capillary and then next generation sequencing methods development and large scale analysis. I am particularly interested in the goal of completely and accurately characterizing individual cancer genomes, and the potential impact that could have upon clinical medicine.

B. Positions and Honors

1996-2000 Software Architect, Sapient Corporation, Cambridge MA

2000-2001 Principal Architect, Vertica Systems, Medford MA

2001-2003 Technical Lead, Sun Microsystems, Marlborough, MA

2003-2004 Technical Lead, DeNovis, Lexington, MA

2004-2008 Senior Software Engineer, Broad Institute, Cambridge MA

2008-2013 Computational Biologist, Broad Institute, Cambridge MA

2013-Present Assistant Director, Broad Institute, Cambridge MA

Honors: Dean's List, John McMullen Dean's Scholar, Tau Beta Pi Honor Society

C. Recent relevant publications (Selected from 43 peer-reviewed publications, * denotes equal contributions)

6. **Cibulskis K**, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. **Nat Biotechnol**. 2013 Mar;31(3):213-9.
7. **Cibulskis K**, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross- contamination of human samples in next-generation sequencing data. **Bioinformatics**. 2011 Sep 15;27(18):2601-2
8. Banerji S*, **Cibulskis K***, Rangel-Escareno C*, Brown KK* et al., Sequence analysis of mutations and translocations across breast cancer subtypes. **Nature**. 2012 Jun 20;486(7403):405-9
9. Carter SL, **Cibulskis K**, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. **Nat Biotechnol**. 2012 May;30(5):413-21
10. M. Lawrence, P.Stojanov, P.Polak, et al, Mutational heterogeneity in cancer and the search for new cancer-associated genes, **Nature** 499, 2013

Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. ***CELL* 152,** 714–726 (2013).

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrative analysis of germline and somatic alterations

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gad Getz (Mass General Hospital, Broad Institute of MIT and Harvard)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Adam Kiezun, Broad Institute of MIT and Harvard Paz Polak, Mass General Hospital	

Background and preliminary data

The molecular basis of cancer is very complex with multiple contributing factors, including inherited genetic variants and acquired somatic alterations. Although the interaction of germline and somatic factors have long been appreciated, it has not been fully elucidated, and many recent cancer studies fall strictly into one of two categories germline OR somatic. Examples of such interactions include increased predisposition to certain somatic mutations (EGRF in lung cancers in East Asians, JAK2 V617F in germline carriers of specific JAK2 haplotypes) and 'two-hit' scenarios in which the first hit is an inherited variant and the second hit is a somatic alteration such as mutation, copy number variant (CNV), loss of heterozygosity (LOH), or epigenetic silencing. We propose to systematically search for and analyze such interactions, analyze germline contribution to somatic mutational patterns and processes, and to integrate germline variation into analysis for cancer driver genes.

In preliminary investigations, we examined 2-hit interactions of rare germline loss of function (LOF) variants and somatic LOH with the goal of targeting tumor suppressor genes. The analysis focused on rare germline LOF variants in genes that are subject to somatic second hits and uses a statistical method to find genes that show a significant enrichment of the 2-hit signature. Our preliminary results of around 680 breast cancer exome samples from TCGA show that this approach is effective - it readily identifies known tumor suppressor genes in breast cancer: BRCA1 and BRCA2 as top genes with genome-wide significance with a well-calibrated test even though no control data was used for germline analysis and few somatic mutations were present in those two genes.

Timelines & resources dedicated to project

Timeline:

- By August 2014: germline callset for ICGC/TCGA exomes and selected regions from WGS
- By October 2014: Identified copy number segmentation for tumor samples and annotated somatic LOH status for germline variants. Annotated LOH status of germline variants in WGS samples (using annotated segmented copy number data from the integrative copy-number and rearrangement analysis by Imielinski et al)
- By End of December 2014: Integrated germline and somatic analysis
- By May 2015: Manuscript preparation

Research proposal

We will perform an integrative germline and somatic analysis which will have the following subtasks:

1) **Identification of germline variants in a joint calling of all ICGC/TCGA exomes (and selected regions of WGS samples).** High-quality detection and genotyping of germline variants is significantly aided by analyzing all samples jointly. We will create a germline callset using Broad's best practices, which uses a local-assembly based method (HaplotypeCaller) and variant quality score recalibration (VQSR). Given our experience with several germline projects of large scale and complexity, we expect a high degree of heterogeneity in germline samples, therefore very stringent quality control (QC) criteria will be used. Stringent QC is particularly important when analyzing rare LOF because those variants are enriched for artifacts. Main ethnic clusters will be identified to stratify or subset samples for more powerful analysis. Germline variants will be annotated for their impact (gene, variant type, prediction algorithms e.g., PolyPhen). We will also jointly call and process variants in selected (all coding and selected non-coding) regions of WGS samples
2) **Identification of SCNAs and regions of LOH.** To detect SCNAs in tumor exome data, we will apply a method that uses tumor copy ratio data normalized by normal variation inferred from many normal samples (tangent normalization), followed by copy number segmentation. Next, we will integrate germline variation and SCNAs to infer homolog-specific absolute copy numbers for each segment using ABSOLUTE (Carter et al, 2012), regions of LOH and (by correcting for impurity – fraction of normal cells in tumor sample) the lost allele for all germline heterozygous sites in regions of LOH. For WGS data, we will use segmented copy-number data produced by the integrated copy number and rearrangement method (see abstract of Imielinski et al.)
3) **Identification of monoallelic expression of germline variants.** In some cases, even though no genetic or epigenetic alteration may be detectable, the wild-type allele may not be transcribed. To detect such cases, we will use RNASeq data whenever available and infer monoallelic expression from RNAseq reads after correcting for impurity.
4) **2-hit analyses.** We will integrate germline and somatic data in a statistical test designed to detect unusually recurrent but otherwise highly improbable combinations of germline variants and somatic mutations, somatic epigenetic silencing, allele-specific somatic LOH events that lead to preferential loss of the wild-type allele, or copy number changes that preferentially amplify germline variants.
5) **Mapping germline determinants of features of somatic mutational processes.** We will identify and quantify spectra and patterns of somatic mutations and use quantitative trait analysis for mapping them using germline variants. Example features of interest include: density of mutations in DNase hypersensitive sites (DHS), strand asymmetry in mutations associated to transcription coupled repair, mutational spectrum (the relative abundance of particular single nucleotide substitution) spatial clustering of mutations (such as those associated with APOBEC activity), total mutation rate, mutation rates at early and late-replicating regions, known hotspots of somatic mutations. We will apply quantitative genetics methods to identify genetic alterations associated with quantitative characteristics of somatic mutations. Of particular interest are the extremes of the trait distributions. The analysis will account for tumor types and somatic alterations that may confound the germline associations. We will also examine candidate genes in DNA repair pathway, chromatin remodeling pathway or genes involved in familial cancer syndromes. In WGS and WES we will search for patients with such variants either germline or somatic and investigate if there is a difference in the mutational spectrum than patients who don't carry mutations in these genes. Moreover, a pan-cancer scan can be helpful in identifying potential tissue specific shifts in mutation rates for example is the impact of POLE or MSH6 is the same in all cancers or the signature vary among tissues.
6) **Search for germline/somatic association in cis.** We will analyze association of germline variants (gene-based or single-variant) with somatic variants in the same gene aiming to find germline variants predisposing to specific mutations. We will focus on cis associations to limit multiple testing.
7) **Integrated somatic and germline analysis of mutation significance.** We will continue work on a method that jointly analyses germline variants and somatic mutations across many individuals to detect cancer genes. This effort also includes integrating data from thousands of non-cancer controls for which we have approval to use as controls in cancer studies.

Legacy plans

1. Germline callset will be made available
2. Summary (allele count) germline data from ICGC/TCGA will be made available publicly (subject to approval)
3. LOH annotations, including inferred lost allele, for germline heterozygous variants will be made available
4. Software codes developed in the course of the project will be made available

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University

2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

NAME
Adam Kiezun, PhD

CONTACT INFORMATION

akiezun@broadinstitute.org

POSITION TITLE
Group Leader
Computational Methods Development
Cancer Genome Analysis
The Broad Institute, Cambridge, MA

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Warsaw University, Warsaw, Poland	B.Sc	1995-1998	Computer Science
Warsaw University, Warsaw, Poland	M.Sc	1999-2001	Computer Science
Massachusetts Institute of Technology, Cambridge	Ph.D.	2003-2009	Computer Science
Brigham and Women's Hospital / Harvard Medical School	Postdoc	2009-2011	Medical and Population Genetics, Computational Biology

A. Personal Statement

My research interests lie in the understanding of cancer through the application of computational methods at large scale. My training and early career were focused on software engineering in high volume transaction processing systems in the financial industry. Later, I shifted that focus to the genome sequencing domain where I applied my engineering skills to computational problems in both capillary and then next generation sequencing methods development and large scale analysis. I am particularly interested in the goal of completely and accurately characterizing individual cancer genomes, and the potential impact that could have upon clinical medicine.

B. Positions and Honors

1996-2000 Software Architect, Sapient Corporation, Cambridge MA

2000-2001 Principal Architect, Vertica Systems, Medford MA

2001-2003 Technical Lead, Sun Microsystems, Marlborough, MA

2003-2004 Technical Lead, DeNovis, Lexington, MA

2004-2008 Senior Software Engineer, Broad Institute, Cambridge MA

2008-2013 Computational Biologist, Broad Institute, Cambridge MA

2013-Present Assistant Director, Broad Institute, Cambridge MA

Honors: Dean's List, John McMullen Dean's Scholar, Tau Beta Pi Honor Society

C. Recent relevant publications (Selected from 43 peer-reviewed publications, * denotes equal contributions)

1. **Cibulskis K**, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. **Nat Biotechnol**. 2013 Mar;31(3):213-9.
2. **Cibulskis K**, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross- contamination of human samples in next-generation sequencing data. **Bioinformatics**. 2011 Sep 15;27(18):2601-2
3. Banerji S*, **Cibulskis K***, Rangel-Escareno C*, Brown KK* et al., Sequence analysis of mutations and translocations across breast cancer subtypes. **Nature**. 2012 Jun 20;486(7403):405-9
4. Carter SL, **Cibulskis K**, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. **Nat Biotechnol**. 2012 May;30(5):413-21
5. M. Lawrence, P.Stojanov, P.Polak, et al, Mutational heterogeneity in cancer and the search for new cancer-associated genes, **Nature** 499, 2013

Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. ***CELL*** **152,** 714–726 (2013)

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Paz Polak	
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency trai if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Technion, Israel	B.Sc.	2000	Physics and Mathematics
Technion, Israel	M.Sc	2003	Applied Mathematics
Weizmann Institute of Science, Israel	M.Sc.	2006	Physics (computational biolc
Max Planck Institute For molecular Genetics/ Free university Berlin, Germany	Ph.D	2010	Computational Biology

B. Personal Statement

C. Positions and Honors

Positions:

2006-2011 Max Planck Institute for Molecular Genetics, Berlin, Germany. PhD student
2011- Brigham and Women's Hospital and Harvard Medical School. Postdoc

Honors:

1996-1999 Dean's excellence list B.Sc. Technion, Haifa, Israel
2000 President's excellence list. B.Sc. Technion, Haifa, Israel
2006 IMPRS-CBSC PhD fellowship by Max Planck Institute

D. Selected Peer-reviewed Publications

1. Polak P and Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. 2006. BMC Genomics 7 p. 133
2. Polak P and Arndt PF. Transcription induces strand-specific mutations at the 5' end of human genes,Genome Research. 2008. 18, 1216-1223.
3. Polak P and Arndt PF. Long Range bi-directional strand asymmetries originate at CpG islands in the human genome. Genome Biology and Evolution 2009, 189
4. Polak P, Querfurth R, Arndt PF. The evolution of transcription-associated biases of mutations across vertebrates. 2010. BMC Evolutionary Biology 10
5. Lawrence MS*, Stojanov P*, **Polak P*,** …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process
6. **Polak P***, Lawrence M.S.*, Haugen E, Stoletzki N, Stojanov P., Thurman R.E., Garraway L.A., Mirkin S., Getz G., John A Stamatoyannopoulos J.A., Sunyaev S. Reduced relative mutation density in regulatory DNA of cancer genomes linked to DNA repair. Nature Biotechnology (accepted)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identify causal pathways associated with specific cancer subtypes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Dr. Gad Getz

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yingchun Spring Liu	

Background and preliminary data

Cancer is a genetics disease with various genetic disorders. A consequence of the various genetic disorders in cancer genome is the deregulation of signaling pathways that regulate cell proliferation and apoptosis. Often pathways are identified by the enrichment of the genes with disorders in the pathways. However, the mutations may not be driver mutations or be functional, or the copy number alterations may not be correlated with gene expression. A pathway can be activated by a single mutation or phosphorylation. These scenarios make enrichment based pathway identification difficult. In addition enrichment based pathway analysis is likely to identify the downstream affected pathways rather than the causal pathways. I will introduce a strategy that considers the expression signatures of a molecular subtype as the target genes of pathways that are regulated by a set of transcription factors (TFs) involved in the pathways. When a pathway is activated, the expression level of the genes regulated by the pathway will always be affected regardless of which genetic disorders.

This strategy was able to identify the E2F, MYC, and RAS pathways from their mRNA expression signatures obtained by over-expressing these genes, respectively. It also identified a set of genes associated with the JNK pathway that show distinct expression patterns in the TCGA ovarian proliferative subtype.

Timelines & resources dedicated to project

This method depends on the classification of molecular cancer subtypes by the pan-can analysis groups, which can be based on mutation, copy number alteration, or mRNA expression. For known subtypes of a specific cancer type,

- **Identify expression signatures of each subtype by March 2014.**
- **Identify the causal pathways associated with each signature by August 2014.**

With this method one can potentially identify the causal pathways rather than the observed downstream effect of the perturbation of causal pathways for a particular pan-can subtype.

Research proposal

In this method, each pathway is characterized by a set of transcription factors (TFs) that regulate the pathway. The TFs of each pathway are extracted from public databases.

The target genes of a TF are preliminarily defined by whether the TF binds to these genes. This information can be retrieved from the Encode project. Up to date there are binding sites data for over 180 TFs that mapped to 158 Biocarta pathways. For a particular cancer subtype, the target genes of a TF are further refined by whether their expression is correlated or anti-correlated with the expression of the TF.

The expression signature of a molecular cancer subtype is identified by proper statistical tests or by selecting markers genes associated with this cancer subtype. Then apply the method to these expression signatures to identify the causal pathways whose deregulation might have caused the observed molecular patterns.

Finally identify pathway molecular components that are associated with clinical features or drug candidates.

This method can be extended to identify causal pathways for each individual sample as well.

Legacy plans

This method will provide an R software application for identifying causal pathways related to cancer molecular subtypes or individual tumor samples.

YINGCHUN (SPRING) LIU, PH.D.

WORK EXPERIENCE

11/2011-Present: Broad Institute, Computational Biologist, Cambridge, USA
Cancer Programme, Principal Investigator: Dr. Gad Getz

- Analyzed RNA, whole exome and genome sequencing data to identify gene expression signatures, fusion genes, and driver mutations in human cancers (TCGA projects http://www.genome.gov/17516564) and mouse models.
- Identified genetic disorders correlated with clinical variables of interest.

7/2009-11/2011: Dana-Farber Cancer Institute / Harvard Medical School, Research Scientist, Boston, USA
Belfer Insitute, Principle Investigator: Dr. Lynda Chin

- Analyzed mRNA expression, copy number, miRNA expression, DNA methylation, and mutation data (both array and NGS sequencing) for identifying novel oncogenes.
  Identified oncogenic pathway signatures that are correlated with clinical variables.

12/2007-6/2009: Dana-Farber Cancer Institute / Harvard School of Public Health, Research Fellow, Boston, USA
Department of Biostatistics & Computational Biology, Principle Investigator: Guo-Cheng Yuan

- Discovered novel roles of JMJ and EZH in H3K27me3 modification that were published in *Cell*.
- Invited to write a book chapter on signaling pathways and cancer for the book entitled "Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications "
- Invited speaker at both World Cancer Congress 2008 in Shanghai, China, as well as the ORFeome conference, Harvard Medical School, 2007.

10/2002-11/2007: Lund University, Department of Theoretical Physics, PhD Candidate/Research Associate, Lund, Sweden
Department of Theoretical Physics, Dr. Markus Ringnér and Dr. Carsten Peterson

- Developed a method for identifying the 2 deregulated pathways across all tumor types. Method adopted throughout the dept.
- Developed Java software implementing a linear mixed model for removing protein-specific dye effects in DIGE data and identifying differentially expressed proteins. http://bioinfo.thep.lu.se/digeanalyzer.html
- Identified 4 previously unknown biological groups in array data through developing own Perl software tools. http://cbbp.thep.lu.se/~markus/software/classdiscoverer/

EDUCATION

- Lund University, Theoretical Physics, PhD, Computational Biology, Lund, Sweden, 2007.
- Chalmers University of Technology, MS, Bioinformatics, Gothenburg, Sweden, 2002.
- Tongji University, BS, Applied Chemistry, Shanghai, China, 1997.

TECHNICAL SKILLS

- Languages: Chinese (Native), English (Fluent), Swedish (Basic).
- Computer skills: R & Perl (Experienced); Java, Python & Matlab; Unix, SVN, & Windows environments (Proficient).
- Publications on accompanying page.

PUBLICATIONS

- The Cancer Genome Atlas Research Network (**Liu Y** included). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013, 499: 43-49.
- The Cancer Genome Atlas Research Network (**Liu Y** included). Integrated genomic analyses of ovarian carcinoma. *Nature* 2011, 474: 609–615.
- Ying H, Elpek K, Vinjamoori A, M.Zimmerman S, Chu G, Yan H, Fletcher-Sananikone E, Wang W, Zhang H, **Liu Y**, Zheng H, Kimmelman A, Paik J, Lim C, Perry S, Jiang S, Ivanova E, Protopopov A, Colla S, Xiao Y, Hezel A, Bardeesy N, Turley S, Thayer S, Wang Y, Chin L, and DePinho R. Pten is a major tumor suppressor in pancreatic ductal adenocarcinoma and regulates an NF-kB-cytokine network. *Cancer Discovery* 2011, 1:158-169.
- **Liu Y**. "Cancer and signaling pathway deregulation", Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications*, IGI Global, 2011,* 369-379.
- Gan B, Hu J, Jiang S, **Liu Y**, Sahin E, Wang Y, Chin L, **and DePinho R.** Lkb1 regulates quiescence and metabolic homeostasis of haematopoietic stem cells. *Nature 2010*, 468, 7324:701-4.
- **Liu Y**, Shao Z, and Yuan GC. Prediction of polycomb targets in mouse embryonic stem cells, *Genomics* 2010*,* 96, 1:17-26.
- Shen X, Kim W, Fujiwara Y, Simon MD, **Liu Y**, Mysliwiec MR, Yuan GC, Lee Y, and Orkin SH. Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell 2009*, 139, 7:1303-14.
- Shen X, **Liu Y**, Hsu J, Fujiwara Y, Kim J, Mao X, Yuan GC, and Orkin SH. EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining the identity and pluripotency of mouse embryonic stem cells, *Molecular Cell* 2008*,* 32: 491-502*.*

- **Liu Y** and Ringnér M. Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis, *Genome Biology* 2007, 8, R77.
- Krogh M, **Liu Y**, Bengtsson S, Valastro B and James P. Analysis of DIGE data using a linear mixed model allowing for protein-specific dye effects, *PROTEOMICS* 2007, 7(23): 4235-4244.
- Karlsson G, **Liu Y**, Larsson J, Goumans M-J, Lee J-S, Thorgeirsson S.S, Ringnér M and Karlsson S. Gene expression profiling demonstrates that TGF-beta signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-beta signaling, *Physiol. Genomics,* 21: 396-403, 2005.

Liu Y and Ringnér M. Multiclass discovery in array data, *BMC Bioinformatics* 2004*,* 5:70.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Predicting the tissue-of-origin of cancer using regional profiles of mutation rates and its implication for treatments

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Gaddy Getz (Broad Institute and MGH)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Paz Polak (MGH) **Mi**chael Lawrence (Broad Institute)	Rosa Karlic (Zagreb University)

Background and preliminary data

Mutational rates vary along cancer genomes, forming a regional mutational signature specific for the tumor. Regional mutation patterns along the genomes have been shown to be correlated with replication timing (Lawrence et al, 2013) and used to improve background models for occurrence analysis. Schuster-Bockler & Lehner (2012) showed that profiles of histone marks from blood CD34+ T-cells could explain up to 55% of mutational variation in a few non-blood cancers. Recently, the Epigenome Roadmap project produced genome-wide maps of histone marks in more than 100 normal tissues, providing an unprecedented representation of the epigenomes of normal human tissues. We, in a recent unpublished work, used this epigenome atlas to show that 73-88% of regional variation (R^2) in the genomes of colon cancer, liver cancer, multiple myeloma, melanoma and lung adenocarcinoma could be explained by the new epigenome data. Remarkably, the epigenetic features which were most informative of mutation patterns came from tissues which were the closest to the (normal) cell of origin. These findings helped us to determine the tissue of origin in about 121 cancer genomes in our recent study.

Timelines & resources dedicated to project

- September, 2014: collecting WGS and epigenomic data from different tissues that we will use for the genome wide analysis. Developing predictive framework for predicting mutation patterns along cancer genomes. Writing an R package and a web application (python) for predicting mutational patterns and identifying the cell of origin of cancer.
- End of December, 2014: Finalizing the analysis of epigenetic data and somatic analysis
- February, 2015: Manuscript preparation

Research proposal

Using 2000 genomes spanning 25 different types of cancers we suggest to investigate how regional variation in these cancers can be predicted from epigenomic data. In particular, we suggest exploring tissue specificity i.e. whether regional variation in the cancer can be explained by epigenomic data measured in the tissue which is most closely related to the cell-of-origin. We plan to use data from the Roadmap Epigenomics project and any other epigenomes that will be available (with emphasis on extending our epithelial epigenomes representation). The results will be used to develop a method of identifying the cell of origin of individual cancers, which could be particularly useful in improving diagnostics and treatment decisions in patients with cancer of unknown primary origin.

In order to determine the cell of origin of individual cancer genomes, we will develop methodologies based on relating epigenomic features of normal tissues to mutational patterns of different cancer types. Briefly, we will use epigenomic data measured in different normal tissues to predict mutational patterns along cancer genomes. Furthermore, we will assess the contribution of each epigenomic feature to the prediction accuracy and calculate enrichment of different tissues among the top-ranked features to infer the most likely tissue of origin of each individual cancer. We already tested this simple method on data from 121 individual cancer genomes, belonging to four different types of cancer, and were able to correctly determine the cell of origin for 89% of the analyzed genomes.

Inclusion of a large number of cancer types and individual cancer genomes will enable us to optimize the simple methodology described above. The availability of ~2000 cancer genomes will let us compare and integrate our method to other methods of identifying cell of origin based on sequence (using mutational signatures e.g. context dependency, mutational spectrum or list of genes that are mutated in the cancer). Furthermore, the availability of ~1500 cancer transcriptomes and ~200 cancer methylome data sets will allow us to compare our method to other types of established methods used to identify cell of origin of cancer, namely those based on the comparison of mRNA expression in cancer samples and normal tissues, as well as those based on comparing the methylation status of CpG dinucleotides in normal cells and cancer cells. We will establish whether our method, based on DNA sequence alone, gives comparable results to methods using information beyond the DNA sequence.

Moreover, the ~200 cancer methylome data sets can be compared to methylomes of normal cells. If these two data sets are significantly different, we plan to test whether methylation data from normal cells or cancer cells are more predictive of the mutation patterns of different cancers. In this way we would be able to determine which stage of cancer development mutational patterns and methylation patterns reflect.

A second approach of predicting the cell of origin will be by clustering cancer genomes according to their regional mutational profiles, similar to the method employed by Lawrence at al. (2013) and Alexandrov et al (2013) for mutational spectra in exomes. Classifying cancers by their regional variation can be helpful in revealing the genetic component involved in shaping these profiles. Although individual cancers of different cancer types often have similar mutational profiles, some of them can be expected to be outliers, not following the same distribution of mutations along the genome. It would be interesting to check whether the mutational profiles of outliers are associated with either somatic or germ-line mutations in DNA repair pathways or histone remodeling pathways, which can then be linked to treatments based on DNA repair inhibitors such as drugs that inhibit PARP1.

Legacy plans

The scripts that will be used in this study will be part of a python library that will added to gitHub. We will develop web applications that will provide a user-friendly interface for clinicians/pathologists to upload MAF files and retrieve the epigenomic profiles, the suggested tissue of the origin and possibly links to clinical trials.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science

2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Paz Polak	
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency trai if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Technion, Israel	B.Sc.	2000	Physics and Mathematics
Technion, Israel	M.Sc	2003	Applied Mathematics
Weizmann Institute of Science, Israel	M.Sc.	2006	Physics (computational biolc
Max Planck Institute For molecular Genetics/ Free university Berlin, Germany	Ph.D	2010	Computational Biology

B. Personal Statement

C. Positions and Honors

Positions:

2006-2011 Max Planck Institute for Molecular Genetics, Berlin, Germany. PhD student
2011- Brigham and Women's Hospital and Harvard Medical School. Postdoc

Honors:

1996-1999 Dean's excellence list B.Sc. Technion, Haifa, Israel
2000 President's excellence list. B.Sc. Technion, Haifa, Israel
2006 IMPRS-CBSC PhD fellowship by Max Planck Institute

D. Selected Peer-reviewed Publications

1. Polak P and Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. 2006. BMC Genomics 7 p. 133
2. Polak P and Arndt PF. Transcription induces strand-specific mutations at the 5' end of human genes,Genome Research. 2008. 18, 1216-1223.
3. Polak P and Arndt PF. Long Range bi-directional strand asymmetries originate at CpG islands in the human genome. Genome Biology and Evolution 2009, 189
4. Polak P, Querfurth R, Arndt PF. The evolution of transcription-associated biases of mutations across vertebrates. 2010. BMC Evolutionary Biology 10
5. Lawrence MS*, Stojanov P*, **Polak P*,** …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process
6. **Polak P***, Lawrence M.S.*, Haugen E, Stoletzki N, Stojanov P., Thurman R.E., Garraway L.A., Mirkin S., Getz G., John A Stamatoyannopoulos J.A., Sunyaev S. Reduced relative mutation density in regulatory DNA of cancer genomes linked to DNA repair. Nature Biotechnology (accepted)

Michael S. Lawrence, Ph.D.

Computational Biologist
Broad Institute of Harvard and MIT
7 Cambridge Center
Cambridge, MA 02142
(617) 875-0420
lawrence@broadinstitute.org

EDUCATION

Ph.D., Biology, Massachusetts Institute of Technology, 2005
"RNA Polymerase Ribozymes"
Prof. David P. Bartel, advisor

B.A., Biochemistry; Linguistics and Cognitive Science, Brandeis University, 1998
Summa Cum Laude
Departmental High Honors awarded for thesis:
"Deletion Analysis of a Prokaryotic Potassium Channel"
Prof. Christopher Miller, advisor

EMPLOYMENT

Harvard University Department of Chemistry
postdoctoral training, laboratory of Prof. David R. Liu. 2005 – 2008.

Broad Institute of Harvard and MIT
Computational Biologist, Cancer Program 2008 – 2014.

SELECTED PUBLICATIONS

Lawrence MS, Stojanov P, Mermel C, et al. (2014), "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, doi: 10.1038/nature12912.

Lawrence MS, Stojanov P, Polak P, et al. (2013) "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature* **499**, 214-8.

Polak P, Lawrence MS, et al. (2013), "Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair," *Nature Biotech*., doi: 10.1038/nbt.2778.

Roberts SA, Lawrence MS, et al. (2013) "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers," *Nature Genet*. **45**, 970-6.

Drier Y, Lawrence MS, et al. (2012) "Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability," *Genome Res*. **23**, 228-35.

Wang L, Lawrence MS, et al. (2011) “SF3B1 and other novel cancer genes in chronic lymphoctic leukemia,” *N*. *Engl*. *J*. *Med*. **365**, 2497-506.

Bass AJ, Lawrence MS, et al. (2011) “Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion,” *Nature Genet*. **43**, 964-8.

Berger MF, Lawrence MS, et al. (2011) “The genomic complexity of primary human prostate cancer,” *Nature* **470**, 214-20.

Chapman MA, Lawrence MS, et al. (2011) “Intial genome sequencing an analysis of multipe myeloma,” *Nature* **471**, 467-72.

Liu JM, Livny J, Lawrence MS, Kimball MD, Waldor MK, Camilli A (2009) “Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion, and parallel sequencing,” *Nucleic Acids Res*. **37**, e46.

Lawrence MS, Phillips KJ, Liu DR (2007) "Supercharging proteins can impart unusual resilience," *J*. *Am*. *Chem*. *Soc*. **129**, 10110-2.

Lawrence MS, Bartel DP (2005) "New ligase-derived RNA polymerase ribozymes," *RNA* **11**, 1173-80.

Lawrence MS, Bartel DP (2003) "Processivity of ribozyme-catalyzed RNA polymerization," *Biochemistry* **42**, 8748-8755.

Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP (2001) "RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension," *Science* **292,** 1319-1325.

Curriculum Vitae

PERSONAL INFORMATION

Name and surname	**Rosa Karlić**
Work Address	Horvatovac 102A, Zagreb, Croatia
Work Phone	+385 / 1 / 4606276
E-mail	rosa@bioinfo.hr
Personal web page	http://www.bioinfo.hr
Citizenship	Croatian
Date and place of birth	March 20th, 1981; Zagreb, Croatia

EDUCATION

2011	**PhD in Bioinformatics**, Freie Universitaet, Berlin, Germany Thesis: Influence of histone modifications on mRNA abundance and structure Grade: Summa cum laude (with highest honor)
2006	**Diploma in Molecular biology**, University of Zagreb, Zagreb, Croatia Thesis: Computational classification of protein folds Grade: 5 (excellent)

WORK EXPERIENCE

2011 -	**Postdoc** , University of Zagreb, Faculty of Science, Zagreb, Croatia
2006 - 2011	**PhD student/teaching assistant**, University of Zagreb, Faculty of Science, Zagreb, Croatia
2007 - 2011	**PhD student**, Max Planck Institute for Molecular Genetics, Berlin, Germany

TEACHING

UNDERGRADUATE	Bioinformatics (2006 - 2007)
MASTERS LEVEL	Statistics and Machine Learning (from 2009); Computational Genomics (from 2012)
DOCTORAL LEVEL	Computational biology (2009 – 2011; MPI-MG, Berlin, Germany)

LANGUAGES

MOTHER TONGUE	Croatian
OTHER LANGUAGES	English (proficient), German (basic), French (basic)

AWARDS AND RECOGNITIONS

2011	L'Oreal Adria-UNESCO national fellowship "For women in science", Croatia
2010	Annual Award for PhD Students, Faculty of Science, Zagreb, Croatia

PUBLICATIONS

Karlić, R., H.R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron, Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A, 2010. 107(7): p. 2926-31.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identification and characterization of non-coding somatic mutations

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gad Getz (Broad Institute/MGH)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Esther Rheinbay (Broad Institute/MGH), Petar Stojanov (Broad Institute/DFCI), Gregory Kryukov (Broad Institute/DFCI)	

Background and preliminary data

The characterization of thousands of tumor exomes to date has greatly advanced our understanding of somatic driver events in different tumor types, leading to the identification of molecular tumor subtypes, and the discovery of oncogenes that occur at low frequencies in the patient population.

However, many tumors remain for which no driver event can be identified in the coding exome, suggesting the presence of oncogenic genomic alterations in the non-coding part of the genome. The recent discovery of somatic mutations in the promoter region of the TERT gene has set a precedent for a recurrent mutation that occurs in multiple tumor types. This mutation generates transcription factor binding sites in the TERT promoter, leading to increased expression levels. In recent years, chromatin signatures have been utilized to define distal regulatory elements (enhancers) in a variety of human normal and malignant cell lines. These enhancers are bound by a repertoire of chromatin regulating proteins and transcription factors, many of which are mutated in cancer. Recently acquired 3D-interaction maps have shown that enhancer-promoter interactions can vary between cell types, that enhancers can sometimes be mega-bases away from their promoter targets and even skip genes that lie between the regulatory site and the target gene. A special class of large enhancer domains, termed “superenhancers” appears to control lineage determining genes as well as oncogenes. Somatic mutations in enhancer sites and concomitant changes in transcription factor and chromatin regulator binding affinities may thus greatly affect the regulation of oncogenes but due to the complex interactions between distal sites and their targets, robust functional annotation of mutations is of even greater importance than in protein-coding genes. In addition to changes in the non-coding genome that directly affect chromatin structure and DNA-binding affinities, somatic mutations in non-coding but transcribed RNAs (lincRNAs, miRNAs and tRNAs) may affect protein complex composition, targeting to DNA sites, mRNA regulation and translational efficiency.

Given our inability to identify coding driver events in a large subset of malignancies, it is to be expected that somatic driver events in regulatory elements and recurrent lesions in non-coding transcripts are likely to exist in many, if not all, tumor types.

Timelines & resources dedicated to project

Timeline: Collection and definition of non-coding and possibly tissue-specific feature sets: February 2014. Development of statistical models for non-coding analysis: July 2014. Identification of candidate non-coding driver events: September 2014. Plans for experimental follow-up: October 2014. Manuscript preparation: January 2015.

Resources: 2000 whole genomes from ICGC/TCGA; epigenomic profiles for histone modifications, DNAse, and 3D-interaction data from ENCODE, Roadmap Epigenomics and other public resources. RNA-seq data for matched tumor/normal pairs; lincRNA catalogs from GENCODE and Cabili et al, 2011. Transcription factor binding motif matrices from Transfac, UNIPROBE, JASPAR, HOMER and recently published large-scale studies.

Research proposal

To identify high-confidence recurrent non-coding mutations and characterize them functionally, we plan the following:

1. Generation of high-confidence genomic features sets outside the coding exome. These sets include non-coding RNAs (lincRNA, miRNA), and regulatory regions such as promoters, 5'UTRs, 3'UTRs and enhancer elements. Features will be collected from Gencode, the Broad lincRNA catalog and mirBase. Epigenomic maps for regulation-associated histone modifications and DNAse hypersensitivity available through the ENCODE and Roadmap Epigenomics consortia will be used to define lineage-specific sets (where appropriate matched data are available) of regulatory sites as well as a comprehensive union set of features. This step will serve to enrich somatic mutations for those believed to be functional.

2. Development of a rigorous statistical framework for detection of significantly mutated genomic features. For this we will adapt the MutSig algorithm (Lawrence et al, 2013, developed in house) for analysis of non-coding mutations. MutSig uses a gene and patient-specific background model as well as genomic covariates to assess the significance of genes with recurrent mutations through evaluation of silent and non-coding mutations associated with coding genes. We will need to develop appropriate background models and identify relevant covariates for the non-coding feature sets.

3. Development of a tool to automatically evaluate functional impact of non-coding mutations. This tool will include evolutionary conservation, especially in promoter sites. Because enhancers and lincRNA tend to be poorly conserved even across mammals, the value of evolutionary conservation for these features will have to be determined first. In promoter and enhancer sites, creation and disruption of DNA-encoded sequence motifs (transcription factor binding sites, motifs related to transcriptional regulation) will be analyzed based on position-specific weight matrices available for many transcription factor families. ChIA-PET-derived interaction data between promoters and enhancers will serve as guide to identify regulatory targets of enhancers. Matched RNA-Seq (or other gene expression data) will be used to derive regulatory consequences of promoter and enhancer mutations. Impact of somatic mutations on lincRNAs will be evaluated based on RNA-Seq data for the lincRNA as well as it's protein target(s) (where known) in matched tumor/normal pairs. Where available, predictions on structural changes will be made.

4. **Experimental validation and functional characterization** of putative driver mutations through 3D-interaction experiments (3C, ChIA-PET), genome editing (CRISPR/TALEN) in human cell lines. For altered TF binding sites, ChIP of binding-motif associated TFs will be used to evaluate differential binding affinities.

Legacy plans

Tools and methods developed as part of this study will be made available to the research community. A database of putative non-coding somatic alterations will be published with the associated manuscript.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University

2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Esther Rheinbay	Postdoctoral associate
eRA COMMONS USER NAME (credential, e.g., agency login) esther@broadinstitute.org	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Tübingen, Germany	Diplom (M.Sc.)	07/2006	Informatik (Bioinformatik)
Boston University	Ph.D	05/2012	Bioinformatics
Massachusetts General Hospital		05/2012	Computational Biology
Broad Institute		05/2013	Computational Biology

A. Personal statement

My research interests revolve around changes in the chromatin regulatory landscape in normal and malignantly transformed cells, including the identification and functional study of regulatory elements that are dysregulated in cancer, and somatic driver mutations in these elements that lead to malignant transformation. My focus lies on analysis of both bulk tissue samples as well as cancer stem cells (cells with tumor-propagating potential) as the driving force of many cancer types. To accomplish these goals, I develop strategies for chromatin and genomic data management, processing and integrated analysis.

B.
Positions:
2003 Summer intern at Astra Zeneca Molndal, Sweden
2006-2012 PhD student with Dr. Simon Kasif, Boston University
2006-2012 PhD student with Dr. Bradley Bernstein, Massachusetts General Hospital/Broad Insititute
2012-2013 Postdoctoral fellos with Dr. Bradley Bernstein, Massachusetts General Hospital/Broad Institute
2013- Postdoctoral associate with Dr. Gad Getz, Broad Institute/Massachusetts General Hospital
Honors:
Verein fur Bildung und Begabung e.V. summer school scholarship (1999), e-fellows.net scholarship (2001-2011), Boston University Presidential Fellowship (2006-2007), Keystone Symposia Scientific meeting scholarship (2009, 2011).

C. Selected Peer-Reviewed Publications
Suva ML*, Rheinbay E*, Gillespie SM, Patel AP, Chi AS, Riggi N, Wakimoto H, Rabkin SD, Martuza RL, Rivera MN, Rossetti N, Beik S, Kasif S, Wortman I, Shalek A, Rozenblatt-Rosen O, Regev A, Louis DN and Bernstein BE: Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell,* revised manuscript submitted.

Rheinbay E*, Suva ML*, Gillespie SM, Oksuz O, Wakimoto H, Patel AP, Shahid M,
Rabkin SD, Martuza RL, Rivera MN, Louis DN, Kasif S, Chi AS and Bernstein BE: An Aberrant Transcription Factor Network Essential for Wnt Signaling and Stem Cell Maintenance in Glioblastoma. *Cell Reports*, 2013.

Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H,Carr SA, Bernstein BE: H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome Biology*, 2012.

Janiszewska M, Suva ML, Riggi N, Houtkooper RH, Auwerx J, Clément-Schatlo V, Radovanovic I, Rheinbay E, Provero P, Stamenkovic I: Imp2 controls oxidative phosphorylation and is crucial for preserving glioblastoma cancer stem cells. *Genes Dev*, 2012.

Rheinbay E, Louis DN, Bernstein BE, Suva ML: A tell-tail sign of chromatin: histone mutations drive pediatric glioblastoma. *Cancer Cell*, 2012.

Presser Aiden A*, Rivera MN*, Rheinbay E, Ku M, Coffman EJ, Truong TT, Vargas SO, Lander
ES, Haber DA and Bernstein BE: Wilms tumor chromatin profiles highlight stem cell properties and a renal developmental network. *Cell Stem Cell*, 2010.

Ku M*, Koche RP*, Rheinbay E*, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, Adli M, Kasif S, Ptaszek LM, Cowan CA, Lander ES, Koseki H, Bernstein BE: Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLOS Genetics*, 2008.

Meraldi P*, McAinsh AD*, Rheinbay E and Sorger PK: Structural and phylogenetic analysis of centromeric DNA and kinetochore proteins. *Genome Biology*, 2006.
*equal contribution

Petar Stojanov
Broad Institute Cancer Program
301 Binney Street,
Cambridge, MA 02141
Mobile: 845-901-6951
petar.stojanov@gmail.com

OBJECTIVE

Admission to a doctoral program in the field of computational systems biology and bioinformatics

EDUCATION

- **College**, Bachelor of Arts, major in Computer Science, Bard College, Annandale-on-Hudson, NY, 2006-2010
- **High School,** Pelham Memorial High School, Pelham, NY (Senior Year) 2005-2006
- **High School**, Rade Jovcevski Korcagin, Skopje, Macedonia, 2002-2005

RESEARCH AND WORK EXPERIENCE

Broad Institute Cancer Program/DFCI, 10/2010 – Present

- **Associate Computational Biologist** at the Cancer Genome Analysis group, under mentorship and supervision of Dr. Gad Getz and Dr. Michael Lawrence. Main focus was development of algorithms for analysis of Next-Generation sequencing data and their application in high impact whole exome and whole genome sequencing projects of different tumor types. Selected projects during this experience:

    - Developed a prototype workflow using modifications of pre-existing methods, for analysis of paraffin embedded (FFPE) samples and classifying mutations in somatic vs. germline without a matched normal, by taking advantage of the fact that tumor samples are contaminated with normal cells.

    - Assisted in the analysis of 5000 cancer exomes across 21 tumor types to determine the how far we are from making a full catalog of all driver mutations in cancer. We performed rigorous statistical analysis to detect all known and 33 novel cancer genes, and we carried out down-sampling and saturation experiments, and we determined that the power to detect new genes has not plateaued with 5000 patients, and that there are more cancer genes to be discovered by increasing the data set. This work was accepted in ***Nature*** and is currently in press.

    - Helped develop a statistical significance analysis method with Dr. Lawrence and Dr. Getz, which models a genome-wide background mutation frequency. We applied this method to 3000 tumor-normal pairs of whole exomes to elucidate the different mutational

processes (mutation spectra) that are prevalent in cancer. We also demonstrated a greatly improved ability to identify key driver genes and eliminate false positive candidates in lung squamous cancer using this background model. Along with Michael Lawrence and Paz Polak, we published these findings in ***Nature*** as co-first authors (Lawrence, Stojanov, Polak et al. 2013). (Please see "Publications" on last page).

- Along with Dr. Austin Dulak, Dr. Shouyong Peng and the lab of Dr. Adam Bass, I took the lead on analyzing the first large set of 150 whole exomes of esophageal adenocarcinoma. We discovered a novel mutational spectrum defined by A->C transversions in AA di-nucleotides. We also identified putative novel driver genes for esophageal adenocarcinoma that were significantly mutated in our cohort (such as ELMO1 and DOCK2 from the RAC1 pathway). We published these findings in ***Nature Genetics*** as co-first authors (Dulak, Stojanov, Peng et al. 2013)

- Worked with Dan-Avi Landau, Scott Carter, and Catherine Wu on elucidating the clonal penetrance of driver mutations in chronic lymphocytic leukemia. We analyzed a cohort of 149 patients, where we used our statistical significance algorithms to nominate driver genes. We used ABSOLUTE (Carter, Cibulskis, Helman et al. 2012) to infer the purity, ploidy, and absolute allelic copy number at each locus. This information helped us calculate a cancer cell fraction for each mutation that we used to classify mutations in clonal and subclonal. We used this approach to compare pre-treatment and post-treatment stages for certain patients, and we observed that in samples that were treated, subclonal driver mutations underwent clonal evolution, thus overtaking the population of cells. We published these results in ***Cell*** as co-first authors (Landau, Carter, Stojanov et al. 2013).

- Along with Dr. Jens Lohr and Dr. Todd Golub, took the lead on analyzing the largest multiple myeloma sequencing effort to date. We analyzed 203 tumor-normal pairs to determine significantly mutated genes and clonal penetrance of driver mutations. We found that oncogenic driver mutations (KRAS, NRAS, BRAF) can be subclonal together in the same population. We also discovered clonal EGR1 mutations near the 3' end that are enriched in WRCY motifs indicating somatic hypermutation and suggesting a potentially novel immunoglobulin (IG) locus translocation. Dr. Lohr and myself will publish this work as co-first authors, as it was accepted in ***Cancer Cell*** and is currently in press.

- Collaborated with Dr. Lohr and Dr. Golub on one of the first high-throughput DNA sequencing projects on diffuse-large B-cell lymphoma (DLBCL). We analyzed 49 tumor-normal pairs and used our statistical significance analysis that I developed (described below), to nominate genes that are significantly mutated in our cohort, with mutations clustered in specific evolutionarily conserved hotspots. We discovered several new potential driver genes, and observed somatic hyper-mutation and putative negative selection of mutations harbored by BCL2. We published these results in ***PNAS*** in early 2012 (Lohr, Stojanov, Lawrence et al. 2012).

- Developed a statistical tool that ranks genes according to the positional configuration of their somatic mutations and the base-level evolutionary conservation of each base. This method has been developed under mentorship of Michael Lawrence, and used in multiple high-profile projects (including the DLBCL project described above),

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Landscape of somatic mutations affecting eQTLs and meQTLs

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gaddy Getz, Broad Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Ayellet V. Segrè, Broad Institute; David DeLuca, Broad Institute	

Background and preliminary data

Tumors acquire multiple genetic and epigenetic alterations that substantially influence transcript abundance in tumors, such as somatic copy number changes or DNA methylation changes. However, the extent to which somatic mutations that alter gene expression drive tumor progression compared to changes in protein structure or function, is not well understood. Furthermore, the functional consequences of somatic mutations in noncoding regions are hard to interpret from sequence alone. One approach to identifying regulatory effects of germline or somatic mutations is to conduct genome-wide association tests of DNA variants with gene expression levels (eQTL) or epigenetic mark status, such as DNA methylation at CpG sites (meQTL). A few studies suggest that genetic associations with cancer risk (found through genome-wide association studies, GWAS of germline DNA variants) may be exerting their causal effect through eQTLs (E.g., Li et al, Cell 2013). In this work, we will integrate ICGC/TCGA whole-exome, whole-genome, and RNA sequencing data with external functional genomics data to (1) evaluate the extent to which somatic mutations may be functioning through alteration or creation of eQTLs or meQTLs in different human cancers, and to (2) further our understanding of coding and non-coding somatic mutations that affect the expression of protein coding genes and noncoding RNAs in the pan-cancer tumors.

The main resource of eQTLs we plan to use for this project will be from the Genotype-Tissue Expression (GTEx) project that is co-led by Gaddy Getz and Kristin Ardlie at the Broad Institute, where the Laboratory, Data Analysis and Coordinating Center (LDACC) resides (https://commonfund.nih.gov/GTEx, http://www.broadinstitute.org/gtex/). The goal of GTEx is to generate a comprehensive catalog of eQTLs and tissue-specific transcriptomes for a wide range of normal human tissues. Up to ~50 tissues are being collected from hundreds of postmortem donors, including 10 sections of the brain when possible, with the goal of reaching 900 donors over the next two years. Genome-wide genotyping and whole exome sequencing is conducted on the individuals' DNA, and RNA-sequencing is conducted on mRNA (polyA libraries) from all tissue samples, including noncoding RNA genes. Almost all tissues studied in the pan-cancer project are represented in the GTEx project. Genome-wide profiling of epigenetic marks, such as DNA methylation, are also planned for the coming year in a subset of the GTEx tissues. In addition, to GTEx, for obtaining meQTL profiles in normal tissues or cell lines will will use the ENCODE and Roadmap Epigenomics projects.

To analyze the RNA-sequencing data matched to WGS pan-cancer samples, we will use a RNA-Seq quality control (QC) method developed by David DeLuca and Gaddy Getz (DeLuca et al, Bioinformatics 2012) that was installed into our analysis infrastructure, Firehose, developed at the Broad Institute. This RNA-Seq QC pipeline was applied to thousands of RNA-Seq data sets in GTEx. We have also implemented an eQTL analysis pipeline in Firehose that is being used in GTEx (Matrix eQTL). We will extend the eQTL detection method, currently set up for point mutations, for analyzing indels and copy number variants, and for detecting meQTLs.

Several computational methods have been developed by our groups at the Broad (Getz lab and Meyerson lab) for calling somatic point mutations (MuTect) or copy number changes in whole exome tumor-normal pair sequences and have been applied to thousands of pan-cancer tumors. These methods are being extended and optimized by our group for whole genome sequences (MuTect 2.0, Kristian Cibulskis et al.; Indel detection, Adam kiezun et al.; Copy number alterations, Marcin Imielinski et al.). We plan to use somatic mutation calls from these methods for the ~2000 pan-cancer tumor-normal whole genome sequences (WGS).

With the tools and resources we have, or will develop, in house we will be able to integrate WGS, RNA-Seq data, and eQTL and meQTL data, to help shed light on potential regulatory mechanisms of somatic mutations in human cancers.

Timelines & resources dedicated to project

Timeline: eQTL annotation and analysis of somatic mutations of up to 2000 tumor-normal genomes by Sept 2014; meQTL annotation of somatic mutations of 1200-2000 genomes by Dec 2014; Validation of results with RNA-Seq data by March 2014; Manuscript preparation/submission by June 2015.

Resources: Tumor and normal whole exome and whole genome read data, copy number segmentation, RNA-Seq data, DNA methylation data.

Research proposal

We will characterize the landscape of somatic mutations that affect or create expression quantitative trait loci (eQTLs) and DNA methylation quantitative trait loci (meQTLs) across the pan-cancer tumor types, through integrative analysis of ICGC/TCGA whole genome sequences of tumor-normal pairs and RNA sequencing data, a reference panel of eQTLs from dozens of normal human tissues (the GTEx project), and meQTL profiles from various resources. **To address this we propose the following goals:**

1) Characterize overlap of somatic mutations with eQTLs found in normal tissues. We will assess the extent to which somatic mutations in noncoding and coding regions in the different pan-cancer tumor types affect eQTLs that naturally occur in the human population in relevant healthy tissues. **We will characterize the number, density and location of somatic mutations that are in linkage disequilibrium to eQTLs in relevant normal tissues,** genome-wide and in regions with recurrent somatic mutations within a tumor type or across tumors. **We will analyze somatic point mutations, small insertions and deletions (indels), and copy number changes in the ~2000 whole genome-sequence (WGS) tumors**, that will be called by methods being developed for WGS by our group at the Broad (MuTect 2.0, Kristian Cibulskis et al.; Indel detection, Adam kiezun et al.; Copy number alterations, Marcin Imielinski et al.). **To evaluate the significance of the overlap**, **we will test for enrichment of eQTLs amongst somatic mutations genome-wide, and near the subset of cancer driver genes** identified in the pan-cancer whole exome study. **For normal tissue eQTLs, we will use the eQTLs that are being generated in the Genotype-Tissue Expression (GTEx) project for ~50 tissues**. We will begin by investigating *cis* eQTLs; we will test *trans* eQTLs as GTEx sample sizes increase from 100-150 to >300-500. For our first round of ICGC/TCGA WGS analyses, we will use eQTLs idenitfied for 9 tissues in the Pilot Phase of GTEx (~175 donors, 80-160 samples / tissue). Of these, blood, lung, and sun-exposed skin match pan-cancer tumors with WGS. eQTLs for the other 40 tissues will become available over the next year or two.

2) Identify somatic mutations that create new eQTLs or alter existing ones by analyzing RNA-sequencing data from tumors and their matched normal tissues. Having RNA-seq data from tumor samples and their normal tissue counterpart provides a unique **opportunity to identify eQTLs created *de novo* in tumors that are not present in their matched normal tissue and eQTLs that are altered or lost in tumors.** To do so, we will use our RNA-Seq quality control (DeLuca *et al*., 2012), microarray normalization, and eQTL analysis pipelines we installed in Firehose, to analyze the RNA-seq or cDNA microarray data matched to WGS for ~1,500 cancer samples, and to detect eQTLs in tumor versus normal tissues. We will characterize the differences in number and nature of eQTLs between tumor and normal tissues, similarly to that proposed in Aim 1.

3) Investigate the underlying regulatory mechanism through which somatic mutations may be altering or creating eQTLs. Due to linkage disequilibrium, identifying the causal mutation/s driving the eQTLs is not trivial. We will investigate the underlying mechanisms of eQTL alteration, such as disruption of regulator binding sites, by testing for enrichment of somatic mutations in eQTL regions amongst various genomic features predicted to have a regulatory effect, taken from the ENCODE and Roadmap Epigenomics projects. We will collaborate with Esther Rheinbay in our group who has expertise in functional characterization of noncoding regions.

4) Test for Loss of Heterozygosity in tumor suppressor gene regions with heterozygous eQTLs. We will test the hypothesis that individuals with germline heterozygous eQTLs for tumor suppressor genes, are more likely to obtain only one somatic mutation hit compared to homozygous individuals, by undergoing loss of heterozygosity (LOH), where the expressing allele of the tumor suppressor gene is lost. We will test a similar hypothesis for oncogenes, though in the opposite direction of expression. We will corroborate our results in pan-cancer samples with RNA-Seq data and WGS.

5) Apply similar framework to examine intersection and enrichment of somatic mutations with DNA methylation quantitative trait loci (meQTLs). We will extend the analytical pipeline that we will develop for eQTLs, to address similar questions with somatic mutations and meQTLs. We will use meQTLs that we will detect in the ~1200-1400 pan-cancer samples with DNA methylation profiles, as well as meQTLs that will be profiled in GTEx, and from the ENCODE and Roadmap Epigenomics projects.

This work will be a first step towards testing the extent to which somatic mutations may be driving tumorigenesis through altered regulation of gene expression versus alteration of protein function. Our analyses should help shed light on the functional consequences of noncoding somatic mutations and regulatory coding mutations, and may have important therapeutic implications.

Legacy plans

We will make available to the community our annotations of somatic mutations that overlap eQTL or meQTL regions found in normal or tumor samples, and their putative functional consequences based on overlap with known or predicted regulatory elements or chromatin marks.

eRA COMMONS USER NAME (credential, e.g., agency login)
GADGETZ

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

C. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

Curriculum Vitae December 30, 2013

Ayellet Vered Segrè (Maiden name: Falcovitz)
Broad Institute of Harvard and MIT, 301 Binney Street, Cambridge, MA 02142, USA
Work: 617-714-7836, e-mail: asegre@broadinstitute.org

CURRENT POSITION

9/2007 – 2/2013 **Computational Biologist, Broad Institute of Harvard and MIT,** Cancer Program, Group of Gaddy Getz.

EDUCATION AND TRAINING

2007-2013 **Postdoctoral fellow**, Broad Institute of Harvard and MIT, Program of Medical and Population Genetics, Lab of David Altshuler.

2001-2007 **Ph.D., Genetics and Genomics, Harvard University,** Dept. of Molecular and Cellular Biology, Cambridge, MA. Advisor: Prof. Andrew Murray.

1998-2001 **M.Sc. in cancer research, Weizmann Institute of Science**, Israel, Dept. of Molecular Cell Biology. Advisor: Prof. Varda Rotter.

1995-1998 **B.Sc., Hebrew University of Jerusalem**, Israel, Faculty of Life Sciences, *magna cum laude*.

1997-1998 Hebrew University of Jerusalem, Israel, Dept. of Genetics, **Undergraduate research assistant**, Lab of Prof. Nissim Benvenisty.

1996 Technion, Israel, Dept. of Biology, **Summer student**, Lab of Prof. David Gershon.

1993-1995 Military service in the Israel Defense Forces.

FELLOWSHIPS AND HONORS

2011 Young Investigator Grant Award for Presenters at the American Diabetes Association 71st Scientific Sessions.

2008 American Diabetes Association Mentor-Based Postdoctoral Fellowship.

2008 Awarded American Diabetes Association-TAKEDA Cardiovascular Postdoctoral Fellowship

2003 Financial aid award for Ph.D. studies, Harvard University: Michael and Anna Vranos Graduate Fellowship Fund in the Life Sciences.

1998 Feinberg Graduate School Fellowship for M.Sc. studies.

1996/97, 1997/98 Dean's List (B.Sc. degree).

1997 Amos De Shalit Summer Program for outstanding undergraduate students, Weizmann Institute of Science, Israel.

SELECTED PEER-REVIEWED PUBLICATIONS

1. Andrew Morris*, Ben Voight*, Tanja Teslovich*, Teresa Ferreira*, **Ayellet V. Segrè***, Valgerdur Steinthorsdottir, [.. 220 authors ..], Josée Dupuis, James B. Meigs, David Altshuler, Michael L. Boehnke, Mark I. McCarthy for the DIAGRAM consortium. *Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes*. **Nature Genetics,** 2012, Sep 44(9): 981-90.
***Equally contributing first co-authors.**

2. Josep M Mercader, Montserrat Puiggros, **Ayellet V. Segrè**, et al. *Identification of novel type 2 diabetes candidate genes involved in the crosstalk between the mitochondrial and the insulin signaling systems.* **PLoS Genetics,** 2012, 8(12):e1003046.

3. Hao Zhu*, Ng Shyh-Chang*, **Ayellet V. Segrè**, et al. *The Lin28/let-7 axis regulates glucose metabolism.* **Cell**, 2011, 147(1): 81-94.

4. **Ayellet V. Segrè,** et al. *Common Inherited Variation in Mitochondrial Genes is not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits.* **PLoS Genetics**, 2010, Aug 12; 6(8). pii: e1001058.

5. **Ayellet V. Segrè**, Andrew W. Murray and Jun-Yi Leu: *High-Resolution Mutation Mapping Reveals Parallel Experimental Evolution in Yeast*. **PLoS Biology**, 2006, 4(8): e256.

BROAD INSTITUTE OF HARVARD AND MIT, 301 BINNEY STREET, CAMBRIDGE, MA 02142, USA
PHONE 617-714-8362• E-MAIL DDELUCA@BROADINSTITUTE.ORG

DAVID S. DELUCA

CURRENT POSITION

2011 to present — The Broad Institute

Computational Biologist, Group of Gad Getz

EDUCATION AND TRAINING

2008 to 2011 — Harvard Medical School

Harvard Research Fellow – Cancer Vaccine Center, Dana-Farber Cancer Institute

2005 to 2008 — Universität Hannover

D.Sc. Chemical Engineering and Immunogenetics

- German doctoral degree, *Dr. rer. nat.*
- Dissertation title, "Computational immunogenetics in allogeneic immunotherapy"; work performed at Hannover Medical School

2002 to 2005 — Universität Hannover

M.Sc. Life Science

- Major concentrations: Bioinformatics & Bio Process Engineering
- Minor concentrations: Molecular Biology & Natural Product Chemistry

1998 to 2002 — Carnegie Mellon University

B.Sc. Computational Biology

- Originally titled, *Biology with a Computer Science Track*

PROFESSIONAL EXPERIENCE

2002 to 2008 — Hannover Medical School

Research Associate

- Developed classification algorithms, statistical optimizations and databases for immunogenetic research and applications (for predicting MHC-peptide binding, minor histocompatibility antigens, haplotype analysis, HLA typing, PCR primer design)
- Supervised database programmer in design and implementation of data warehouse. Supervision of student for web and relational database development.
- Implementation of dynamic websites and database-driven stand-alone applications.

2009 to 2010 — Boston University Metropolitan College

Visiting Scholar – Computer Science Department

- Co-development and teaching of new Biomedical Informatics Course
- Development of interdisciplinary research projects
- Online course facilitation training and participation

AWARDS RECEIVED

2005 European Federation for Immunogenetics Best Abstract: "Comprehensive Peptide Binding Prediction by Developing a Modular Concept for HLA Peptide Binding Pockets"

2007 European Federation for Immunogenetics bursary (support for young researchers)

APPOINTMENTS AND MEMBERSHIPS

2010 to 2011 International Immunomics Society
Secretary General

2010 International Conference on Bioinformatics
Program Committee Member

2007 to 2008 Data Interoperability Steering Committee, DAIT, NIAID
Ontology developer on the HLA Subcommittee

2005 to 2008 European Federation for Immunogenetics
Member, conference attendee and frequent presenter

2004 to 2008 German Federation for Immunogenetics
Member, conference attendee and frequent presenter

SELECTED PEER-REVIEWED PUBLICATIONS

1. **DeLuca D. S.**, Levin, J.Z. Sivachenko A, Fennell T, Nazaire M.D., Williams C, Reich M, Winckler W, Getz G. RNA-SeQC: RNA-seq metrics for quality control and process optimization *Bioinformatics* 28 (11), 1530-1532

2. **DeLuca, D. S.**, Eiz-Vesper, B., Ladas, N., Khattab, B. A., and Blasczyk, R. (2009) High-throughput minor histocompatibility antigen prediction, *Bioinformatics 25*, 2411-2417.

3. **DeLuca, D. S.**, Beisswanger, E., Wermter, J., Horn, P. A., Hahn, U., and Blasczyk, R. (2009) MaHCO: an ontology of the major histocompatibility complex for immunoinformatic applications and text mining, *Bioinformatics 25*, 2064-2070.4

4. **DeLuca, D. S.**, Khattab, B., and Blasczyk, R. (2007) A modular concept of HLA for comprehensive peptide binding prediction, *Immunogenetics 59*, 25-35.

5. **DeLuca, D. S.**, and Blasczyk, R. (2007) The immunoinformatics of cancer immunotherapy, *Tissue Antigens 70*, 265-271.

6. Elsner, H.A., **DeLuca, D.**, Strub, J. & Blasczyk, R. HistoCheck: rating of HLA class I and II mismatches by an internet-based software tool. *Bone Marrow Transplant* **33**, 165-169 (2004).

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 31th December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrated genomic analysis of cancer drivers and pathways

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gad Getz, MGH / Broad Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Lihua Zou, Broad Institute	

Background and preliminary data

Timelines & resources dedicated to project

Worldwide cancer genome projects such as TCGA and ICGC have revealed a large number of genetic alternations by sequencing primary tumors of thousands of patients. One fundamental problem of cancer research is to classify 'driver' from 'passenger' event based on whether it confers selective advantage on tumor initiation and progression. Current computational methods typically rank drivers based on recurrence of somatic events from a patient cohort based on a single platform. However, it remains challenging to identify many of the low-frequency cancer drivers. One promising approach is to perform integrative analysis across multiple platforms to improve the statistical power beyond single platform based approach. Toward this goal, we have developed an integrative analysis framework to reverse engineer dis-regulated cancer pathways and to rank novel cancer drivers (NetSig).

Timeline:
2014.01-03: extend current network analysis methods (NetSig) to incorporate WGS data
2014.03-09: joint analysis of mutation and transcriptome/epigenome of ICGC/TCGA data
2014.09-12: manuscript preparation

Resources:
Somatic variant calls of ICGC/TCGA WGS data; Copy number (GISTIC) calls of ICGC/TCGA tumor data; Tumor matched Level-3 RNAseq/Methylation/RPPA data for the WGS data

Research proposal

We plan to develop an integrative analysis pipeline using the ICGC/TCGA dataset to address several questions:

1) to prioritize low-frequency drivers that tend to be missed by cohort-based approach. We plan to extend the network-module based approach (NetSig) developed in house to increase the statistical power for calling novel driver events.

2) to reverse engineer dis-regulated cancer pathways by integrating somatic mutations, transcriptome and epigenome to gain insights about the cellular context of the driver events in tumor development and progression.

3) to develop a web portal to collect dis-regulated cancer pathways 'mined' from the ICGC/TCGA data and make it available to the entire cancer research community for visualization and easy exploratory analysis.

Legacy plans

We will create a resource to collect and curate the dis-regulated pathways identified from the Pan-cancer study. We will develop a web portal for visualization and exploratory analysis of the collected cancer pathways and make it freely accessible for the cancer research community.

eRA COMMONS USER NAME (credential, e.g., agency login)
GADGETZ

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

B. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding

sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Lihua Zou	Computational Biologist
eRA COMMONS USER NAME (credential, e.g., agency login) lihuazou	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Beijing Institute of Light Industry, China	B.A.	1997	Biochemical Engineering and Computer Science
University of Illinois at Urbana-Champaign	Ph.D.	2004	Biophysics

A. Personal Statement

My research background and expertise are in statistical modeling and computational biology. In particular, I am interested in developing statistical methods to integrate multi-dimensional genomic data to understand dis-regulated pathways in cancer and common genetic diseases. In addition, I am working on questions to identify biomarkers for predicting drug and therapeutic response in cancer.

B. Positions and Honors

Positions:

1999-2004: Univ. Illinois at Urbana-Champaign/Center of Biophysics. PhD student
2004-2006: Harvard University/Department of Statistics. Postdoc
2006-2009: Dana-Farber Cancer Institute/Harvard Cancer Center. Research Scientist.
2009-2010: Yale University/Yale Medical School. Research Scientist.
2010- Broad Institute of MIT and Harvard. Computational Biologist.

C. Selected Peer-reviewed Publications

1. Ge H, Player CM, **Zou L.** Toward a global picture of development: lessons from genome-scale analysis in Caenorabditis elegans embryonic development. ***Dev Dyn***. 2006 Aug; 235(8): 2009-17.
2. Shioda T, Chesnes J, Coser KR**, Zou L**, Jingyung Hur, Kathleen L. Dean, Carlos Sonnenschei, Ana M. Soto, Kurt J. Isselbacher, Importance of dosage standardization for interpreting
3. Sarah Walker, Erik Nelson, **Lihua Zou**, Mousumi Chaudhury, David Frank, Reciprocal effects of STAT5 and STAT3 in breast cancer, ***Molecular Cancer Research*** 2009 Jun; 7(6): 966-76.
4. Patrycja V. Missiuro, Kesheng Liu**, Lihua Zou**, Brian C. Ross, Guoyan Zhao, Jun S. Liu, and Hui Ge, Information flow analysis of interactome networks, ***PLoS Computational Biology*** 5(4): e1000350, 2009.
5. **Lihua Zou,** Sira Sriswasdi, Brian Ross, Patrycja V. Missiuro, Jun Liu, Hui Ge, Systematic Analysis of Pleiotropy in *C. elegans* Early Embryogenesis, ***PLoS Computational Biology*** 4(2): e1000003, 2008.
6. Hailing Cheng, Pixu Liu*, Zhigang C. Wang*, **Lihua Zou***, Stephanie Santiago, Ole V. Gjoerup, J. Dirk Iglehart, Alexander Miron, Andrea L. Richardson, William C. Hahn and Jean J. Zhao, SIK1 couples LKB1 to p53-dependent anoikis and suppresses metastasis, ***Science Signaling*** 2009 Jul 21; 2(80): ra35
7. Yang Li*, **Lihua Zou***, Qiyuan Li, Ruiyang Tian, Yan Li, Zoltan Szallasi, Benjamin Haibe-Kains, Christine Desmedt, Christos Sotiriou, J. Dirk Iglehart, Andrea L. Richardson and Zhigang Charles Wang, Amplification of LAPTM4B and

YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer, ***Nature Medicine*** 2010 Jan 24; 16(2): 214-218.

8. Irie HY, Shrestha Y, Selfors LM, Frye F, Iida N, Wang Z, **Zou L**, Yao J, Lu Y, Epstein CB, Natesan S, Richardson AL, Polyak K, Millis GB, Hann WC, Brugge JS, PTK6 regulates IGF-1 dependent anchorage-independent survival of breast and ovarian cancer cells, ***PLoS One*** Jul 23; 5(7): e11729, 2010.
9. Sang Hyun Lee, George Poulogiannis, Saumyadipta Pyne, Shidong Jia, **Lihua Zou**, Sabina Signoretti, Massimo Loda, Lewis Cantley and Thomas M. Roberts, A constitutively activated allele of the p110 beta catalytic subunit of PI3-kinase induces prostatic intraepithelial neoplasia in mice, ***Proc Natl Acad Sci USA*** Jun 15; 107(24): 11002-7. Epub 2010 Jun 1.
10. Gonzalez-Malerva L, Park J, **Zou L**, Hu Y, Pearlberg J, Sawyer J, Stevens H, Harlow E, LaBaer J, High throughput ectopic expression screen identifies HSPB8, a kinase that blocks autophagy and confers tamoxifen resistance, ***Proc Natl Acad Sci USA*** February 1, 2011 vol. 108 no. 5 2058-2063.
11. Regulation of TFEB and V-ATPases by mTORC1. Peña-Llopis S, Vega-Rubin-de-Celis S, Schwartz JC, Wolff NC, Tran TA, **Zou L**, Xie XJ, Corey DR, Brugarolas J. ***EMBO J.*** 2011 Jul 29;30(16):3242-58.
12. Jay E. Mittenthal, **Lihua Zou**, Multi-input networks are more likely to signal a conjunction through disinhibition than activation, ***Mathematical Biosciences.*** May;231(1):69-75. Epub 2011 Feb 15.
13. Descovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole exome sequencing. ***Proc Natl Acad Sci USA***. Mar 6; 109(10): 3879-84 2012
14. Sequence analysis of mutations and translocations across breast cancer subtypes. ***Nature***. Jun 20; 486(7403): 405-9. 2012
15. Comprehensive genomic characterization of squamous cell lung cancers. ***Nature*** Sep 28; 489(7417):519-25 2012
16. Comprehensive molecular portaits of human breast tumors. ***Nature.*** Oct 4; 490(7418):61-70. 2012
17. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. ***N Engl J Med***. May 30; 368(22):2059-74. 2013
18. Integrated genomic characterization of endometrial carcinoma. ***Nature*** Aug 8; 500-7. 2013
19. Mutational heterogeneity in cancer and the search for new cancer-associated genes. ***Nature***. Jul 11; 499(7457): 214-8 2013
20. The Cancer Genome Atlas Pan-Cancer Project. ***Nature Genetics*** Oct; 45(10: 1113-20 2013
21. The Somatic Genomic Landscape of Glioblastoma. ***Cell*** Oct 10; 155(2): 462-77 2013

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The use of integrative whole genome sequencing for precision cancer medicine

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Gad Getz (Broad Institute/MGH)
Levi A. Garraway (Broad Institute/DFCI)

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Eliezer M . Van Allen (Broad Institute/DFCI) Nikhil Wagle (Broad Institute/DFCI)	Paz Polak (Broad Institute/MGH)

Background and preliminary data

Numerous targeted agents directed as specific genetic vulnerabilities have been deployed clinically and led to improved patient outcomes. Examples include EGFR inhibitors in EGFR-mutant lung adenocarcinoma, and RAF inhibitors in BRAF-mutant melanoma. Broader use of prospective genomic profiling may impact patient care in oncology across tumor types, and towards that end, the scope of genomic profiling has widened to capture increasing amounts of clinically relevant events for a given patient. Clinical genomic profiling that began with hotspot mutation detection has transitioned towards targeted sequencing using large (n = 200-1000) gene panels for prospective clinical management. Furthermore, the use of whole exome sequencing to guide precision cancer medicine has recently been demonstrated as feasible and potentially informative for individual patients (Van Allen, Wagle et al, *in press*, Nature Medicine). In this study, we demonstrated how the use of whole exome sequencing to guide patient care resulted in clinically meaningful outcomes for individual patients, including enrollment on clinical trials of targeted therapies that would otherwise not be recognized. Moreover, our preliminary studies on 3,277 whole exomes from TCGA patients demonstrated that 89.4% (2,928/3,277) of these patients harbor at least one alteration in a clinically relevant cancer gene (Yuan, Van Allen, et al, *in review*). In this study, we defined clinically relevant genes as those that, when altered, may predict response or resistance to therapies, or have a diagnostic and/or prognostic utility. Among this cohort, we demonstrate that rare clinically relevant alterations can emerge in unexpected settings, such as an activating *BRAF* mutation in a renal clear cell cancer or activating *MTOR* mutations across numerous tumor types.

However, the use of whole exome sequencing alone is still limited in its ability to fully characterize an individual patient's tumor from a clinical perspective. For instance, this data does not include information about clinically relevant translocations (e.g. BCR-ABL), nor can it be used for discovery of non-coding alterations in clinically relevant cancer genes that may have important impact. Finally, the ability to classify tumors from a diagnostic perspective based on mutational profiling (rather than histology techniques) remains out of reach, but could result in improvements in diagnostic accuracy and subsequent clinical management for tumors that are reclassified. A thorough assessment of whole genome sequencing data from a clinical perspective may lead to the discovery of additional clinically relevant alterations across tumor types, and inform its relative increased utility (and when integrated with other datasets) for improving individualized patient care in oncology from a predictive and diagnostic perspective.

Timelines & resources dedicated to project

Timeline:

March, 2014: Identification of non-coding alterations in clinically-relevant cancer genes. Identification of clinically actionable translocations across tumor types.

July, 2014: Development of integrative algorithms to assess the functional impact of nominated non-coding alterations

September, 2014: Application of tissue-of-origin algorithm to identify clinically-relevant classifications

January, 2015: Manuscript preparation

Research proposal

We will perform an integrative whole genome clinically-focused study to address the following topics:

1) **Identification functionally relevant non-coding alterations in clinically actionable cancer genes.** Efforts to identify alterations in clinically relevant cancer genes have been limited to exon/coding regions that carry a specific effect (e.g. *BRAF*^{V600E} in melanoma). We will utilize the ICGC/TCGA whole genome data set to identify non-coding alterations in clinically relevant cancer genes (n = 125; Van Allen, Wagle et al, *in press*, Nature Medicine). Then, we will develop algorithms to predict which of these non-coding alterations may result in the predicted functional impact that leads to clinically relevant effects by integrating these findings with corresponding transcriptome data. For instance, non-coding alterations in *PTEN* from tumors that are otherwise considered *PTEN* wild-type (e.g. without non-synonymous loss-of-function alterations in exons or focal deletions) will be assessed for decreased PTEN expression and increased expression of other PI3-kinase pathway members (and pathway activation). This effort should translate into a candidate set of non-coding alterations that will be functionally assessed in collaboration with members of the Garraway Lab.
2) **Landscape of clinically relevant alterations that extend the "long tail".** Initial studies of pan-cancer data that identify alterations in clinically relevant cancer genes have been limited to exome sequencing data and miss many critical alterations. With whole genome and transcriptome data, we will utilize a heuristic algorithm to identify all clinically relevant alterations across the > 2,000 ICGC/TCGA cases. This will include translocations, non-coding alterations, and methylation data in combination with previously characterized events (e.g. exon-focused point mutations, short insertion/deletions, and copy number alterations). By describing the landscape of such alterations in this expansive fashion, we hypothesize there will be a "long tail" of clinically relevant alterations that predict response or resistance to therapies and occur at low frequencies in unexpected tumor types. Some of these events involve processes we could not previously test for (e.g. rare BRAF translocations in multiple tumor types not identifiable with exome data). Identification of these events may inform future clinical trial design and patient testing.
3) **The use of whole genome data for diagnostic purposes across tumor types.** Tumors are often categorized diagnostically using histologic techniques (e.g. transitional cell carcinoma versus squamous cell carcinoma in urothelial malignancies). However, the use of mutation and epigenetic patterns can reclassify tumor types as behaving separately from those they are histologically related to, which may have clinical relevance for treatment selection. Furthermore, some tumor types are of unknown primary, and whole genome based methodology that can be used for diagnostic purposes may have immediate clinical impact in these cases. We will apply an algorithm to ICGC/TCGA whole genome sequencing data to reclassify tumors based on mutational patterns rather than histologic subtype to determine whether such classification conflicts occur and address the therapeutic relevance of these events.

Legacy plans
Tools and methods developed for this study will be made available to the research community. Code will be published for download, and relevant databases will be made available online.

eRA COMMONS USER NAME (credential, e.g., agency login)
GADGETZ

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

C. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

C. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science
2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

D. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding

sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

CURRICULUM VITAE

Date Prepared: October 8, 2013

Name: LEVI ALEXANDER GARRAWAY

Office Address: Dana Building, Room 1542,
Dana-Farber Cancer Institute
44 Binney Street
Boston, MA 02115

Home Address: 363 Walnut Street
Newton, MA 02460

Work Phone: 617-632-6689

Work E-Mail: levi_garraway@dfci.harvard.edu

Work FAX: 617-582-7880

Place of Birth: Oakland, California

Education

1990	A.B.	Biochemical Sciences	Harvard College, Cambridge, MA
1999	M.D.	Medicine	Harvard Medical School, Boston, MA
1999	Ph.D.	Biological Chemistry and Molecular Pharmacology (Ph.D. Adviser: Dr. Stephen M. Beverley)	Harvard Graduate School of Arts & Sci., Cambridge, MA

Postdoctoral Training

06/93-06/98	Research Assistant	Biological Chemistry and Molecular Pharmacology	Harvard Medical School
06/99-06/01	Resident	Internal Medicine	Massachusetts General Hospital, Boston, MA
07/99-11/02	Clinical Fellow	Medicine	Harvard Medical School
07/01-06/05	Clinical Fellow	Medical Oncology	Dana-Farber Cancer Institute, Boston, MA
07/01-06/05	Clinical Fellow	Medicine	Brigham and Women's Hospital Boston, MA
01/03-12/03	Chief Resident	Medicine	Massachusetts General Hospital

Faculty Academic Appointments

07/05-05/07	Instructor	Medicine	Harvard Medical School
06/07-	Assistant Professor	Medicine	Harvard Medical School

10/11			
11/11-	Associate Professor	Medicine	Harvard Medical School

Appointments at Hospitals/Affiliated Institutions

Current			
07/05-	Active Staff	Medical Oncology	Dana-Farber Cancer Institute
07/05-	Associate Physician	Medicine	Brigham and Women's Hospital
07/06-	Associate Member	Cancer Program	Broad Institute, Cambridge, MA
06/07-	Member	BBS Graduate Program (BCMP Dept.)	Harvard Medical School
10/10-	Senior Associate Member		Broad Institute, Cambridge, MA

Other Professional Positions

2007-	Consultant	Novartis Institutes for Biomedical Research, Cambridge, MA
07/08-07/09	Consultant (Genetic Technologies)	Clinical & Translational Research Center, Harvard Medical School
2009-	Scientific Task Force	LAM Treatment Alliance
2010-	Consultant	Foundation Medicine, Inc., Cambridge, MA
2011-	Member	Board of Scientific Advisors, Memorial Sloan-Kettering Cancer Center
2011-	Member	External Scientific Advisory Board, The Ohio State University Comprehensive Cancer Center
2013-	Member	External Scientific Advisory Board, MD Anderson Cancer Center

Major Administrative Leadership Positions

2011-	Co-Leader, Cancer Genetics Program	Dana-Farber/Harvard Cancer Center

Committee Service

- ***Local***

1998-99	M.D.-Ph.D. Program Review Committee	Member	Harvard Medical School
1999-2001, 2003	Teaching and Training Council	Member	Massachusetts General Hospital
2003	Training Program Committee	Member	Massachusetts General Hospital
2003	Length-of-stay Review Committee	Member	Massachusetts General Hospital
2003	Search Committee, Medical Firm Chief	Member	Massachusetts General Hospital
2003	Curriculum Committee, Medical Residency Program	Member	Massachusetts General Hospital

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Eliezer Van Allen, MD	**Instructor in Medicine**
eRA COMMONS USER NAME (credential, e.g., agency login) EMVANALLEN	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Stanford University, Stanford, CA	B.S.	06/03	Symbolic Systems
UCLA School of Medicine, Los Angeles, CA	M.D.	06/07	Medicine
UCSF, San Francisco, CA	Residency	06/10	Internal Medicine
Dana-Farber Cancer Institute, Boston, MA	Fellowship	07/10	Medical Oncology

A. Personal Statement

Dr. Van Allen is currently an Instructor in Medicine at Harvard Medical School, a clinician at Dana-Farber/Partners Cancer Care and a post-doctoral fellow in the lab of Dr. Levi Garraway. His research focuses on computational cancer genomics, the application of new technologies such as massively parallel sequencing to personalized cancer medicine, and resistance to targeted therapeutics.

B. Positions and Honors

Positions and Employment

2007 – 2010 Internship and Residency in Internal Medicine, UCSF, San Francisco, CA
2008 – 2010 Laboratory of Dr. Elad Ziv, UCSF, San Francisco, CA. Thesis: "*SMAD2* and biologically associated genes in the Latina breast cancer population: a case-control genome study"
2010 – 2013 Fellowship in Medical Oncology, Dana Farber/Partners Cancer Care, Boston, MA
2011 – Laboratory of Dr. Levi Garraway, Dana-Farber Cancer Institute, Boston, MA
2013 – Instructor in Medicine, Harvard Medical School

Other Experience and Professional Memberships

2008 – 2010 Member, PRIME (Program in Residency Investigation Methods and Epidemiology)
2010 – Member, Massachusetts Medical Society
2010 – Member, American Society of Clinical Oncology
2011 – Member, American Association for Cancer Research
2011 – 2012 Clinical Investigator Seminar at Dana-Farber Cancer Institute
2012 CEC Molecular and Translational Oncology Workshop
2012 – Clinical Sequencing Exploratory Research Working Groups
2013 AACR Molecular Biology in Clinical Oncology Workshop

Honors

2003 Phi Beta Kappa, Stanford University
2006 Alpha Omega Alpha Honor Medical Society, UCLA
2007 Award of Excellence of the Department of Medicine Clinical Faculty Association, UCLA
2009 Graduation Clinical Teaching Award, UCSF
2010 Reza Gandjei Humanism in Medicine Award, UCSF
2012 New England Journal of Medicine Gold Scholar
2012 Conquer Cancer Foundation Merit Award
2012 NIH Loan Repayment Program (NHGRI)
2013 Conquer Cancer Foundation Merit Award
2013 AACR-Millennium Prostate Cancer Fellowship (Awarded)

2013 ASCO Young Investigator
2013 Prostate Cancer Foundation Young Investigator

C. Selected Peer-reviewed Publications

1. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat J, White TA, Stojanov P, **Van Allen EM**, Stransky N, Nickerson E, Chae S, BoysenG, Auclair D, Onofrio R, Park K, Kitabayashi N, MacDonald TY, Sheikh K, Vuong T, Guiducci C, Cibulskis K, Sivachenko A, Carter SL, Saksena G, Voet D, Hussain WM, Ramos AH, Winckler W, Redman MC, Ardlie K, Mosquera JM, Rupp N, Wild PJ, Moch H, Morissey C, Nelson PS, Kantoff PW, Gabriel SB, Golub TR, Meyerson M, Lander ES, Getz G, Rubin MA, and Garraway LA. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nature Genetics 2012 May 20;44(6):685-9. PMCID: PMC3673022
2. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, Macdonald TY, Ghandi M, **Van Allen E**, Kryukov GV, Sboner A, Theurillat JP, Soong TD, Nickerson E, Auclair D, Tewari A, Beltran H, Onofrio RC, Boysen G, Guiducci C, Barbieri CE, Cibulskis K, Sivachenko A, Carter SL, Saksena G, Voet D, Ramos AH, Winckler W, Cipicchio M, Ardlie K, Kantoff PW, Berger MF, Gabriel SB, Golub TR, Meyerson M, Lander ES, Elemento O, Getz G, Demichelis F, Rubin MA, Garraway LA. Punctuated evolution of prostate cancer genomes. Cell. 2013 Apr 25; 153(3):666-77. PMCID: PMC3690918
3. **Van Allen EM**, Wagle N, Levy MA. Clinical analysis and interpretation of cancer genome data. J Clin Oncol. 2013 May 20; 31(15):1825-33. PMID: 23589549
4. **Van Allen EM**, Wagle N, Carter SC, Sucker A, Farlow D, Hodis E, Taylor-Weiner A, Berking C, Egberts F, Hassel JC, Gogas H, Gutzmer R, Goldinger SM, Loquai C, Urgurel S, Zimmer L, Gabriel SB, Getz G, Garraway LA, Schadendorf D. The genetic landscape of clinical resistance to RAF inhibition in melanoma. J Clin Oncol 31, 2013 (suppl; abstr 11009) [Oral Abstract Session, Tumor Biology, 2013 ASCO Annual Meeting].
5. **Van Allen EM**, Wagle N, Stojanov P, Perrin D, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich D, Kryukov G, Carter S, Rosenberg M, Fostel J, McKenna A, Sivachenko A, Kiezun A, Voet D, Lawrence M, Lichtenstein L, Gentry J, Huang F, Farlow D, Barbie D, Lander E, Gray S, Joffe S, Janne P, Garber J, MacConaill L, Lindeman N, Rollins B, Kantoff P, Fisher S, Gabriel S, Getz G, Garraway L. Whole-exome sequencing and clinical interpretation of FFPE tumor samples to guide precision cancer medicine. Nat Med. *In Press.*
6. Wagle N**, Van Allen EM**, Treacy D, Frederick DT, Cooper ZA, Taylor-Weiner A, Rosenberg M, Goetz EM, Sullivan RJ, Farlow DN, Friedrich D, Anderka K, Perrin D, Johannessen CM, McKenna A, Cibulskis K, Kryukov G, Hodis E, Lawrence DP, Fisher S, Getz G, Gabriel SB, Carter SL, Flaherty KT, Wargo JA, Garraway LA. MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. Cancer Discov. *In Press.*
7. **Van Allen EM**, Wagle N, Sucker A, Treacy D, Johannessen C, Goetz EM, Place CS, Taylor-Weiner A, Whittaker S, Kryukov G, Hodis E, Rosenberg M, McKenna A, Cibulskis K, Farlow D, Zimmer L, Hillen U, Gutzmer R, Goldinger SM, Ugurel S, Gogas HJ, Egberts F, Berking C, Trefzer U, Loquai C, Weide B, Hassel JC, Gabriel SB, Carter SL, Getz G, Garraway LA, Schadendorf D. The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. Cancer Discov. *In Press.*

D. Research Support

Prostate Cancer Foundation Young Investigator Award 07/01/2013-06/30/2016
Principle Investigator: Eliezer Van Allen, M.D.
Dissecting clinical resistance to abiraterone acetate in prostate cancer
Role: Instructor

American Cancer Society Post-doctoral Fellowship 07/01/2013-06/30/2016
Principle Investigator: Eliezer Van Allen, M.D.
Dissecting clinical resistance to PI3 kinase inhibitors
Role: Instructor

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Nikhil Wagle, MD	**Instructor in Medicine**
eRA COMMONS USER NAME (credential, e.g., agency login) NWAGLE11	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Harvard College, Cambridge, MA	A.B.	06/99	Biochemical Sciences
Harvard Medical School, Boston, MA	M.D.	06/05	Medicine
Brigham and Women's Hospital, Boston, MA	Residency	06/08	Internal Medicine
Dana-Farber Cancer Institute, Boston, MA	Fellowship	06/12	Medical Oncology

A. Personal Statement

Dr. Wagle is a medical oncologist at Dana-Farber Cancer Institute, an Instructor in Medicine at Harvard Medical School, and an Associate Member at the Broad Institute. His research focuses on clinical and translational cancer genomics, resistance to targeted therapeutics, and precision (or "personalized") cancer medicine. Dr. Wagle uses systematic genomic profiling approaches to comprehensively characterize tumor samples from patients with cancer in order to better understand the molecular determinants of tumorigenesis, characterize mechanisms of therapeutic response and resistance, and identify actionable genomic alterations to aid with clinical decision-making. In 2011, he published a novel approach for precision cancer medicine using targeted massively parallel sequencing for high-throughput mutation profiling of ~150 clinically relevant cancer genes from "real-world" archival tumor samples (Wagle et al, *Cancer Discovery*, 2011). Dr. Wagle is currently leading the effort to implement a prospective clinical sequencing program to guide clinical management of cancer patients, known as CanSeq, at Dana-Farber Cancer Institute, Brigham and Women's Hospital, and the Broad Institute. Dr. Wagle's research also focuses on characterization of patients with cancer who develop resistance to targeted therapies. He previously identified activating mutations in *MEK1*, the kinase downstream from BRAF, as a novel mechanism of resistance to the targeted therapy vemurafenib in metastatic melanoma (Wagle et al, J*ournal of Clinical Oncology*, 2011). He continues to study mechanisms of resistance in metastatic melanoma, as well as resistance to targeted therapies in breast cancer, lung cancer, and other solid tumors to identify novel ways to treat refractory advanced cancer. In addition, Dr. Wagle has recently focused his efforts on understanding the genomic mechanisms of extraordinary responses to cancer therapies, studying patients with exquisite sensitivity and/or unexpected durable responses to targeted agents. Overall, his research has made important contributions to the field of precision cancer medicine and translational cancer genomics.

B. Positions and Honors

Positions and Employment

1999 – 2000 RAND Health Program, RAND, Santa Monica, CA
2001 – 2003 Laboratory of Dr. Anindya Dutta, Brigham and Women's Hospital / Harvard Medical School, Boston, MA. Thesis: "The Role of DNA Replication Protein Cdc6 in DNA Damage Checkpoint Activation"
2005 – 2008 Internship and Residency in Internal Medicine, Brigham and Women's Hospital, Boston, MA
2008 – 2012 Fellowship in Medical Oncology, Dana Farber/Partners CancerCare, Boston, MA.
2010 – 2011 Chief Resident, Department of Medicine, Brigham and Women's Hospital, Boston, MA
2009 – Laboratory of Dr. Levi Garraway, Dana-Farber Cancer Institute, Boston, MA.
2012 – Instructor in Medicine, Harvard Medical School and Dana-Farber Cancer Institute, Boston, MA
2013 – Associate Member, Broad Institute, Cambridge, MA

Other Experience and Professional Memberships

2000 – Member, Massachusetts Medical Society
2009 – Member, American Society of Clinical Oncology
2009 – Member, American Association for Cancer Research

Honors

1998 Harvard College Research Program Grant
1999 Thomas Temple Hoopes Prize for excellence in undergraduate research
1999 *Magna Cum Laude* with Highest Honors in Biochemical Sciences, Harvard University
2002 – 2003 Howard Hughes Medical Institute Research Training Fellowship
2003 – 2005 Howard Hughes Medical Institute Fellowship for Completion of Medical Studies
2005 *Cum Laude* in a Special Field, Harvard Medical School
2008 Resident Mentor Award, Brigham and Women's Hospital
2008 Golden Stethoscope Award, Harvard Medical School
2010 Best Resident/Fellow Teacher Award; Harvard Medical School Class of 2010
2011 Partners in Excellence Award (for AMI Redesign Team), Partners Healthcare
2012 ASCO-Conquer Cancer Foundation Young Investigator Award
2013 Landon Foundation-AACR INNOVATOR Award for Research in Personalized Cancer Medicine
2013 Next Generation Fund of the Broad Institute of Harvard and MIT Award

C. Selected Peer-reviewed Publications (selected from 18**)**

1. Buschmann T, Minamoto T, **Wagle N**, Fuchs SY, Adler V, Mai M, and Ronai Z. Analysis of JNK, Mdm2, and $p14^{ARF}$ contribution to the regulation of mutant p53 stability. Journal of Molecular Biology**.** 2000 Jan 28;295(4):1009-21.
2. Goldman, DP, Schoenbaum M, Potosky AL, Weeks JC, Berry SH, Escarce JJ, Weidmer BA, Kilgore ML, **Wagle N,** Adams J, Figlin RA, Lewis JH, Cohen J, Kaplan R, and McCabe M. Measuring the incremental cost of clinical cancer research: The Cost of Cancer Treatment Study. Journal of Clinical Oncology. 2001 Jan 1;19(1):105-10.
3. **Wagle N**, Goldman DP, and Kilgore ML. Re: Surveys identify barriers to participation in clinical trials. Journal of the National Cancer Institute. 2001 Feb 7;93(3):238-9.
4. Jonsson ZO, Dhar SK, Narlikar GJ, Auty R, **Wagle N,** Pellman D, Pratt RE, Kingston R, and Dutta A. Rvb1p/Rvb2p: A new chromatin remodeling factor that regulates transcription of over 5% of yeast genes. Journal of Biological Chemistry. 2001 May 11;276(19):16279-88. Epub 2001 Feb 5.
5. Vaziri C, Saxena S, Jeon Y, Lee C, Murata K, Machida Y, **Wagle N**, Hwang DS, and Dutta A. A p53 dependent checkpoint pathway prevents re-replication. Molecular Cell. 2003 Apr;11(4):997-1008.
6. Punnoose LR, Roh JD, Hu S, Udell JA, **Wagle N**, Kirshenbaum JM, and LaCasce AS. Cardiac presentation of anaplastic large-cell lymphoma. Journal of Clinical Oncology. 2010 Jul 1;28(19):e314-6. Epub 2010 Jun 1 1;28(19):e314-6.
7. Ross JS, Torres-Mora J, **Wagle N**, Jennings TA, and Jones DM. Biomarker-based prediction of response to therapy for colorectal cancer: current perspective. American Journal of Clinical Pathology. 2010 Sep;134(3):478-90.
8. **Wagle N**, Emery C, Berger MF, Davis MJ, Sawyer A, Pochanard P, Kehoe S, Johannessen CM, MacConaill LE, Hahn WC, Meyerson M, and Garraway LA. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. Journal of Clinical Oncology. 2011 Aug 1;29(22):3085-96. Epub 2011 Mar 7.
9. **Wagle N**, Berger MF, Davis MJ, Blumenstiel B, DeFelice M, Pochanard P, Ducar M, Van Hummelen P, MacConaill LE, Hahn WC, Meyerson M, Gabriel SB, and Garraway LA. High-Throughput Detection of Actionable Genomic Alterations in Clinical Tumor Samples by Targeted, Massively Parallel Sequencing. Cancer Discovery. 2012 Jan 1;2(1):82-93. Epub 2011 Nov 7.

BIOGRAPHICAL SKETCH

NAME Paz Polak	POSITION TITLE
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Technion, Israel	B.Sc.	2000	Physics and Mathematics
Technion, Israel	M.Sc	2003	Applied Mathematics
Weizmann Institute of Science, Israel	M.Sc.	2006	Physics (computational biology)
Max Planck Institute For molecular Genetics/ Free university Berlin, Germany	Ph.D	2010	Computational Biology

A. Personal Statement

B. Positions and Honors

Positions:

2006-2011 Max Planck Institute for Molecular Genetics, Berlin, Germany. PhD student
2011- Brigham and Women's Hospital and Harvard Medical School. Postdoc

Honors:

1996-1999 Dean's excellence list B.Sc. Technion, Haifa, Israel
2000 President's excellence list. B.Sc. Technion, Haifa, Israel
2006 IMPRS-CBSC PhD fellowship by Max Planck Institute

C. Selected Peer-reviewed Publications

1. Polak P and Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. 2006. BMC Genomics 7 p. 133
2. Polak P and Arndt PF. Transcription induces strand-specific mutations at the 5' end of human genes,Genome Research. 2008. 18, 1216-1223.
3. Polak P and Arndt PF. Long Range bi-directional strand asymmetries originate at CpG islands in the human genome. Genome Biology and Evolution 2009, 189
4. Polak P, Querfurth R, Arndt PF. The evolution of transcription-associated biases of mutations across vertebrates. 2010. BMC Evolutionary Biology 10
5. Lawrence MS*, Stojanov P*, **Polak P*,** …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process
6. **Polak P***, Lawrence M.S.*, Haugen E, Stoletzki N, Stojanov P., Thurman R.E., Garraway L.A., Mirkin S., Getz G., John A Stamatoyannopoulos J.A., Sunyaev S. Reduced relative mutation density in regulatory DNA of cancer genomes linked to DNA repair. Nature Biotechnology (accepted)

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

APOBEC Mutagenesis in Human Cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Dr. Dmitry Gordenin, National Institute of Environmental Health Sciences (NIEHS, NIH), RTP, NC 27709, USA.; (Gordenin@niehs.nih.gov) Member of TCGA Analysis Network AWGs - BLCA, CESC, KICH, Pan-cancer.

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Dr. Steven Roberts, NIEHS, RTP, NC 27709, USA (Robertssa2@niehs.nih.gov)	**Dr. Shamil Sunyaev**, Harvard Medical School, Boston, MA 02115, USA (ssunyaev@rics.bwh.harvard.edu)

Background and preliminary data

We and others have demonstrated that the mutagenesis by APOBEC cytidine deaminases is widespread in several types of human cancers. In whole-genome mutation data sets from multiple myelomas, prostate, and head and neck cancers, we have revealed non-random distributions of mutations in the genome space – clusters of closely spaced strand coordinated base substitutions changing only cytosines (or only guanines) of the same strand (Roberts et al., *Mol. Cell*, 46:424,2012). [Similar clusters termed *kataegis* were found in a parallel study (Nik-Zainal et al. *Cell*, 149:1,2012.)] Clustered mutations carried the signature of a subclass of APOBEC cytidine deaminases (tCw→tTw or tCw→tGw) determined in independent mechanistic studies. Exquisite specificity of APOBECs for single-stranded (ss) over double-stranded (ds) DNA agreed with clustering and with C- or G-strand-coordination. Frequent co-localization of APOBEC-enriched clusters with rearrangement breakpoints also agreed with clusters stemming from ssDNA prone to breakage. Based on a combination of prior mechanistic data and the mutation signature derived from C- and G-coordinated clusters, we developed analytical steps for calculating enrichment with APOBEC signature mutagenesis and associated sample-specific q-values. (Roberts et al., *Nature Genet*. 45:970,2013). Since this mechanism-based hypothesis was sufficiently stringent, the analysis was applicable not only to individual samples with whole-genome but also for samples with whole-exome mutation calls. Analysis of 2680 TCGA exomes highlighted 6 out of 14 cancer types containing high prevalence of samples with APOBEC mutagenesis. [Note: this agreed well with a parallel publication by Alexandrov et al., *Nature* 500:415,2013, which identified mutation signatures by statistical extraction from large multi-type cancer data set.] In demonstration of the special utility of sample-specific q-values produced by our analysis, we highlighted one of the four subtypes of breast cancer (HER2 enriched) as more populated by APOBEC-mutated samples. As a result of our and of the parallel studies APOBEC mutagenesis emerged as one of the powerful sources of genome change in cancer. Pan-cancer project is the ideal setting to determine in what cancers APOBEC mutagenesis is over-represented and find potential reasons of such over-representation.

Timelines & resources dedicated to project

The starting point for our analysis with each cancer type would be the list of cancer-specific mutation calls (preferably in MAF format) for each sample and the list of rearrangement breakpoints. The outputs are: (i) MAF files in which every mutation is annotated by APOBEC mutation signature, inclusion in a cluster, co-localization with rearrangement breakpoint(s); and (ii) a set of summary tables with sample-specific mutation statistics centered on evaluating the APOBEC mutagenesis pattern including sample specific q-values. The estimated time from MAF and rearrangement inputs to outputs including graphics is up to three months. By this time material will be organized for easy interactive exploration of multiple hypotheses by our as well as by collaborating teams. The team dedicated to this interaction consists of the PI and a junior investigator, both experts in mutagenesis and genome instability as well as up to 6 bioinformatics professionals, experts in code development with experience in developing codes for this specific research through extensive collaboration with the PI. All team members already participated in work of BLCA, CESC and KICH TCGA AWGs. Collaborating Investigator, Dr. Sunyaev is an expert in analyzing mutational heterogeneity in cancer genomes.

Research proposal
Our analysis will produce: binary annotation of each mutation for APOBEC signature, location and types of mutation clusters, co-localization of each cluster with rearrangement breakpoint(s) and sample-specific q-values for the enrichment with APOBEC mutation signature over expected random mutagenesis. These will be used to explore the first tier of questions: (**i**) Which cancer types are enriched with samples showing APOBEC mutagenesis? (**ii**) Is there subtype specificity within cancer types enriched with APOBEC-mutated samples? (Known subtypes or subtypes highlighted in this project through mRNA expression clustering will be used.); (**iii**) Is there correlation between mutagenesis and APOBEC expression? (**iv**) Is there correlation between mutagenesis and expression of 5-15 genes chosen based on prior mechanistic knowledge; (**v**) Are there groups of genes with differential expression between samples with and without APOBEC mutagenesis? (**vi**) Does mutational heterogeneity within cancer genomes (replication timing, transcription, etc.) differ between APOBEC and non-APOBEC mutations in samples with high levels of APOBEC mutagenesis? (**vii**) Is there correlation of APOBEC mutation clusters with replication timing and transcription level? (More genomic features may be included into analysis based on interactive discussions with other groups). (**viii**) Does the frequency of co-localization of APOBEC mutation clusters with rearrangement breakpoints differ between cancer types? Outcomes from the first tier analyses will serve as a "discovery set" and will become a subject for validation using already existing exome data from TCGA as well as the 8000 exomes anticipated in this study. Second tier analysis will be based on the results of initial exploration and on alliances with other research groups based on their results and interests. For example, if mutational heterogeneity profile for APOBEC mutations in the APOBEC-enriched samples would differ from other mutations, it may be of interest to identify a subset of "significantly mutated genes" based on separate analysis of mutations with APOBEC signature. Other potential explorations could involve correlation of APOBEC mutagenesis with germline SNPs, viral presence, retrotranspositions and new pseudogenes. Since APOBEC signature mutagenesis can be defined so stringently, it can be also used to dissect mutagenesis pathways and identify mutator effects that are obscured by the presence of APOBEC mutagenesis. This is especially true for cancer types where APOBEC mutagenesis is abundant, such as bladder and cervical cancers as well as subgroups of head and neck, lung and breast cancers. Third tier explorations, possibly going beyond this project, will utilize the outputs produced by our analysis and made available for cancer and genome instability researchers. We also anticipate that the software and algorithms developed in the study of APOBEC mutagenesis will be easily applicable for hypothesis-based exploration of other mutation signatures developed from a combination of mechanistic knowledge with the mutation signatures extracted from analysis of large cancer datasets (e.g., Alexandrov et al., 2013 *ibid* and/or this project).

Legacy plans
We have already developed two software components that are integrated in a pipeline workflow and will be refined in the course of the study: code for analysis of spatial mutation clustering and code for evaluation of enrichment with APOBEC mutagenesis signature in clusters and in WGS, including production of sample specific q-values. The final versions of these packages will be set to automatically produce graphical outputs included for publication as well as for quick evaluation of additional hypotheses. We shall provide an executable multi-module code for this data-processing pipeline, written mostly in R and documented sufficiently to enable replication by the third parties. Software will be distributed locally and/or by depositing in distributed resources such as github (https://github.com/). The packages will be designed to be easily modifiable for analysis of sample enrichment with other mutation signatures which emerge form large-scale pattern recognition studies or/and from mechanistic research.

Dmitry A. Gordenin, Ph.D.
National Institute of Environmental Health Sciences (NIEHS), NIH
Research Triangle Park, NC 27709, USA
gordenin@niehs.nih.gov

A. EDUCATION
St. Petersburg State University, USSR Ph.D. in Genetics, 1978

B. POSITIONS HELD
1981-1988 - St. Petersburg State University, St. Petersburg, Russia, Senior Research Fellow, Group Leader
1988-1997- St. Petersburg State University, St. Petersburg, Russia, Supervising Research Fellow/ Research Group Leader
1997-Present -Senior Associate Scientist National Institute of Environmental Health Sciences (NIH), Laboratory of Molecular Genetics

C. PROFESSIONAL AND EDITORIAL BOARDS
1996 - present Editorial Board of *Mutation Research, Fundamental and Molecular Mechanisms of Mutagenesis* section
2011 – present Next Generation Sequencing Committee, National Institute of Environmental Health Sciences

D. FUNDING - NIEHS (NIH) Intramural Research Program

E. TCGA Analysis Working Groups AFFILIATION
Bladder Cancer (BLCA); Cervical Cancer (CESC), Kidney Chromophobe Cancer (KICH), Pan-cancer.

F. SELECTED RELEVANT PUBLICATIONS

Yang Y., Sterling J., Storici F., Resnick M.A., and **Gordenin D.A.** Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast Saccharomyces cerevisiae. (2008). PLoS Genetics, 2008 v. 4 (11):e1000264

Burch L.H., Yang, Y, Sterling J.F., Roberts S.A, Chao F.G., Xu H., Zhang L., Walsh J, Resnick M.A., Mieczkowski P.A., **Gordenin D.A.** Damage-Induced Localized Hypermutability (2011). Cell Cycle, v. 10:1073-1085.

Roberts S.A., Sterling J., Thompson C., Harris S., Mav D., Shah R., Klimczak L.J., Kryukov G.V., Malc E., Mieczkowski P.A., Resnick M.A. and **Gordenin D.A.** (2012). Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol. Cell, v. 46:424-435

Kin Chan, Joan F. Sterling, Steven A. Roberts, Ashok S. Bhagwat, Michael A. Resnick, and **Dmitry A. Gordenin**. (2012). Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. PLoS Genetics, 8(12): e1003149. doi:10.1371/journal.pgen.1003149.

Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, **Dmitry A. Gordenin**, Shamil Sunyaev, Eric S. Lander, Gad Getz. (2013). Mutational heterogeneity in cancer and the search for new cancer genes. Nature, v. 499: 214-218

Steven A. Roberts, Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V. Kryukov, Scott L. Carter, Gordon Saksena, Shawn Harris, Ruchir R. Shah, Michael A. Resnick, Gad Getz, and **Dmitry A. Gordenin**. (2013). An APOBEC Cytidine Deaminase Mutagenesis Pattern is Widespread in Human Cancers. Nature Genetics v. 45:970-976

Chan, Kin., Michael A. Resnick, and **Dmitry A. Gordenin**. (2013). The choice of nucleotide inserted opposite random abasic sites formed within chromosomal DNA is indicative of the polymerases participating in translesion DNA synthesis. DNA Repair. v. 12:878– 889

The Cancer Genome Atlas Research Network. (2013). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature, accepted in principle.

Steven A. Roberts, Ph.D.
National Institute of Environmental Health Sciences (NIEHS), NIH
Research Triangle Park, NC 27709, USA
robertssa2@niehs.nih.gov

A. EDUCATION

University of North Carolina, Chapel Hill, NC. Ph.D. in Biochemistry and Biophysics, 2008
Bowling Green State University, Bowling Green, OH. BS in Chemistry/Biochemistry and in Biology, 2003

B. POSITIONS HELD

2008-2009: University of North Carolina, Chapel Hill, NC. Post-doctoral researcher. Laboratory of Dale A. Ramsden, Ph.D.
2009-present: National Institute of Environmental Health Sciences, Research Triangle Park, NC. Mentor: Dmitry A. Gordenin, Ph.D.

C. GRANTS

NIH Pathway to Independence Award (K99/R00), June 4, 2013. Funding Institute: NIEHS

D. AWARDS

NIEHS Fellow of the Year	2013
Best Poster Presentation, NIEHS Science Day	2013
NIEHS Paper of the Year (Roberts, *et al. Mol Cell*, 2012)	2012
NIH Fellows Award for Research Excellence	2012
Poster Award, Toxicogenomics Integrated with Environmental Sciences conference	2011
Best Poster Presentation, Gordon Research Conference in Genetic Toxicology	2011
Best Oral Presentation, Genetics and Environmental Mutagenesis Society Fall Meeting	2010

E. RELEVANT PUBLICATIONS

Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Saksena, G., Harris, S., Shah, R.R., Resnick, M.A., Getz, G., and Gordenin, D.A. (2013) An APOBEC Cytidine Deaminase Mutagenesis Pattern is Widespread in Human Cancers. Nature Genetics. 45(9):970-6. doi:10.1038/ng.2702. PMID: 23852170

Roberts, S.A., Sterling, J, Thompson, C, Harris, S, Mav, D, Shah, R, Klimczak, L.J., Kryukov, G.V., Malc, E, Mieczkowski, P.A., Resnick, M.A., and Gordenin, D.A. (2012) Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. Molecular Cell. 46(4):424-35. PMCID: PMC3361558

The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of urothelial carcinoma of the bladder. Nature. Accepted in principle Nov. 7, 2013.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., **Roberts, S.A.**, Kiezun, A., Hammerman, P.S., McKenna, A., Drier, Y., Zou, L., Ramos, A.H., Pugh, T.J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M.L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D.I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A.M., Lohr, J., Landau, D., Wu, C.J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S.A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S.B., Roberts, C.W.M., Biegel, J.A., Stegmaier, K., Bass, A.J., Garraway, L.A., Meyerson, M., Golub, T.R., Gordenin, D.A., Sunyaev, S., Lander, E.S., Getz, G. (2013) Mutational heterogeneity in cancer and the search for new cancer genes. Nature. 499(7457):214-8. PMID: 23770567

Chan, K., Sterling, J.F., **Roberts, S.A**., Bhagwat, A.S., Resnick, M.A., and Gordenin, D.A. (2012) Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. PLoS Genetics. 8(12):e1003149. PMID: 23271983

Burch L.H., Yang, Y., Sterling J.F., **Roberts S.A.**, Chao F.G., Xu H., Zhang L., Walsh J., Resnick M.A., Mieczkowski P.A., and Gordenin D.A., (2011) Damage-induced localized hypermutability. Cell Cycle. Apr 1;10(7):1073-85. PMCID: PMC3100884

Shamil Sunyaev. Ph.D
Harvard Medical School and Brigham & Women's Hospital, Boston, MA 02115
ssunyaev@rics.bwh.harvard.edu

A. EDUCATION
1998 Ph.D. - Moscow Institute of Physics and Technology, Moscow, Russia
B. POSITIONS and EMPLOYMENT
2002-2009 - Assistant Professor of Medicine, Harvard Medical School, Boston, MA
2002-2009 - Research Staff, Associate Geneticist, Division of Genetics, Brigham and Women's Hospital,
2003- Divisional Appointment as Member, Harvard-M.I.T. Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA
2009- Associate Member, Broad Institute of MIT and Harvard, Cambridge, MA.
2009- Research Staff, Geneticist, Division of Genetics, Brigham and Women's Hospital,
2009- Associate Professor of Medicine, Harvard Medical School, Boston, MA
C. PROFESSIONAL AND EDITORIAL BOARDS
2007-2010 - American Journal of Human Genetics Editorial Board Member
2005- present - Biology Direct Editorial Board Member
D. CURRENT FUNDING
NIH grants (PI or co-PI): RO1 DK095721-01, U01 HG006500-01, U01 DE017018-08, R01 GM078598-04, R01MH101244-01
E. SELECTED RELEVANT PUBLICATIONS
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, **Sunyaev SR.** Human mutation rate associated with DNA replication timing. Nat Genet. (2009), v. 41:393-395.

Kryukov GV, Shpunt A, Stamatoyannopoulos JA, **Sunyaev SR.** Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci U S A. (2009) v. 106:3871-3876.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, **Sunyaev SR.** A method and server for predicting damaging missense mutations. Nat Methods. (2010) v. 7:248-9.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, **Sunyaev SR.** Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. (2010) v. 86:832-838

Zuk O, Hechter E, **Sunyaev SR**, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012 Jan 24; 109 (4) :1193-1198.

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM, **Sunyaev SR.** Exome sequencing and the genetic basis of complex traits. Nat Genet. (2012) v. 44:623-630.

Leshchiner I, Alexa K, Kelsey P, Adzhubei I, Austin-Tse CA, Cooney JD, Anderson H, King MJ, Stottmann RW, Garnaas MK, Ha S, Drummond IA, Paw BH, North TE, Beier DR, Goessling W, **Sunyaev SR.** Mutation mapping and identification by whole-genome sequencing. Genome Res. (2012) v. 22:1541-1548.

Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, **Sunyaev S.** Sequencing studies in human genetics: design and interpretation. Nat. Rev. Genet. (2013) v.14:460-470.

Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, van Duijn CM, Slagboom PE, van Ommen GJ, Wijmenga C; Genome of the Netherlands Consortium, de Bakker PI, **Sunyaev SR.** Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. (2013) v. 9(2):e1003301.

Nusinow DP, Kiezun A, O'Connell DJ, Chick JM, Yue Y, Maas RL, Gygi SP, **Sunyaev SR.** Network-proteomic mixtures using SNIPE. Bioinformatics. (2012). v. 28:3115-3122.

Polak P, Lawrence MS , Haugen E, Stoletzki N, Stojanov P, Thurman RE, Garraway LA, Mirkin S, Getz G, Stamatoyannopoulos JA, **Sunyaev SR.** Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nature Biotech. (2013). Accepted.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

David Haussler, University of California Santa Cruz, TCGA GDAC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Jingchun Zhu, U. of California Santa Cruz	

Background and preliminary data

ICGC and TCGA holds promise for a comprehensive understanding of human cancer. These projects are producing the most uniform and comprehensive catalog of cancer-specific genomic aberrations to date. However to derive the most value from this large amount of data, there must be powerful, flexible tools to allow researchers and clinicians both inside and outside the consortia to efficiently access and view this data, as well as integrate data derived from their own research, thus allowing connections with genomic aberrations, cancer subtypes and clinical features such as survival and response to therapy. There are four major challenges: 1) retrieval of the most relevant subsets of data across the entire data collection based on specific biological queries on meta-data (clinical parameters, phenotypes, etc.) and genomic patterns. 2) visualize the returned data in a biological meaningful way 3) easy manipulation of queries and results, allowing the formation of iterative loops of refinement, questions and data 4) seamless integration and cross-analysis of ICGC/TCGA data with other seminal datasets such as *in vitro* drug perturbation as well as researcher's private data.

To begin to address these challenges, we developed the UCSC Cancer Genomics Browser (Cancer Browser). It provides interactive visualization and exploration of functional genomics, phenotypic, and clinical data. Researchers explore the impact of genomic alterations on phenotypes by visualizing gene and protein expression, copy number, DNA methylation, somatic mutation and pathway inference data alongside clinical features, Pan-Cancer subtype classifications and genomic biomarkers. Integrated Kaplan–Meier survival analysis helps investigators to assess survival stratification by any type of data. Summary views and online statistical analysis allow easy comparisons across subgroups. Custom clinical data allow users to integrate their own sample annotations into existing datasets. The UCSC Cancer Browser currently hosts an expanding set of searchable data, including 409 datasets from The Cancer Genome Atlas, as well as data from Cancer Cell Line Encyclopedia, Connectivity Map, Stand Up To Cancer and selected datasets from literature.

Timelines & resources dedicated to project

Timelines: We will display the core, higher-level calls derived using the raw sequencing data on the browser as the data is generated. Higher-level calls include VCF data provided by ICGC, accompanied expression, DNA methylation calls on gene, exon and probe level, and de-identified clinical data. In addition to the core dataset, users will be able to display their own genomic and clinical data alongside public data using the browser data server API and an instance of cancer browser built in the cloud. This will allow users to access to the new data search, view, and analysis capabilities as described in the research proposal.

Resources: We depend on high quality genomic data derived from raw sequencing data provided by ICGA and other groups in the consortium.

Research proposal

We will use UCSC Cancer Genomics Browser to provide data visualization for the Pan-cancer core datasets. The type of data we will host are higher-level calls derived from raw sequencing data, such as base substitutions and indels calls, copy number level estimates in VCF formats, and the accompanied gene-, exon- expression, and DNA methylation estimations, as well as clinical data. We expect to function as a major data visualization portal for the high-level core dataset.

We will leverage the cloud computing resource provided by the consortium to have Cancer Browser instances easily initiated. We expect researchers use their own instance to view and analyze their own data, together with the core data already accessible from the main browser. We have recently developed the browser data server API, which allows data from multiple sources (both users own data and the core data) can be displayed securely in a single web page.

In addition to our current capabilities, we will develop the Cancer Browser to allow slicing of the data using any combination of the following criteria: genomic data, genomic signatures, structured and unstructured clinical, phenotypic and meta-data information (such as pdf pathological reports). This functionality is particularly useful for large and expanding data such as those produced by the ICGC/TCGA project. The results will be displayed in heatmaps and summary views along three representations: genomic regions, collections of genes of user interest, and pathways. Interactive manipulation of results and queries will allow adding and removing any of the above data, which can be used to form a user-driven interactive process of data exploration.

Legacy plans

The UCSC Cancer Genomics Browser is accessible at https://genome-cancer.ucsc.edu. The browser development team has been developing and supporting it since 2009. The Browser source code and virtual machine images will be updated regularly and made freely available to biomedical researchers and educators in the non-profit sector, such as institutions of education, research institutions, and government laboratories.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Zhu, Jingchun	Research Scientist, University of California Santa Cruz

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Fudan University, China	B.S.	07/1995	Biochemistry
State University of New York at Stony Brook	M.S.	05/1998	Biology
University of California, San Francisco	Ph.D.	03/2006	Biological and Medical Informatics
University of California, Santa Cruz	Postdoctoral	10/2010	Genomics & Bioinformatics

Selected Peer-reviewed Publications

Cline MS, Craft B, Swatloski T, Goldman M, Ma S, Haussler D, **Zhu J**. Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser. Sci Rep 2013. In Press.

Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Joshua M. Stuart JM. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nature 2013. In Press.

Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M, Ma S, Wilks C, Stuart J, Haussler D, **Zhu J**. The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res. 2012 : gks1008v1-gks1008.*

Easwaran H., Johnstone S, VanNeste L., Ohm J., Mosbruger T., Wang Q., Aryee M., Joyce P., Ahuja N., Weisenberger D., Collisson E., **Zhu J**, Yegnasubramanian S., Matsui W., Baylin S. (2012) A DNA Hypermethylation Module for the Stem/Progenitor Cell Signature of Cancer. *Genome Reserch 22*(5), 837–849. doi:10.1101/gr.131169.111

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A., Koebbe, B. C., Nielsen C., Hirst M., Farnham P., Kuhn R., **Zhu J.**, Smirnov I., Kent W.J., Haussler D., Madden P., Costello J., Wang T. (2011). The Human Epigenome Browser at Washington University. *Nature methods*. 8(12), 989–990

Raab, J. R., Chiu, J., **Zhu, J.,** Katzman, S., Kurukuti, S., Wade, P. A., Haussler, D., et al. (2011). Human tRNA genes function as chromatin insulators. *The EMBO journal*. 31(2), 330–350.

Sanborn J.Z., Benz S., Craft B., Szeto C., Kober K., Meyer L., Vaske C., Goldman M., Smith K., Kuhn R., Karolchik D., Kent W.J., Stuart J., Haussler D. and **Zhu J.** The UCSC cancer genomics browser: update 2011. (2011) *Nucleic Acids Res.* 39(suppl 1):D951-D959.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, **Zhu J**, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. (2010) *Bioinformatics* 26(12): i237-45.

Zhu J, Sanborn J.Z, Benz S, Hsu F, Szeto C, Kuhn R, Karolchik D, Archie J, Lenburg M.E, Esserman L.J, Kent W.J, Haussler D, and Wang T. The UCSC Cancer Genomics Browser. (2009) *Nature Methods* 6(4): 239-40.

Zhu J, Sanborn J. Z, Diekhans M, Lowe CB, Pringle TH, Haussler D. Comparative genomics search for losses of long-established genes on the human lineage. (2007) PloS Computational Biology 3(12):e247.

Zhu J, Jambhekar A., Sarver A. and DeRisi JL. A Bayesian network driven approach to model the transcriptional response to nitric oxide in *Saccharomyces cerevisiae*. (2006) *Plos ONE* 1:e94 (the inaugural issue).

Bozdech Z, Llinas M, Pulliam B, Wong E, **Zhu J** and DeRisi JL. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. (2003) *PLoS Biol. Oct;1(1):E5*. (the inaugural issue)**.**

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identification and characterization of amplification-associated rearrangements and gene fusions across cancer types

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

David Haussler, TCGA UCSC-Buck Institute GDAC

Sofie Salama, TCGA GBM AWG, TCGA LGG AWG

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Mia Grifford, UCSC **Olena Morozova, UCSC**	

Background and preliminary data

Cancer is a genetic disease and gene fusion transcripts represent an important mechanism for generating critical driver mutations. A well-known example is the BCR-ABL fusion in chronic myelogenous leukemia that produces an oncogenic kinase sufficient to drive the malignancy. Gene fusions represent attractive therapeutic targets as they allow for the specific targeting of cancer cells. Recently we have identified rearrangement-based receptor tyrosine kinase fusions in glioblastoma mutliforme (GBM), and IDH wildtype lower grade gliomas (IDHwt LGGs),(Brennan, et al. *Cell.* 2013. 155:462-77, TCGA LGG AWG, in preparation). We have also found that in about one quarter of GBM and IDHwt LGG patients, rearrangements are associated with double minute chromosomes or homogenously staining regions (DMs/HSRs) (Sanborn, et al. *Cancer Res.* 2013. 73:6036-45, TCGA LGG AWG, in preparation). These DMs/HSRs contain one or more oncogenes and become highly amplified, possibly driving the progression of the cancer. In other patients, rearrangements involving receptor tyrosine kinases are associated with copy number neutral events or focal amplifications.

Until now it has been difficult to explore the molecular mechanisms of genome rearrangements that underlie fusion transcripts because whole genome sequencing data was not available for most tumor samples with RNA sequencing data. The WGS-Pan cancer project gives us a unique opportunity to combine and integrate fusion transcript and DNA rearrangement analyses to determine the features of the DNA rearrangement breakpoints associated with individual fusion transcripts. In addition, by having data representing a diversity of tumor types we can explore the effect of cell of origin on both the genes involved in fusion transcripts and the types of rearrangements leading to fusion transcripts. The overall goal is to determine whether specific rearrangement mechanisms are associated with particular oncogenic driver mutations and/or particular tumor types.

Timelines & resources dedicated to project

We will use the whole genome, whole exome, and RNA sequencing data generated by the TCGA and ICGC projects. In the first phase (Spring, 2014), BamBam, nFuse and deFuse, will be used to generate lists of high quality DNA rearrangements, DNA copy number estimates, and RNA fusions. Next (Summer, 2014), we will focus on associating fusion events with specific rearrangements as well as identifying and reconstructing DMs/HSRs. We will also determine the clonality and order of rearrangements using our recently developed CN-AVG tool. Finally (Fall 2014, Winter 2015), we will integrate the DNA and RNA analyses to determine the genomic basis of the gene fusions. This analysis will focus on understanding the relationship between copy number changes and molecular features associated with the underlying rearrangements across recurrent gene fusion partners and tumor types. Most work will be performed by graduate student, Mia Grifford (DNA analysis) and postdoctoral fellow, Olena Morozova (RNA analysis) with support from the Cancer Genomics Research Group (Haussler and Stuart labs at UCSC) as well as collaborator Dr. J. Zachary Sanborn at Five3 Genomics. Computing resources necessary for this work are available in the Haussler lab.

Research proposal

We plan to use the pan cancer whole genome sequencing (WGS) data, the RNA sequencing data (RNA-Seq), and the whole exome sequencing (WES) data to catalogue the type and frequency of rearrangement-based fusions in each cancer. The RNA-Seq data allows us to identify situations where a DNA rearrangement results in an expressed product, whereas the whole genome sequencing data helps us determine the mechanism underlying the fusion.

We will use the WES data to determine the frequency and characteristics of circular amplicons, which correspond to double minute chromosomes and/or homogenously staining regions in multiple cancer types. We will use WGS data, whenever they are available, to reconstruct in detail potential double minute chromosomes we identify. The structure of double minute chromosomes is typically inferred by combining multiple sources of data, including sets of breakends, copy-number variations and expression data. This inference has significant uncertainty, particularly in predicting the overall layout of the amplicon. We will apply novel sampling methodology to the inference process to quantify our confidence in predictions, and explore the evolutionary history of these products. This will also allow us to estimate the prevalence of these amplicons in tumor sub-clones, and determine at what point in tumorigenesis they were created.

We will run the deFuse algorithm (McPherson, et al, PLoS Comput Biol. 2011. 7:e1001138) on the RNA sequencing data to identify gene fusions. We will then run filtering algorithms based on features of validated rearrangements and exon-level expression data to get a list of low quality and high quality gene fusions. This will tell us how often gene fusions occur in multiple cancer types and what genes are involved. We will also learn if there are any recurrent gene fusions across cancer types.

In collaboration with Five3 Genomics, we will run a tool we developed called BamBam on the WGS and WES data to find DNA rearrangements and estimate DNA copy numbers (Sanborn, *Cancer Res*, 2013). The rearrangements will be filtered based on mapping quality and read support, resulting in both low-quality and high-quality lists of rearrangements. The BamBam rearrangements and copy number calls from the WES data will be used to find circular amplicons, while the BamBam results from the WGS data will be used to reconstruct such amplicons. We will then be able to determine the frequency and characteristics of DMs/HSRs in each cancer type.

As there is uncertainty in double minute chromosome prediction, the CN-AVG tool, which we have recently developed, will be used to refine and expand our analysis. CN-AVG takes as input the raw variations predicted by BamBam, in the form of copy-number and break end calls. It outputs, at the lowest level, evolutionary histories that explain the input data, in the form of an order of specific rearrangement events. Integrating across multiple sampled histories (typically tens of thousands), we will quantify our confidence in the predicted structure of DMs, and estimate the timing of the emergence of these amplicons.

We will also integrate the DNA and RNA analyses to reconstruct the genomic basis for the observed gene fusions. This will be done by two methods. First we will search for breakpoints in the BamBam output that support transcript fusions predicted by deFuse. Second, we will run nFuse (McPherson, et al. Genome Res. 2012. 22:2250-61), a method that combines WGS and RNA-Seq data to precisely reconstruct fusion transcripts. These data will allow us to determine the features associated with breakpoints underlying oncogenic transcript fusion events. Understanding the rearrangement mechanisms associated with specific genes fusions and tumor types will be an important advance in our understanding of the etiology of cancer.

Legacy plans

All of the software and algorithms are either already published or are in preparation and will be made available to the public.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Haussler, David	Investigator, Howard Hughes Medical Institute Professor, Biomolecular Engineering, University of California, Santa Cruz
eRA COMMONS USER NAME (credential, e.g., agency login) HAUSSL	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Connecticut College, New London, CT	BA	1975	Mathematics
California Polytechnic State University, San Luis Obispo	MS	1979	Applied Mathematics
University of Colorado, Boulder	PhD	1982	Computer Science

Positions and Employment

1982-1986 Assistant Professor, Mathematics & Computer Science, University of Denver, CO
1986-2004 Assistant Professor to Professor, Computer Science, University of California, Santa Cruz, CA
2000- Investigator, Howard Hughes Medical Institute, University of California, Santa Cruz, CA
2004- Distinguished Professor, Biomolecular Engineering, University of California, Santa Cruz, CA

Selected Peer-reviewed Publications

Daniel R. Zerbino, Benedict Paten, Glenn Hickey, David Haussler. An algebraic framework to sample the rearrangement histories of a cancer metagenome with double cut and join, duplication and deletion events. arXiv:1303.5569v1 [q-bio.GN] 22 Mar 2013.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments. Bioinformatics. 2013 Mar 16. PMID: 23505295

Cancer Genome Atlas Research Network, Hammerman PS, Hayes DN, Wilkerson MD, Schultz N, Bose R, Chu A, Collisson EA, Cope L, Creighton CJ, Getz G, Herman JG, Johnson BE, Kucherlapati R, Ladanyi M, Maher CA, Robertson G, Sander C, Shen R, Sinha R, Sivachenko A, Thomas RK, Travis WD, Tsao MS, Weinstein JN, Wigle DA, Baylin SB, Govindan R, Meyerson M. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep 27;489(7417):519-25. doi: 10.1038/nature11404. PMID: 22960745

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Sep 23;490(7418):61-70. doi: 10.1038/nature11412. PMID: 23000897

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. PMID: 22955616; PMCID: PMC3439153

Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F, Bayani N, Hu Z, Billig JI, Dueregger A, Lewis S, Jakkula L, Korkola JE, Durinck S, Pepin F, Guan Y, Purdom E, Neuvial P, Bengtsson H, Wood KW, Smith PG, Vassilev LT, Hennessy BT, Greshock J, Bachman KE, Hardwicke MA, Park JW, Marton LJ, Wolf DM, Collisson EA, Neve RM, Mills GB, Speed TP, Feiler HS, Wooster RF, Haussler D, Stuart JM, Gray JW, Spellman PT. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc Natl Acad Sci U S A. 2012 Feb 21;109(8):2724-9. PMID: 22003129

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010 Jun 15;26(12):i237-45. PMID: 20529912. PMCID: PMC2881367

Katzman S, Kern AD, Pollard KS, Salama SR, Haussler D. GC-Biased Evolution Near Human Accelerated Regions. PLoS Genet. 2010 May 20;6(5):e1000960. PMID: 20502635

Zhu J, Sanborn JZ, Benz S, Szeto C, Hsu F, Kuhn RM, Karolchik D, Archie J, Lenburg ME, Esserman LJ, Kent WJ, Haussler D, Wang T. The UCSC cancer genomics browser. Nature Methods. 2009 Apr;6(4):239-40. PMID: 19333237.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2. Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Sofie Reda Salama	UCSC Research Associate HHMI Research Specialist
eRA COMMONS USER NAME Salama	

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
University of Illinois at Urbana-Champaign	BS	1989	Chemistry
University of California, Berkeley	PhD	1995	Molecular and Cell Biology
Mass. General Hospital/Harvard Medical School	Postdoc	1998	Molecular Oncology

Positions and Employment

1995-1998 Postdoctoral Fellow, Laboratory of Molecular Oncology, Mass. General Hospital, Charlestown, MA. Mentor: Dr. Edward Harlow

1998-2001 Senior Scientist, Microbia, Inc., Cambridge, MA

2001-2002 Director of Core Technology, Microbia, Inc. Cambridge, MA

2002-2004 Visiting Researcher, Department of Environmental Toxicology, UC Santa Cruz, Santa Cruz, CA. Mentor: Dr. Fitnat Yildiz

2004-present Research Specialist: Center for Biomolecular Science and Engineering and Howard Hughes Medical Institute, University of California at Santa, Santa Cruz, CA

2008-present Research Associate in Biomolecular Engineering, University of California at Santa Cruz, Santa Cruz, CA

Selected Peer-reviewed Publications

1. Classon M, **Salama S**, Gorka C, Mulloy R, Braun P, Harlow E. 2000. Combinatorial roles for pRB, p107, and p130 in E2F-mediated cell cycle control. Proc Natl Acad Sci. **97**(20):10820-5.
2. Fingar DC, **Salama S**, Tsou C, Harlow E, Blenis J. 2002. Mammalian cell size is controlled by mTOR and its downstream targets S6K1 and 4EBP1/eIF4E. Genes Dev. **16**(12):1472-87.
3. Bejerano G, Lowe C, Ahituv N, King B, Siepel A, **Salama SR**, Rubin EM, Kent, JW, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. **441**(7089):87-90.
4. Pollard KS, **Salama SR**, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M Jr, Vanderhaeghen P, Haussler D. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. **443**(7108):167-72.
5. Pollard KS, **Salama SR**, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, Rosenbloom KR, Kent J, Haussler D. 2006. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. **2**(10):e168.
6. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, **Salama SR**, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. Science. **317**(5840):915.
7. Katzman S, Kern AD, Pollard KS, **Salama SR** and Haussler D. 2010. GC-biased evolution near human accelerated regions. PLoS Genet. 6(5):e1000960.
8. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, **Salama SR**, Haussler D. 2010. FragSeq: transcriptome-wide RNA structure probing using thigh-throughput sequencing. Nat. Methods. 7(12):995-1001.
9. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, **Salama SR**, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. Science. 19;333(6045):1019-24.
10. Sanborn JZ, **Salama SR**, Grifford M, Brennan CW, Mikkelsen T, Jhanwar S, Katzman S, Chin L, Haussler D. Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. Cancer Res. 2013 Oct 1;73(19):6036-45.

    Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, **Salama SR**, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L; TCGA Research Network. The somatic genomic landscape of glioblastoma. Cell. 2013 Oct 10;155(2):462-77.

MIA S GRIFFORD

EMAIL:MGRIFFORD@SOE.UCSC.EDU

211 21ST AVE • SANTA CRUZ CA 95062 • PHONE: (302) 562.0511

EDUCATION

2009-Current University of California, Santa Cruz Santa Cruz, CA

- Ph.D. Biomolecular Engineering and Bioinformatics

2009-2011 University of California, Santa Cruz Santa Cruz, CA

- M.S. Biomolecular Engineering and Bioinformatics

2005-2009 University of Delaware Newark, DE

- B.S. Computer Science, Concentration in Bioinformatics, GPA 3.98

2001-2005 Delaware Technical & Community College Wilmington, DE

- A.A.S Computer Information Systems, Graduated Summa Cum Laude, GPA 4.0
- A.A.S Computer Network Engineering, Graduated Summa Cum Laude, GPA 4.0

RESEARCH EXPERIENCE

Apr 2010-Current Dr. David Haussler University of California, Santa Cruz

- Using high-throughput sequencing to study the progression of cancer

Jan 2010-Mar 2010 Dr. Camilla Forsberg University of California, Santa Cruz

- The role of Flk2 in hematopoiesis

Sept 2009-Dec 2009 Dr. Joshua M. Stuart University of California, Santa Cruz

- Predicting clinical outcomes using gene module information

Feb 2008-May 2009 Dr. Prasad Dhurjati University of Delaware

- Personalized pharmacological effects of drugs: combining biological and chemical data to predict side effects

Oct 2008-Dec 2008 Dr. Li Liao University of Delaware

- Predicting drug-target interactions using chemogenomic approaches and support vector machines

May 2007-Sept 2007 Dr. Li Liao University of Delaware

- Inferring functional relationships from co-evolutionary information

SELECTED PEER-REVIEWED PUBLICATIONS

- Sanborn JZ, Salama SR, **Grifford M**, Brennan CW, Mikkelsen T, Jhanwar S, Katzman S, Chin L, Haussler D. Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. Cancer Res. 2013 Oct 1;73(19):6036-45.
- Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L; TCGA Research Network. The somatic genomic landscape of glioblastoma. Cell. 2013 Oct 10;155(2):462-77.
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. Nature. 2012 Oct 4;490(7418):61-71.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2. Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Olena Morozova	California Institute for Regenerative Medicine Postdoctoral Scholar
eRA COMMONS USER NAME	

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
University of Toronto, Toronto, ON, Canada	BS (Hons)	2006	Molecular Genetics and Biology
University of British Columbia, Vancouver, BC, Canada	PhD	2012	Bioinformatics
University of California Santa Cruz	Postdoc	Present	Cancer Genomics, Stem Cell Biology

Positions and Employment

2013-present Postdoctoral Scholar, California Institute for Regenerative Medicine, University of California Santa Cruz. Mentor: Dr. David Haussler

2013 Postdoctoral Scholar, Howard Hughes Medical Institute, University of California Santa Cruz. Mentor: Dr. David Haussler

Selected Peer-reviewed Publications

1. **Morozova O** and Marra MA. (2008) From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors. *Biochemistry and Cell Biology*, **86**(2):81-91.
2. **Morozova O,** Morozov V, Hoffman B, Helgason C, and Marra MA. (2008) A seriation approach for visualization-driven discovery of co-expression patterns in Serial Analysis of Gene Expression (SAGE) data. *PLoS ONE* **3**(9):e3205.
3. **Morozova O** and Marra MA. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**(5):255-64.
4. **Morozova O,** Hirst M and Marra MA. (2009) Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics* **10**:135-51.
5. **Morozova O**, Vojvodic M, Grinshtein N, Hansford LM, Blakely KM, Maslova A, Hirst M, Cezard T, Morin RD, Moore R, Smith KM, Miller F, Taylor P, Thiessen N, Varhol R, Zhao Y, Jones S, Moffat J, Kislinger T, Moran MF, Kaplan DR, Marra MA. (2010) Systems-level analysis of tumor-initiating cells implicates AURKB as a novel drug target for neuroblastoma. *Clinical Cancer Research* **16**(18):4572-82.
6. Yip S, Butterfield YS, **Morozova O**, Chittaranjan S, Blough MD, An J, Birol I, Chesnelong C, Chiu R, Chuah E, Corbett R, Docking R, Firme M, Hirst M, Jackman S, Karsan A, Li H, Louis DN, Maslova A, Moore R, Moradian A, Mungall KL, Perizzolo M, Qian J, Roldan G, Smith EE, Tamura-Wells J, Thiessen N, Varhol R, Weiss S, Wu W, Young S, Zhao Y, Mungall AJ, Jones SJM, Morin GB, Chan JA, Cairncross JG, Marra MA. (2012) Concurrent *CIC* mutations, *IDH* mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers. *Journal of Pathology* 226(1):7-16.
7. Pugh TJ*, **Morozova O***, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K, Hanna M, Kiezun A, Kim J, Lawrence MS, Lichenstein L, McKenna A, Pedamallu CS, Ramos AH, Shefler E, Sivachenko A, Sougnez C, Stewart C, Ally A, Birol I, Chiu R, Corbett RD, Hirst M, Jackman SD, Kamoh B, Khodabakshi AH, Krzywinski M, Lo A, Moore RA, Mungall KL, Qian J, Tam A, Thiessen N, Zhao Y, Cole KA, Diamond M, Diskin SJ, Mosse YP, Wood AC, Ji L, Sposto R, Badgett T, London WB, Moyer Y, Gastier-Foster JM, Smith MA, Guidry Auvil JM, Gerhard DS, Hogarty MD, Jones SJM, Lander ES, Gabriel SB, Getz G, Seeger RC, Khan J, Marra MA, Meyerson M and Maris JM. *Authors contributed equally. (2013) The genetic landscape of high-risk neuroblastoma. *Nature Genetics* **45**(3):279-84.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identification of Somatic Mutations and RNA-Editing Events in Cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

David Haussler and Jingchun Zhu

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Amie Radenbaugh

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Much of our current understanding of cancer has come from investigating how normal cells are transformed into cancerous cells through the acquisition of somatic mutations. In order to identify genetic alterations that are essential for cancer development and progression, researchers have focused on determining the genetic differences between tumor cells and normal cells in the same individuals. Just a few years ago, projects such as The Cancer Genome Atlas (TCGA) began providing high-throughput sequencing data for both tumor and matched-normal genomic DNA for multiple types of cancers. Many groups participating in TCGA analysis, including our own, evaluated the two DNA datasets to predict germline variants and somatic mutations. With the development of new technologies such as RNA-Seq, projects like TCGA began providing an additional high-throughput sequencing dataset for the tumor RNA. These three datasets consisting of the tumor DNA, matched-normal DNA, and tumor RNA have become the new standard in cancer genomics. It is now possible to investigate the consequences of genomic changes in the actual RNA transcripts and to better characterize 1) germline variants, 2) somatic mutations, and 3) variants in the RNA that are not found in the DNA that could be the result of RNA-Editing. We have developed a method called RADIA (RNA and DNA Integrated Analysis) to identify and characterize alterations to DNA and RNA in cancer using high-throughput sequencing data. Somatic mutation calls from RADIA have been submitted for eight different cancer types from TCGA. The WGS Pan-cancer dataset would be an ideal dataset for further somatic mutation calling and for investigating the role of RNA-Editing in cancer.

Timelines & resources dedicated to project

The key intermediate milestones for this analysis would be to run RADIA on as many normal DNA, tumor DNA, and tumor RNA triplets as possible. The identification of RNA-Editing events would limit the analysis to those patients where RNA-Seq data is available.

Research proposal
Traditionally, somatic mutation calling has been done on tumor and matched-normal DNA pairs. The ability to accurately detect somatic mutations is hindered by both biological and technical artifacts. In addition, it is difficult to obtain both high sensitivity and high specificity. Different algorithms predicting somatic mutations on the same data often have discernible differences due to the trade-off between sensitivity and specificity. This is especially true for somatic mutations with low variant allele frequencies. By creating a method that utilizes both DNA and RNA, we have increased our power at detecting somatic mutations, especially at low variant allele frequencies. **Over the past decade, RNA-editing events have been identified in a variety of cancer tumors: brain, kidney, prostate, lung, breast, and AML to name a few. The RNA-editing events have been shown to be significant to the development and progression of cancer. Projects that provide high-throughput sequencing datasets from both the DNA and the RNA from the same patients, make it possible to search for RNA-Editing events at a genome-wide scale.** **Using the tumor and matched-normal genomes and the RNA-Seq data, we will identify RNA-Editing events across multiple cancer types. All putative RNA-Editing events will be assessed according to the most common types of RNA-Editing, such as the deamination of adenosine into inosine (A-to-I) or the conversion of cytosine into uracil (C-to-U).** **As a positive control, we will confirm RNA-editing events previously discovered experimentally from the literature, such as an A-to-I conversion in the protein tyrosine phosphatase PTPN6 gene in AML patients. The PTPN6 gene is recognized as a tumor suppressor gene and is important for the down-regulation of growth-promoting receptors. The A-to-I conversion causes the splicing mechanism to ignore a splicing junction, leading to a non-functional PTPN6 protein via the inclusion of an intron in the mature RNA transcript. Using RNA-Seq data, we will identify such RNA-Editing events and analyze the functional impact of the event. In addition, we will report novel RNA-Editing events across multiple cancer types and look for patterns that may be cancer specific or globally relevant to cancer development.**
Legacy plans
We will make our code available to the research community and visualize our results in the UCSC Cancer Genomics Browser.

AMIE RADENBAUGH

606 Almaden Walk Loop
San Jose, CA 95125

(781) 354-4286
Amie.Radenbaugh@gmail.com

BIOINFORMATICIAN

PhD candidate in Bioinformatics in David Haussler's lab at the University of California, Santa Cruz working on the identification and characterization of alterations to DNA and RNA in cancer using high-throughput sequencing data. The sequencing of tumor and matched-normal genomes and tumor RNA-Seq data has become the new standard in cancer genomics projects such as The Cancer Genome Atlas (TCGA). We have created a pipeline that uses both the tumor and matched-normal DNA and the tumor RNA to detect germline variants, Loss Of Heterozygosity (LOH) events, somatic mutations and RNA-Editing events. We have submitted our calls to numerous TCGA Analysis Working Groups.

TECHNICAL SKILLS

LANGUAGES:

- **EXPERT:** Python, R, Java, J2EE (JSP, Servlets, EJBs), J2SE (JFC/Swing, RMI), Perl, JDBC, ODBC, ANT, JUnit, PL/SQL, SQL, HTML, XML, XSLT, PHP, CGI, C++, C, Cold Fusion, ILOG JRules, Scheme, LISP, Prolog, SmallTalk
- **AVERAGE:** ABAP, BAPIs, BADIs, RFCs, AJAX, JavaScript, CSS

METHODOLOGIES: UML, RUP, Design Patterns, Extreme Programming
DATABASES: Oracle 9i/9.x, Oracle 10g and 11g, RDBMS, MS-SQL
APP SERVERS: Apache, Tomcat, JSERV, JBoss, Jonas, Haht
ERP SYSTEMS: SAP, Primavera, MS Project, ArcGIS, WebMethods
VERSION CONTROL: Git, SNV
PLATFORMS: Unix, Windows

PROFESSIONAL EXPERIENCE

PHD CANCER RESEARCH STUDENT Sep 2009 – Present
University of California, Santa Cruz, Center for Biomolecular Science and Engineering, Santa Cruz, CA.

- Identification of somatic mutations and RNA-Editing events in cancer HTS data
- Classification of breast cancer subtypes using pathway level gene expression and CNV
- Prediction of pathway signatures for breast cancer diagnosis, prognosis, and therapy
- Assembly of a hyperthermophilic archaeon using Sanger 454 and SOLiD paired-end data

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana* and provides software tools used in research on *Arabidopsis thaliana*.

BIOINFORMATICIAN June 2007 – Aug 2009
TAIR, Carnegie Institution for Science, Department of Plant Biology, Palo Alto, CA.

- Implemented a GBrowse visualization tool for the *Arabidopsis thaliana* genome
- Developed Perl and SQL scripts to convert data from database scheme into GFF format
- Used BioPerl GBrowse component to create GBrowse MS SQL database
- Configured 15 tracks for GBrowse and integrated with www.arabidopsis.org website
- Enhanced GBrowse by including ability to download FASTA and GFF formatted files

Impress Software, Inc. is the market leader in SAP application integration with products for Enterprise Project Management (EPM) Systems and Geographical Information Systems (GIS).

CUSTOMER APPLICATION ENGINEER Jan 2006 – May 2008
Impress Software Inc., Sunnyvale, CA.
- Analyzed customer business requirements and wrote scope of work documents
- Configured product for customer business processes and implemented custom Java code
- Executed large batches to initially synchronize data in two ERP systems
- Tested customer business processes and supported customer through go-live and beyond

APPLICATION DEVELOPMENT MANAGER Jan 2004 – Dec 2006
Impress Software Inc., Sunnyvale, CA.
- Designed roadmap for multiple simultaneous projects including resource planning
- Managed team of 5 people through entire software development lifecycle
- Collaborated with team members in different parts of US and Europe
- Recruited, interviewed, and hired suitable new members for development team

BUSINESS APPLICATION ENGINEER Jan 2002 – Dec 2004
Impress Software Inc., Sunnyvale, CA.
- Analyzed customer business processes to determine generic product solution
- Designed product for synchronizing data between project management systems
- Researched ERP systems like Primavera, SAP PM, SAP PS and MS Project
- Developed client- and server-side APIs and stored procedures for product
- Involved in complete development lifecycle for product based applications

GUI DEVELOPER Jan 2001 – Dec 2001
Impress Software Inc., Boston, MA.
- Developed Java GUI for business processes of configuration industry
- Designed Administration Tool to visualize server functionality using EJBs
- Technical contact between Project and Sales teams to Research & Development
- Implemented XML/XSLT based tool for easy project front-end generation

CLIENT-SERVER DEVELOPER Oct 1998 – Dec 2000
Impress Software AG, Hanover, Germany
- Developed client-server communication between product and various app servers
- Designed and implemented session handling for in-house product
- Engineered projects connecting SAP functionality to internet using RFCs

Abels & Kemmner, GmbH specializes in supply chain management consulting and has developed simulation and optimization software from expertise in industry business processes.

WEB DESIGNER Sep 1997 – Sep 1998
Abels & Kemmner GmbH, Aachen, Germany
- Created main commercial web site utilizing functionality of Cold Fusion
- Designed database backend to help manage web site content
- Developed virtual secure site for consultants and customers to access private data

EDUCATION AND CREDENTIALS

DEGREES
Master of Science Degree in Bioinformatics: University of California Santa Cruz, CA – 2011
Master of Science Degree in Computer Science: San José State University, San José, CA – 2008
Bachelor of Arts Degree in Computer Science: Macalester College, St. Paul, MN - 1997

PUBLICATIONS

Diversity of Lung Adenocarcinoma Revealed by Integrative Molecular Profiling. The Cancer Genome Atlas Research Network. (November, 2013). Publication under review at Nature.

The Arabidopsis Information Resource (TAIR): gene structure and function annotation. D. Swarbreck, C. Wilks, P. Lamesch, T.Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A.J. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang and E. Huala. (2007). Nucleic Acids Research, 36, D1009-D1014.

CERTIFICATES

Certification in Computer Science: International Summer University, Switzerland – 2008
Certification in Bioinformatics: San José State University, San José, CA – 2008

POSTERS

Chancellor's Graduate Division Research Grand Prize Winner for poster entitled "Identification of DNA and RNA mutations in Cancer Using High-Throughput Sequencing Data" – Graduate Division Symposium – University of California Santa Cruz, Santa Cruz, CA. – 2012

Identification of RNA-Editing Events in Cancer Using High-Throughput Sequencing Data – American Association for Cancer Research (AACR) Annual Meeting – Orlando, FL. – 2011

Genome Browser for *Arabidopsis thaliana* – Summer Research Day – Carnegie Institution for Science, Department of Plant Biology, Palo Alto, CA. – 2007

Are Genetic Algorithms Effective At Solving the Multiple Sequence Alignment Problem on DNA and Amino Acid Sequences? – College of Science Student Research Day – San José State University, San José, CA – 2007

A Web-Driven Database of Beta Globin Mutations Leading to Beta-thalassemia – College of Science Student Research Day – San José State University, San José, CA – 2007

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Pan-cancer RNA sequencing analysis	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Katherine A. Hoadley, University of North Carolina at Chapel Hill, TCGA RNA sequencing Charles M. Perou, University of North Carolina at Chapel Hill, TCGA PI	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

From the Cancer Genome Atlas (TCGA), we have a wealth of data to start looking at similarities across tumor types. In TCGA's Pan Cancer effort, we looked at Illumina mRNA-sequencing data for twelve tumor types representing over 3,500 samples and analyzed them for patterns unique to each tumor type and patterns present across multiple tumor types. Tumor types tested included rectal adenocarcinoma, colon adenocarcinoma, lung squamous cell carcinoma and adenocarcinoma, uterine corpus endometrial carcinoma, ovarian serous cystadenocarcinoma, kidney clear cell, breast ductal and lobular carcinoma, glioblastoma multiforme, bladder carcinomas, acute myeloid leukemia, and head and neck squamous cell carcinoma. Tissue-specific expression was the dominant signature for five tumor types; however, there were similarities across sets of tumors including a common convergent "squamous" subtype that was populated by several different tumors types (lung squamous, head & neck, some bladder). There were several tumors with divergent expression patterns including basal-like breast cancer versus the rest of breast cancers (i.e. luminal), and bladder tumors that were split into three distinct groups. Pathway or Module-based analysis also provided a method for comparisons across tissues and identified distant patterns of estrogen receptor signaling that differed within distinct subtypes that showed estrogen receptor expression. Integration of other data from TCGA strengthens these observations suggesting that the cell of origin, which is not necessarily coincident with the tissue of origin, plays and important role in the disease behavior.

For the ICGC pan cancer analysis, we propose to 1) expand the RNA/expression comparison to include additional tumor types that are available and 2) use the mutation information from the WGS data to explore the transcriptome data to understand how mutations in cancer genes behave similarly or differently across the tumor types.

Timelines & resources dedicated to project

Spring 2014 to determine set of samples with RNA sequencing data for analyses. For the analysis looking for patterns within transcriptome data, this could be a comprehensive list from the data available. For the analysis looking at the transcriptional effect of mutations across tumor types, it will be dependent on the sub set with both WGS and RNA sequencing.

Summer-December 2014 – Data analysis

The analysis of how mutations affect the transcriptome data across the tumor types is dependent on the availability of the WGS data and the results of the mutation calling algorithms by the groups involved with calling mutations.

We have the time and computer resources dedicated to analyzing the transcriptional data, and we require no additional resources to accomplish this goal.

Research proposal
1) For the comparison of the transcriptomes across tumor types, we will follow a similar approach that was used in the TCGA Pan-can subtypes paper (submitted). Using RNA sequencing data, we will first make sure all fastq files are processed by the same mapping algorithm (MapSplice). We will calculate gene-level summaries as well as run algorithms to look at gene fusions and alternative splicing. Gene clustering approaches will be used to find structure in the data set and help set up the groups that will be used for further classification. Minimally, we will use all available TCGA and ICGC mRNA-seq data, which is ideally generated as paired-end Illumina sequencing approach. 2) We would like to look at how mutations in genes relevant to cancer behave in different tumor types. Does the same mutation cause the same or different transcriptional response in tumors from different organs? We propose to restrict to mutations that are frequently mutated across cancer types such as TP53, PIK3CA, PTEN, etc. The list of genes would be determined from the set of samples used in the ICGC/TCGA WGS project. By restricting the list to frequently mutated genes, we hope that we will be powered to make the comparisons across different tumor types. We will be able to compare the variant allele frequency from WGS with that of the RNA sequencing. The first pass of the analysis will be considering mutations on a gene level – mutated or wild type. Depending on the numbers of mutations, we may be able to also look at hotspot mutations within genes. We will used supervised gene expression analyses to look for genes that are significantly different between wild-type and mutated genes within tumor types and across tumor types. Curated pathways will also be used for comparing the effects of mutations across different tumor types.
Legacy plans
All software and visualization tools I use are open source. I will also make all analysis available.

Katherine A. Hoadley

University of North Carolina at Chapel Hill
Lineberger Comprehensive Cancer Center
CB# 7295, 450 West Drive
Chapel Hill, NC 27599
Email: hoadley@med.unc.edu

EDUCATION

2006 **Ph.D. in Genetics and Molecular Biology**, University of North Carolina at Chapel Hill, NC
2001 **B.S. in Biology and B.A. in Chemistry**, West Virginia Wesleyan College, Buckhannon, WV

RESEARCH EXPERIENCE

Research Assistant Professor, University of North Carolina at Chapel Hill, Department of Genetics
October 2013 - current

Research Associate, University of North Carolina at Chapel Hill, Lineberger Comprehensive Cancer Center, September 2009 – September 2013

- Project Management and RNA sequencing analysis for The Cancer Genome Atlas.

Postdoctoral Research Associate, University of North Carolina at Chapel Hill, Lineberger Comprehensive Cancer Center, August 2007 – August 2009. Research Advisors: Charles M. Perou, D. Neil Hayes

- Genomic analysis and comparison of chemotherapy response in cell lines models and clinical cancer trials.

Postdoctoral Research Associate, Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, January 2007 – July 2007. Research Advisor: Rene Bernards

- Functional genomic approaches to understand molecular pathways important in breast cancer.

Graduate Research Assistant, University of North Carolina at Chapel Hill, Curriculum in Genetics and Molecular Biology, August 2001 – December 2006.
Research Advisor: Charles M. Perou

- Studied chemotherapeutic and general stress responses of luminal and basal-like breast cancers using cell lines models and *in vivo* responses for each subtype.

SELECTED PUBLICATIONS

Brennan CW, et al; TCGA Research Network. The somatic genomic landscape of glioblastoma. Cell. 2013 Oct 10;155(2):462-77.

Ciriello G, Sinha R, **Hoadley KA**, Jacobsen AS, Reva B, Perou CM, Sander C, Schultz N. The molecular diversity of Luminal A breast tumors. Breast Cancer Res Treat. 2013 Oct;141(3):409-20.

Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113-20.

Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, McLellan MD, **Hoadley KA**, Mardis ER, Ley TJ, Perou CM, Wilson RK, Ding L. BreakTrans: uncovering the genomic architecture of gene fusions. Genome Biol. 2013 Aug 23;14(8):R87.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013 Jul 4;499(7456):43-9.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012 Oct 4;490(7418):61-70.

Charles M. Perou
Professor of Genetics and Pathology & Laboratory Medicine
University of North Carolina at Chapel Hill

A. Education

Bates College, Lewiston, ME	B.S.	1987	Biology
University of Utah, Salt Lake City, UT	Ph.D.	1996	Cell Biology

B. Positions and Honors

1992-1995	Graduate Student, Jerry Kaplan Lab, University of Utah, Salt Lake City, UT (PhD advisor)
1997-2000	Postdoctoral Fellow, David Botstein Lab, Department of Genetics, Stanford University, CA
1999	Awarded U.S. Patent No. US5952223, "Compositions for the diagnosis and treatment of Chediak-Higashi syndrome"
2000-2007	Assistant Professor of Genetics and Member of the Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill
2001-present	Faculty Director of the UNC LCCC Bioinformatics Group
2002-2007	Adjunct appointment as Assistant Professor of Pathology & Laboratory Medicine
2002-present	Member of the ALLIANCE Breast Cancer Committee (formerly the CALGB)
2006	Awarded U.S. Patent No. US7118853, "Basal Markers in Breast Cancer"
2007-2009	Associate Professor of Genetics, and Pathology & Laboratory Medicine, UNC
2007	Ruth and Phillip Hettleman Prize for Artistic and Scholarly Achievement, UNC
2008	Co-Director of the UNC LCCC Breast Cancer Research Program
2009	AACR Outstanding Investigator Award for Breast Cancer Research
2010	Professor of Genetics, and Pathology & Laboratory Medicine, UNC
2010	AACR Distinguished Lecture on the Science of Cancer Health Disparities
2011	Endowed Chair, May Goldman Shaw Distinguished Professor of Molecular Oncology, UNC
2011	Danaher Scientific and Medical Award, a Susan G. Komen Award for Scientific Distinction
2012	The European Institute of Oncology Breast Cancer Therapy Award
2013	Hyman L. Battle Distinguished Cancer Research Award, UNC

C. Selected Peer-reviewed publications (in chronological order)

1. C. Fan, D.S. Oh, L. Wessels, B. Weigelt, D.S. Nuyten, A. Nobel, L.J. van't Veer, and **C.M. Perou**. Concordance among gene-expression-based predictors for breast cancer, **New England Journal of Medicine**, 355: 560-569 (2006). PMID: 16899776
2. L. A. Carey, **C. M. Perou**, C. A. Livasy, L. G. Dressler, K. Conway-Dorsey, G. Karaca, D. Cowan, M. Troester, C. Kit Tse, S. Edmiston, S. L. Deming, J. Geradts, M. C. U. Cheang, T. O. Nielsen, P. G. Norman, H. Shelton Earp, and R. C. Millikan. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study, **J. of the American Medical Association**, 295: 2492-2502 (2006). PMID: 16757721
3. L.A. Carey, E.C. Dees, L. Sawyer, L. Gatti, D.T. Moore, F. Collichio, D.W. Ollila, C.I. Sartor, M.L. Graham, and **C.M. Perou**. The triple-negative paradox: Primary tumor chemosensitivity of breast cancer subtypes, **Clinical Cancer Research**, 13(8): 2329-2334, (2007). PMID: 1743809
4. J. S Parker, M. Mullins, M.C.U Cheang, S. Leung, D. Voduc, T. Vickery, X. He, Z. Hu, J.F Quackenbush, I.J. Stijleman, S. Davies, C. Fauron, J. Palazzo, J.S. Marron, A.B. Nobel, E. Mardis, T.O. Nielsen, M.J. Ellis, **C.M. Perou**, and P.S. Bernard. A supervised risk predictor of breast cancer based on intrinsic subtypes. **Journal of Clinical Oncology**, Feb 9 (2009). PMCID: PMC2667820
5. A. Prat, J.S. Parker, C. Fan, O. Karginova, C. Livasy, J. Herschkowitz, X. He and **C.M. Perou**. Phenotypic and Molecular Characterization of the Claudin-low Intrinsic Subtype of Breast Cancer, **Breast Cancer Research**., Sep 2;12(5):R68. (2010). PMCID: PMC3096954
6. L.A. Carey, H. Rugo, P.K. Marcom, E. Mayer, F.J. Esteva, C. Ma, M. Liu, A.-M. Storniolo, M. Rimawi, A. Forero, A.C. Wolff, T. Hobday, A.Ivanova, M.Chiu, M. Ferraro, E. Burrows, P. S. Bernard, K.A. Hoadley, **C.M. Perou**, E.P. Winer on behalf of TBCRC investigators. TBCRC001: Randomized Phase II study of cetuximab in combination with carboplatin in Stage IV triple negative breast cancers. **Journal of Clinical Oncology**, June 4th [Epub ahead of print], (2012). PMCID: PMC3413275
7. **The Cancer Genome Atlas Research Network**, Comprehensive molecular portraits of human breast tumors, **Nature,** Sep 23. doi: 10.1038/nature11412. (2012). PMCID: PMC3465532

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Differences and similarities across solid tumor types (ovarian, breast, prostate, pancreatic) in spatial and temporal genomic heterogeneity

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Justin Guinney, Sage Bionetworks, TCGA PanCancer Analysis Working Group & UCSC GDAC

Adam Margolin, Sage Bionetworks, TCGA PanCancer Analysis Working Group & UCSC GDAC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Rodrigo Dienstmann, Sage Bionetworks	

Background and preliminary data

There is a growing recognition that intratumor heterogeneity (primary vs. metastatis) within the same patient is clinically relevant. Investigations of mutations with respect to clonal/subclonal architecture delineate their temporal orders during tumorigenesis. Recent studies have characterized the emergence of treatment resistant subclones that were present at a minor frequency in the primary tumor. However, a comprehensive characterization of intratumor spatial heterogeneity (primary versus synchronous/metachronous metastasis) and temporal heterogeneity (primary versus metachronous metastasis) across multiple solid tumor types is still missing.

Landau DA et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26.

Bedard PL et al. Tumour heterogeneity in the clinic. Nature. 2013 Sep 19;501(7467):355-64.

Timelines & resources dedicated to project

Methods:
Predefined algorithms
- Somatic Single Nucleotide Variations (sSNV)
- "Significantly Mutated Genes" definition
- Somatic Copy Number Alterations (sCNA)

Specific algorithms
- Clonal/subclonal estimation: ABSOLUTE algorithm
- Pathway analysis: Gene Set Variant Analysis - GSVA

Data dependencies:
Clinical correlations: tumor type, tissue of origin of metastatic sample, timing of metastatic sample acquisition (synchronous vs. metachronous), therapies administered from diagnosis to progression (when metachronous metastatic samples are acquired).

Carter SL et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012 May;30(5):413-21.

Hänzelman S et al. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16;14:7.

Research proposal

Differences and similarities across solid tumor types (ovarian, breast, prostate, pancreatic) in spatial and temporal genomic heterogeneity, as assessed by:

- sSNV, aCNA and GSVA (primary vs. metastasis, treatment naïve vs. progressing to adjuvant chemo/radiotherapy – overall and tumor-specific patterns);
- "Significantly mutated genes" (primary vs. metastatic sites – overall and tumor-specific patterns);
- Metastatic-site-specific mutational signatures and GSVA across tumor types;
- Variant allele fractions of "significantly mutated genes" (clonal/subclonal architecture in primary vs. metastatic sites – overall and tumor-specific patterns);
- Effect of treatment on subclonal heterogeneity (differential clonal evolution/selection in synchronous vs. metachronous metastasis ["treated" vs. "untreated" samples] – overall and tumor-specific patterns).

This research proposal entails many "essential projects" of ICGC WGC pan-cancer analysis:

- *Mutation signatures*
- *Landscape of driver mutations*
- *Pathway analysis*
- *Clinical correlations*
- *Temporal evolution of cancer genomes*

Legacy plans

Sage Bionetworks has at its core the value of repeatable and transparent research. Sage Bionetworks is a leader in openness in biomedical research, and has developed its Synapse software platform (www.synapse.org) to facilitate the sharing of data, code, and results among researchers and the wider public. For these reasons, the TCGA PanCancer consortium selected Synapse as its primary platform for organizing data and results among participating researchers. Similarly, all work proposed here will be shared and exposed within Synapse to the research community. All results from analyses proposed here will be available through an interactive and queryable online interface, and the complete provenance of all algorithms and data will be tracked.

Justin Guinney

Education

Duke University Durham, NC, USA
Ph.D. Computational Biology and Bioinformatics

University of Illinois, Urbana-Champaign IL, USA
B.S. Electrical Engineering

University of Pennsylvania Philadelphia, PA, USA
B.A. Intellectual History & Pre-medicine
magna cum laude

Employment History

Sage Bionetworks Seattle, WA, 2009-2013
Principal Research Scientist

Proventys, Inc. Durham, NC, 2008-2009
Computational Biologist

FiveSight Technologies, Inc. Chicago, IL, 1999-2005
Vice president / cofounder

Selected Publications (*first or last author)

- Guinney, Justin* et al, Modeling RAS phenotype in colorectal cancer uncovers novel molecular traits of RAS dependency and improves prediction of response to targeted agents in patients, *Clinical Cancer Research*, 2013.
- Gregory Hannum, Justin Guinney*, et al, Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates, *Molecular Cell*, 2013.
- Sonja Hänzelmann, Robert Castelo, Justin Guinney*, "GSVA: gene set variation analysis for microarray and RNA-Seq data", *BMC Bioinformatics*, 2013.
- Erhan Bilal, Januz Dutkowski, Justin Guinney, et al, Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling, Plos Comp Bio, 2013.
- Xia Xu, et al, Evidence for type II cells as cells of origin of K-Ras – induced distal lung adenocarcinoma,*PNAS*, 2012.
- Greenawalt, et al, Integrating Genetic Association, Genetics of Gene Expression, and Single Nucleotide Polymorphism Set Analysis to Identify Susceptibility Loci for Type 2 Diabetes Mellitus, *American Journal of Epidemiology*, 2012.
- Lamb, et al, Predictive Genes in Adjacent Normal Tissue Are Preferentially Altered by sCNV during Tumorigenesis in Liver Cancer and May Rate Limiting, *PLoS ONE*, 2011.
- Guinney J*, Wu Q, and Mukherjee S. (2010) "Estimating variable structure and dependence for multitask learning via gradients." *Journal Machine Learning*
- Wu Q, Guinney J, et al, "Learning gradients: predictive models that infer geometry." (2010), *Journal Machine Learning Research*,11, 2175-2198.
- Mendiratta P, Mostaghel E, Guinney J, Tewari A, Porrello A, Barry W, Nelson P, and Febbo P. (2009) Genomic Strategy for Targeting Therapy in Castration-Resistant Prostate Cancer, *Journal Clinical Oncology*.

Adam Arne Margolin, PhD
Director, Computational Biology, Sage Bionetworks, Seattle, USA
Email: margolin@sagebase.org

A. Education

University of Pennsylvania, Philadelphia	B.S.	05/02	Information systems
University of Pennsylvania, Philadelphia	M.S.	12/02	Computer science
Columbia University, New York	M. Phil.	02/06	Biomedical informatics
Columbia University, New York	Ph.D.	01/08	Biomedical informatics

B. Recent Professional Experience

2001 Developer, EGenomics, Inc., New York, NY
2002 Gene Expression Omnibus, Developer, National Center for Biotechnology Information, Bethesda, MD
2001-03 Bioinformatics Application Developer, University of Pennsylvania, Abramson Cancer Research Institute, Philadelphia, PA
2005, 06 Functional Genomics and Systems Biology Group, Intern, IBM T.J. Watson Research Center, Yorktown Heights, NY
2003-08 Department of Biomedical Informatics, Ph.D. Student, Columbia University, New York, NY
2008-10 Cancer Program, Postdoctoral Associate, The Broad Institute of Harvard and MIT, Cambridge, MA
2010-11 Group Leader, Genotype-Specific Therapeutics Initiative, The Broad Institute of Harvard and MIT, Cambridge, MA
2011- Director, Computational Biology, Sage Bionetworks, Seattle, WA

C. Recent Peer-reviewd Publications

1. Jang, I. S., Neto, E. C., Friend, S. H., Margolin, AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput (in press).
2. Neto, E. C., Jang, I. S., Friend, S. H., Margolin, AA. The Stream Algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications in high throughput predictive modeling of cancer cell-line pharmacogenomic screens. Pac Symp Biocomput (in press).
3. Omberg, L., Ellrott, K., Yuan, Y., Kandoth, K., Wong, C., The Cancer Genome Atlas Research Network, Friend, S.H., Stuart, J., Liang, H., Margolin, AA. Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. Nat Genet. Nat Genet. 2013;45(10):1121-6.
4. Margolin AA*, Bilal E*, Huang E*, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Science translational medicine. 2013;5(181):181re1. Epub 2013/04/19.
5. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, Sauerwine BA, Shimoni Y, Moen Vollan HK, Mecham BH, Rueda OM, Tost J, Curtis C, Alvarez MJ, Kristensen VN, Aparicio S, Borresen-Dale AL, Caldas C, Califano A, Friend SH, Ideker T, Schadt EE, Stolovitzky GA, Margolin AA. Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. PLoS computational biology. 2013;9(5):e1003047. Epub 2013/05/15.
6. Barretina J.*, Caponigro G.*, Stransky N.*, Venkatesan K.*, Margolin A.A.*, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. Epub 2012/03/31.
7. Wei G.*, Margolin A.A.*, Haery L., Brown E., Cucolo L., Julian B., et al. Chemical genomics identifies small-molecule MCL1 repressors and BCL-xL as a predictor of MCL1 dependency. Cancer cell. 2012;21(4):547-62. Epub 2012/04/21.
8. Margolin, A. A., K. Wang, A. Califano, I. Nemenman, Multivariate dependence and genetic networks inference. IET Syst Biol, 2011. 4(6): 428-40.
9. Margolin, A.A., S.E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, and T.R. Golub, Empirical Bayes analysis of quantitative proteomics experiments. PLoS One, 2009. 4(10): p. e7454.
10. Margolin, A.A., T. Palomero, P. Sumazin, A. Califano, A.A. Ferrando, and G. Stolovitzky, ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc Natl Acad Sci U S A, 2009. 106(1): p. 244-9.
11. Margolin, A.A. and A. Califano, Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci, 2007. 1115: p. 51-72.
12. Margolin, A.A., K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, and A. Califano, Reverse engineering cellular networks. Nat Protoc, 2006. 1(2): p. 662-71.

Rodrigo Dienstmann, M.D.
Research Scientist, Sage Bionetworks, Seattle, USA
Email: rodrigo.dienstmann@sagebase.org

Education:
2001 - School of Medicine – Universidade Federal do Rio Grande do Sul, Brazil.

Professional Experience:
Jan 2002 – Dec 2003: Intership Internal Medicine, Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil
Jan 2004 – Jan 2006: Fellowship Medical Oncology, Brazilian National Cancer Institute, Rio de Janeiro, Brazil
Feb 2006 – Dic 2009: Clinical Research in Oncology, Brazilian National Cancer Institute, Rio de Janeiro, Brazil
Feb 2010 – Dec 2012: Molecular Therapeutic Research Unit/Phase 1 Unit, Vall d´Hebron Institute of Oncology, Barcelona, Spain
Jan 2013 – Sep 2013: Molecular Pathology Lab, Massachusetts General Hospital, Boston, USA

Recent Publications:

The genomic medicine frontier in human solid tumors: prospects and challenges
Dienstmann R, Rodon J, Barretina J, Tabernero J.
J Clin Oncol 2013;31(15):1874-84.

Biomarker-driven patient selection for early clinical trials
Dienstmann R, Rodon J, Barretina J, Tabernero J.
Curr Opin Oncol 2013;25(3):305-12.

Development of PI3K inhibitors: lessons learned from early clinical trials.
Rodon J, Dienstmann R, Serra V, Tabernero J.
Nat Rev Clin Oncol 2013;10(3):143-53.

Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors
Dienstmann R, Rodon J, Prat A, Perez-Garcia J, Adamo B, Felip E, Cortes J, Iafrate AJ, Nuciforo P, Tabernero J
Annals Oncol 2013 Nov 20. [Epub ahead of print]

Molecular prescreening to select patient population in early clinical trials
Rodón J, Saura C, Dienstmann R, Vivancos A, Cajal SR, Baselga J, Tabernero J.
Nat Rev Clin Oncol 2012;9(6):359-66

Molecular profiling of patients with colorectal cancer and matched targeted therapy in Phase 1 clinical trials
Dienstmann R, Serpico D, Rodon J, Saura C, Macarulla T, Elez ME, Alsina M, Capdevila J, Perez-Garcia J, Sánchez-Ollé G, Aura C, Prudkin L, Landolfi S, Hernández-Losa J, Vivancos A, Tabernero J.
Mol Cancer Ther 2012;11(9):2062-71.

Risk-benefit assessment of bevacizumab in the treatment of breast cancer
Dienstmann R, Ades F, Saini KS, Metzger-Filho O.
Drug Saf 2012;35(1):15-25.

Drug development to overcome resistance to EGFR inhibitors in lung and colorectal cancer
Dienstmann R, De Dosso S, Felip E, Tabernero J.
Mol Oncol 2012;6(1):15-26.

BRAF as a target for cancer therapy
Dienstmann R, Tabernero J.
Anticancer Agents Med Chem 2011;11(3):285-95.

Molecular predictors of response to chemotherapy in colorectal cancer
Dienstmann R, Vilar E, Tabernero J.
Cancer J 2011;17(2):114-26.

Personalizing therapy with targeted agents in non-small cell lung cancer
Dienstmann R, Martinez P, Felip E.
Oncotarget 2011;2(3):165-77.

Combined modality therapy of stage IIIC breast cancer
Dienstmann R, Branco LG, Rezende LM, Freitas LC, Lima CF, Rodrigues GJ, Noronha Filho H, Sarmento RMB, Small I, Bines J.
Breast J 2011;17(3):331-3.

Toxicity as biomarker of efficacy of molecular targeted therapies: focus on EGFR and VEGFR inhibiting anticancer agents
Dienstmann R, Braña I, Rodon J, Tabernero J.
Oncologist 2011;16(12):1729-40.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Conjoint modeling of cell lines and patient tumor data to infer disease specific molecular variants of drug sensitivity and resistance

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Justin Guinney, Sage Bionetworks, TCGA PanCancer Analysis Working Group & UCSC GDAC
Adam Margolin, Sage Bionetworks, TCGA PanCancer Analysis Working Group & UCSC GDAC

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Elias Neto, Sage Bionetworks
Rodrigo Dientsmann, Sage Bionetworks

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

The unresponsiveness to anticancer drugs in patients outlines the need to identify novel and robust biomarkers of response to therapy. The recent release of large molecular cell line datasets labeled for drug sensitivity enables the development of such predictors. A key limitation of cell line based models, however, is that they poorly reflect the extent of molecular heterogeneity and of tumor microenvironment relationships existing at the patient level. We aimed to overcome these limitations by combining cell line datasets with patient cohorts to uncover novel molecular traits associated with drug response at the patient level.

As preliminary work, we have developed predictive models of drug sensitivity using the Cancer Cell Line Encyclopedia (CCLE) and the Sanger Genomics of Drug Sensitivity (Sanger), and have applied these models on many of the TCGA disease cohorts to infer molecular traits (somatic mutations / copy-number aberrations) of drug sensitivity. We have developed models for over 140 drugs from the combined CCLE and Sanger data sets, applied on many of the TCGA cohorts. Our models are able to recover many of the gold standards of known drug sensitivity and/or resistance: EGFR mutation for erlotinib (EGFR inhibitor) sensitivity in lung adenocarcinoma; KRAS/NRAS/BRAF mutation for selumitinib (MEK inhibitor) in colorectal cancer; ERBB2 amplification for sensitivity to lapatanib in invasive breast cancer. Moreover, our pipeline is able to uncover many putative markers of drug sensitivity and resistance with preclinical evidence. With our collaborator Rene Bernards at the NKI, we have positively validated several novel genes uncovered by our models as conferring increased sensitivity to MEK inhibition in CRC. Our methodology demonstrates the value of modeling across cell lines and patients conjointly to undercover the mechanisms of resistance or sensitivity to drugs in patients.

Timelines & resources dedicated to project

Our analysis depends on the availability of variant calls across all tumors, including germ line and somatic mutations, copy-number alterations, and translocations. Moreover, corresponding mRNA profiling is required to translate drug sensitivity from cell lines into ICGC data.

Research proposal

In our preliminary investigation, we had focused on identifying exonic mutations and copy-number alterations that are predictive of drug sensitivity/resistance within the TCGA patient cohorts. The data proposed for use in the ICGC consortium will allow us to vastly broaden our investigation, permitting a more extensive study of putative causal variants capable of altering signaling pathways and inducing changes to therapeutic response in patients. While the use of whole genome sequencing will substantially increase both the resolution and the number of variants our models might consider (such as non-exonic mutations and translocations), it will also increase the likelihood of identifying false associations by chance. Therefore there is a need to develop robust models capable of incorporating new variants, while simultaneously penalizing overly complex or non-informative models. We therefore propose the following areas of research to adapt our models to the ICGC data sets:

A. Incorporation of prior knowledge for data dimension reduction.

Many targeted therapies act by inhibiting or disrupting a signaling pathway. Variants that are more likely to impact a pathway corresponding to a drug's mechanism of action should be given priority over other variants. By integrating knowledge from protein-protein interaction (PPI) networks and curated pathway databases, we propose to reduce the effective dimensionality of the data. Hofree, et al (Nature Methods, 2013) recently highlighted the value of this approach in patient stratification of TCGA data. Similarly, we propose to combine variants that may be functionally similar, thereby increasing our power of detection for (de)sensitizing variants.

B. Conjoint models of cell-line and ICGC patient data.

A key assumption of our drug sensitivity models is that mRNA patterns identified in cell lines represent drug sensitivity patterns shared among many tumor types found in TCGA (ICGC). Given the large differences in tumor biology among disease types, this is likely to be an oversimplification. We therefore propose to enhance our modeling pipeline using an approach called "transfer learning": conceptually, the goal is to explicitly adapt each drug model to a single disease of interest (e.g. colorectal cancer), whereby we find common structure between cell lines and the disease that is optimally predictive of drug sensitivity. In addition to enhancing the robustness of our drug sensitivity models, we believe this approach will have a more general utility for data modelers and experimentalists who must translate information between cell lines and patient data.

C. Integration of predictions within a clinical knowledge database

We have developed a structured knowledge database (Kdb) with standardized terminology describing associations that integrate different layers of annotations: tumor types, genes, variants, response/resistance patterns to approved and experimental agents under clinical investigation and PubMed identifiers. Predictive associations are classified in a hierarchical way based on the strength of evidence: (i) late trials; (ii) early trials; (iii) case reports; and (iv) preclinical data. The Kdb has more than 500 unique gene - drug interactions. Our goal is to integrate this knowledge base with the predictions from our drug sensitivity models, providing a large community resource coupling putative markers with experimental and clinical evidence that can be used to guide both experiments and clinical decision-making.

Legacy plans

Sage Bionetworks has at its core the value of repeatable and transparent research. Sage Bionetworks is a leader in openness in biomedical research, and has developed its Synapse software platform (www.synapse.org) to facilitate the sharing of data, code, and results among researchers and the wider public. For these reasons, the TCGA PanCancer consortium selected Synapse as its primary platform for organizing data and results among participating researchers. Similarly, all work proposed here will be shared and exposed within Synapse to the research community. All results from analyses proposed here will be available through an interactive and queryable online interface, and the complete provenance of all algorithms and data will be tracked.

Justin Guinney

Education

Duke University Durham, NC, USA
Ph.D. Computational Biology and Bioinformatics

University of Illinois, Urbana-Champaign IL, USA
B.S. Electrical Engineering

University of Pennsylvania Philadelphia, PA, USA
B.A. Intellectual History & Pre-medicine
magna cum laude

Employment History

Sage Bionetworks Seattle, WA, 2009-2013
Principal Research Scientist

Proventys, Inc. Durham, NC, 2008-2009
Computational Biologist

FiveSight Technologies, Inc. Chicago, IL, 1999-2005
Vice president / cofounder

Selected Publications (*first or last author)

- Guinney, Justin* et al, Modeling RAS phenotype in colorectal cancer uncovers novel molecular traits of RAS dependency and improves prediction of response to targeted agents in patients, *Clinical Cancer Research*, 2013.
- Gregory Hannum, Justin Guinney*, et al, Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates, *Molecular Cell*, 2013.
- Sonja Hänzelmann, Robert Castelo, Justin Guinney*, "GSVA: gene set variation analysis for microarray and RNA-Seq data", *BMC Bioinformatics*, 2013.
- Erhan Bilal, Januz Dutkowski, Justin Guinney, et al, Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling, Plos Comp Bio, 2013.
- Xia Xu, et al, Evidence for type II cells as cells of origin of K-Ras – induced distal lung adenocarcinoma,*PNAS*, 2012.
- Greenawalt, et al, Integrating Genetic Association, Genetics of Gene Expression, and Single Nucleotide Polymorphism Set Analysis to Identify Susceptibility Loci for Type 2 Diabetes Mellitus, *American Journal of Epidemiology*, 2012.
- Lamb, et al, Predictive Genes in Adjacent Normal Tissue Are Preferentially Altered by sCNV during Tumorigenesis in Liver Cancer and May Rate Limiting, *PLoS ONE*, 2011.
- Guinney J*, Wu Q, and Mukherjee S. (2010) "Estimating variable structure and dependence for multitask learning via gradients." *Journal Machine Learning*
- Wu Q, Guinney J, et al, "Learning gradients: predictive models that infer geometry." (2010), *Journal Machine Learning Research*,11, 2175-2198.
- Mendiratta P, Mostaghel E, Guinney J, Tewari A, Porrello A, Barry W, Nelson P, and Febbo P. (2009) Genomic Strategy for Targeting Therapy in Castration-Resistant Prostate Cancer, *Journal Clinical Oncology*.

Adam Arne Margolin, PhD
Director, Computational Biology, Sage Bionetworks, Seattle, USA
Email: margolin@sagebase.org

A. Education

University of Pennsylvania, Philadelphia	B.S.	05/02	Information systems
University of Pennsylvania, Philadelphia	M.S.	12/02	Computer science
Columbia University, New York	M. Phil.	02/06	Biomedical informatics
Columbia University, New York	Ph.D.	01/08	Biomedical informatics

B. Recent Professional Experience

2001 Developer, EGenomics, Inc., New York, NY
2002 Gene Expression Omnibus, Developer, National Center for Biotechnology Information, Bethesda, MD
2001-03 Bioinformatics Application Developer, University of Pennsylvania, Abramson Cancer Research Institute, Philadelphia, PA
2005, 06 Functional Genomics and Systems Biology Group, Intern, IBM T.J. Watson Research Center, Yorktown Heights, NY
2003-08 Department of Biomedical Informatics, Ph.D. Student, Columbia University, New York, NY
2008-10 Cancer Program, Postdoctoral Associate, The Broad Institute of Harvard and MIT, Cambridge, MA
2010-11 Group Leader, Genotype-Specific Therapeutics Initiative, The Broad Institute of Harvard and MIT, Cambridge, MA
2011- Director, Computational Biology, Sage Bionetworks, Seattle, WA

C. Recent Peer-reviewd Publications

1. Jang, I. S., Neto, E. C., Friend, S. H., Margolin, AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput (in press).
2. Neto, E. C., Jang, I. S., Friend, S. H., Margolin, AA. The Stream Algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications in high throughput predictive modeling of cancer cell-line pharmacogenomic screens. Pac Symp Biocomput (in press).
3. Omberg, L., Ellrott, K., Yuan, Y., Kandoth, K., Wong, C., The Cancer Genome Atlas Research Network, Friend, S.H., Stuart, J., Liang, H., Margolin, AA. Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. Nat Genet. Nat Genet. 2013;45(10):1121-6.
4. Margolin AA*, Bilal E*, Huang E*, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Science translational medicine. 2013;5(181):181re1. Epub 2013/04/19.
5. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, Sauerwine BA, Shimoni Y, Moen Vollan HK, Mecham BH, Rueda OM, Tost J, Curtis C, Alvarez MJ, Kristensen VN, Aparicio S, Borresen-Dale AL, Caldas C, Califano A, Friend SH, Ideker T, Schadt EE, Stolovitzky GA, Margolin AA. Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. PLoS computational biology. 2013;9(5):e1003047. Epub 2013/05/15.
6. Barretina J.*, Caponigro G.*, Stransky N.*, Venkatesan K.*, Margolin A.A.*, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. Epub 2012/03/31.
7. Wei G.*, Margolin A.A.*, Haery L., Brown E., Cucolo L., Julian B., et al. Chemical genomics identifies small-molecule MCL1 repressors and BCL-xL as a predictor of MCL1 dependency. Cancer cell. 2012;21(4):547-62. Epub 2012/04/21.
8. Margolin, A. A., K. Wang, A. Califano, I. Nemenman, Multivariate dependence and genetic networks inference. IET Syst Biol, 2011. 4(6): 428-40.
9. Margolin, A.A., S.E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, and T.R. Golub, Empirical Bayes analysis of quantitative proteomics experiments. PLoS One, 2009. 4(10): p. e7454.
10. Margolin, A.A., T. Palomero, P. Sumazin, A. Califano, A.A. Ferrando, and G. Stolovitzky, ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc Natl Acad Sci U S A, 2009. 106(1): p. 244-9.
11. Margolin, A.A. and A. Califano, Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci, 2007. 1115: p. 51-72.
12. Margolin, A.A., K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, and A. Califano, Reverse engineering cellular networks. Nat Protoc, 2006. 1(2): p. 662-71.

Elias Chaibub Neto November 27, 2013

Education

2004 - 2010	University of Wisconsin-Madison	PhD	Statistics
2000 - 2003	University of Sao Paulo (Brazil)	MA	Statistics
1994 - 1998	University of Sao Paulo (Brazil)	BS	Agronomy

Employment

2011 - 2013 Research Fellow, Computational Biology, Sage Bionetworks.
2013 - present Senior Scientist, Computational Biology, Sage Bionetworks.

Research Interests

Machine learning, Bayesian statistics, statistical genetics and genomics.

Selected Publications

1- Chaibub Neto et al (2013) The Stream Algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity. *Pacific Symposium on Biocomputing* 2014 (accepted).

2- Jang et al. (2013) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing* 2014 (accepted).

3- Moon et al. (2013) Bayesian causal phenotype network incorporating genetic variation and biological knowledge. In *Probabilistic Graphical Models in Genetics* (accepted).

4- Chaibub Neto et al. (2013) Modeling causality for pairs of phenotypes in systems genetics. *Genetics* 193: 1003-1013.

5- Chaibub Neto et al. (2012) Quantile-based permutation thresholds for QTL hotspots. *Genetics* 191: 1355-1365.

6- Chaibub Neto et al. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Annals of Applied Statistics* 4: 320-339.

7- Zhao et al. (2009). Obesity and genetics regulate microRNAs in islets, liver, and adipose of diabetic mice. *Mammalian Genome* 20: 476-485.

8- Chaibub Neto et al. (2008). Inferring causal phenotype networks from segregating populations. *Genetics* 179: 1089-1100.

9- Ferrara et al. (2008). Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genetics* 4: e1000034.

11- Keller et al. (2008). A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Research* 18: 706-716.

12- Doksum et al. (2007). Thinking outside the box: statistical inference based on Kullback-Leibler empirical projections. *Statistics and Probability Letters* 77: 1201-1213.

Rodrigo Dienstmann, M.D.
Research Scientist, Sage Bionetworks, Seattle, USA
Email: rodrigo.dienstmann@sagebase.org

Education:
2001 - School of Medicine – Universidade Federal do Rio Grande do Sul, Brazil.

Professional Experience:
Jan 2002 – Dec 2003: Intership Internal Medicine, Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil
Jan 2004 – Jan 2006: Fellowship Medical Oncology, Brazilian National Cancer Institute, Rio de Janeiro, Brazil
Feb 2006 – Dic 2009: Clinical Research in Oncology, Brazilian National Cancer Institute, Rio de Janeiro, Brazil
Feb 2010 – Dec 2012: Molecular Therapeutic Research Unit/Phase 1 Unit, Vall d´Hebron Institute of Oncology, Barcelona, Spain
Jan 2013 – Sep 2013: Molecular Pathology Lab, Massachusetts General Hospital, Boston, USA

Recent Publications:

The genomic medicine frontier in human solid tumors: prospects and challenges
Dienstmann R, Rodon J, Barretina J, Tabernero J.
J Clin Oncol 2013;31(15):1874-84.

Biomarker-driven patient selection for early clinical trials
Dienstmann R, Rodon J, Barretina J, Tabernero J.
Curr Opin Oncol 2013;25(3):305-12.

Development of PI3K inhibitors: lessons learned from early clinical trials.
Rodon J, Dienstmann R, Serra V, Tabernero J.
Nat Rev Clin Oncol 2013;10(3):143-53.

Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors
Dienstmann R, Rodon J, Prat A, Perez-Garcia J, Adamo B, Felip E, Cortes J, Iafrate AJ, Nuciforo P, Tabernero J
Annals Oncol 2013 Nov 20. [Epub ahead of print]

Molecular prescreening to select patient population in early clinical trials
Rodón J, Saura C, Dienstmann R, Vivancos A, Cajal SR, Baselga J, Tabernero J.
Nat Rev Clin Oncol 2012;9(6):359-66

Molecular profiling of patients with colorectal cancer and matched targeted therapy in Phase 1 clinical trials
Dienstmann R, Serpico D, Rodon J, Saura C, Macarulla T, Elez ME, Alsina M, Capdevila J, Perez-Garcia J, Sánchez-Ollé G, Aura C, Prudkin L, Landolfi S, Hernández-Losa J, Vivancos A, Tabernero J.
Mol Cancer Ther 2012;11(9):2062-71.

Risk-benefit assessment of bevacizumab in the treatment of breast cancer
Dienstmann R, Ades F, Saini KS, Metzger-Filho O.
Drug Saf 2012;35(1):15-25.

Drug development to overcome resistance to EGFR inhibitors in lung and colorectal cancer
Dienstmann R, De Dosso S, Felip E, Tabernero J.
Mol Oncol 2012;6(1):15-26.

BRAF as a target for cancer therapy
Dienstmann R, Tabernero J.
Anticancer Agents Med Chem 2011;11(3):285-95.

Molecular predictors of response to chemotherapy in colorectal cancer
Dienstmann R, Vilar E, Tabernero J.
Cancer J 2011;17(2):114-26.

Personalizing therapy with targeted agents in non-small cell lung cancer
Dienstmann R, Martinez P, Felip E.
Oncotarget 2011;2(3):165-77.

Combined modality therapy of stage IIIC breast cancer
Dienstmann R, Branco LG, Rezende LM, Freitas LC, Lima CF, Rodrigues GJ, Noronha Filho H, Sarmento RMB, Small I, Bines J.
Breast J 2011;17(3):331-3.

Toxicity as biomarker of efficacy of molecular targeted therapies: focus on EGFR and VEGFR inhibiting anticancer agents
Dienstmann R, Braña I, Rodon J, Tabernero J.
Oncologist 2011;16(12):1729-40

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Identify genetic patterns of mutations and fusion transcripts through integrative analysis of large-scale cross-cancer genome and transcriptome sequencing data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Adam Margolin, Sage Bionetworks, TCGA PanCancer Analysis Working Group & UCSC GDAC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yin Hu, Sage Bionetworks	

Background and preliminary data

Mutations and fusion transcripts are key biomarkers for cancer. As demonstrated by TCGA pan-cancer data, the collection of mutations across cancer types has provided an unprecedented opportunity to redefine cancers based on their genetic characteristics, rather than just by organ of origin. Separately, for the discovery of cancer biomarkers in transcriptome, effort has been devoted to fusion transcript detection in specific tumor studies. However, there lacks a comprehensive profile of gene fusion events that span multiple types of cancer. The connection between genomic variations and transcriptional aberrations is also not clear. Therefore, we propose to picture the landscape of chimeric events in mRNA transcriptomes across cancer types using the RNA-seq data in the pan-cancer analysis, and to investigate their relation with mutations in the genome. Specifically, we seek for the answers to the following questions:

1. Are fusion transcripts organ-dependent? Which fusion transcripts are common among cancer types and which may correspond to specific types?
2. Are there patterns, such as co-occurrence, between mutations and fusion transcripts? Are these patterns organ-specific?
3. Can mutation-fusion transcript patterns help better organize cancers?

In addition to the integrated analysis of genome and transcriptome, we aim to reshape the current sample-by-sample strategy by jointly modeling thousands of samples from multiple cancer types.

In the past year, we have extensively studied the transcriptomes of the breast cancer (BRCA) samples of TCGA project, including analyses of gene fusion and alternative splicing. Despite the large sample size (>800 RNA-seq samples) and the thousands of putative fusion transcripts found, the recurrence of the fusion transcripts was highly limited (most present in <2% samples). The revealed transcription signals suggested considerable heterogeneity within each subtype. These observations have driven us to seek for further evidence across cancer types. The pan-cancer data will enable the transfer of knowledge and discoveries from different cancer types, which may shed light on common and unique driver biomarkers. The large-scale paired data will support desired statistical power and help alleviate false discoveries. The availability of both genome and transcriptome sequencing data will be essential to the proposed integrative analysis.

Timelines & resources dedicated to project

Key intermediate milestones include reaching a list of fusion transcripts with strong evidence, constructing the co-occurrence patterns of fusion transcripts and mutations, and deriving the clusters of cross-cancer samples that exhibit similar markers of fusion transcripts and mutations together with the fusion-mutation correlations.

The analysis of fusion transcripts will be based on the RNA-seq data sets from the pan-cancer studies. The analysis of mutation-fusion transcript patterns will further rely on the variants called by the pan-cancer studies.

Research proposal

This integrative analysis consists of a series of steps: developing a joint analysis pipeline that finds fusion transcripts in thousands of samples, statistical modeling for highly confident fusion transcripts, filtering steps, confirmation using DNA-seq data, identification of co-occurrence patterns of mutations and fusion transcripts, classification of potential tumor subtypes and branches across organs. The plans for the core components have been detailed below.

Fusion transcript discovery. At least three fusion detection algorithms that take different strategies yet have recognized accuracies will be applied, including the MapSplice package co-developed by the investigator. A comprehensive comparison among different algorithms will be conducted, for an in-depth understanding of the complexity of the cancer transcriptomes. A consensus list of fusion transcripts will be generated according to the comparison, followed by biological filters regarding, for example, the locations of fusion sites, read-throughs, ribosomal genes and pseudo-genes. WGS data may be further used to help validate the presence of some of the candidate fusion transcripts.

Co-occurrence patterns of mutations and fusion transcripts. The presence of detected fusion transcripts will reveal the recurrent fusion transcripts and frequent partner genes. If the recurrence is prominent, the genes connected by fusion junctions may be represented as a fusion gene graph, on which a subgraph mining-based algorithm can be developed and applied for the identification of frequent co-occurred fusion events. Provided with the mutations, the co-occurrence of mutations and fusion transcripts can be determined using an algorithm analogous to frequent item-set mining.

Organizing samples according to fusion transcript-mutation patterns. A cluster analysis will be designed and performed on all samples, combining features from both genomes and transcriptomes.

Legacy plans

Sage Bionetworks has at its core the value of repeatable and transparent research. Sage Bionetworks is a leader in openness in biomedical research, and has developed its Synapse software platform (www.synapse.org) to facilitate the sharing of data, code, and results among researchers and the wider public. For these reasons, the TCGA PanCancer consortium selected Synapse as its primary platform for organizing data and results among participating researchers. Similarly, all work proposed here will be shared and exposed within Synapse to the research community. All results from analyses proposed here will be available through an interactive and queryable online interface, and the complete provenance of all algorithms and data will be tracked.

Adam Arne Margolin, PhD
Director, Computational Biology, Sage Bionetworks, Seattle, USA
Email: margolin@sagebase.org

A. Education

University of Pennsylvania, Philadelphia	B.S.	05/02	Information systems
University of Pennsylvania, Philadelphia	M.S.	12/02	Computer science
Columbia University, New York	M. Phil.	02/06	Biomedical informatics
Columbia University, New York	Ph.D.	01/08	Biomedical informatics

B. Recent Professional Experience

2001 Developer, EGenomics, Inc., New York, NY
2002 Gene Expression Omnibus, Developer, National Center for Biotechnology Information, Bethesda, MD
2001-03 Bioinformatics Application Developer, University of Pennsylvania, Abramson Cancer Research Institute, Philadelphia, PA
2005, 06 Functional Genomics and Systems Biology Group, Intern, IBM T.J. Watson Research Center, Yorktown Heights, NY
2003-08 Department of Biomedical Informatics, Ph.D. Student, Columbia University, New York, NY
2008-10 Cancer Program, Postdoctoral Associate, The Broad Institute of Harvard and MIT, Cambridge, MA
2010-11 Group Leader, Genotype-Specific Therapeutics Initiative, The Broad Institute of Harvard and MIT, Cambridge, MA
2011- Director, Computational Biology, Sage Bionetworks, Seattle, WA

C. Recent Peer-reviewd Publications

1. Jang, I. S., Neto, E. C., Friend, S. H., Margolin, AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput (in press).
2. Neto, E. C., Jang, I. S., Friend, S. H., Margolin, AA. The Stream Algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications in high throughput predictive modeling of cancer cell-line pharmacogenomic screens. Pac Symp Biocomput (in press).
3. Omberg, L., Ellrott, K., Yuan, Y., Kandoth, K., Wong, C., The Cancer Genome Atlas Research Network, Friend, S.H., Stuart, J., Liang, H., Margolin, AA. Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. Nat Genet. Nat Genet. 2013;45(10):1121-6.
4. Margolin AA*, Bilal E*, Huang E*, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Science translational medicine. 2013;5(181):181re1. Epub 2013/04/19.
5. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, Sauerwine BA, Shimoni Y, Moen Vollan HK, Mecham BH, Rueda OM, Tost J, Curtis C, Alvarez MJ, Kristensen VN, Aparicio S, Borresen-Dale AL, Caldas C, Califano A, Friend SH, Ideker T, Schadt EE, Stolovitzky GA, Margolin AA. Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. PLoS computational biology. 2013;9(5):e1003047. Epub 2013/05/15.
6. Barretina J.*, Caponigro G.*, Stransky N.*, Venkatesan K.*, Margolin A.A.*, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. Epub 2012/03/31.
7. Wei G.*, Margolin A.A.*, Haery L., Brown E., Cucolo L., Julian B., et al. Chemical genomics identifies small-molecule MCL1 repressors and BCL-xL as a predictor of MCL1 dependency. Cancer cell. 2012;21(4):547-62. Epub 2012/04/21.
8. Margolin, A. A., K. Wang, A. Califano, I. Nemenman, Multivariate dependence and genetic networks inference. IET Syst Biol, 2011. 4(6): 428-40.
9. Margolin, A.A., S.E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, and T.R. Golub, Empirical Bayes analysis of quantitative proteomics experiments. PLoS One, 2009. 4(10): p. e7454.
10. Margolin, A.A., T. Palomero, P. Sumazin, A. Califano, A.A. Ferrando, and G. Stolovitzky, ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc Natl Acad Sci U S A, 2009. 106(1): p. 244-9.
11. Margolin, A.A. and A. Califano, Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci, 2007. 1115: p. 51-72.
12. Margolin, A.A., K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, and A. Califano, Reverse engineering cellular networks. Nat Protoc, 2006. 1(2): p. 662-71.

Yin Hu, PhD

Research Scientist, Computational Biology, Sage Bionetworks, Seattle, USA

Email: yin.hu@sagebase.org

A. Education

University of Science and Technology of China, Hefei, China	B.S.	06/08	Computer Science
University of Kentucky, Lexington, KY	M.S.	05/13	Statistics
University of Kentucky, Lexington, KY	Ph.D.	10/13	Computer Science

B. Recent Professional Experience

2011.6 – 2011.8	Visiting student, University of North Carolina at Chapel Hill, Chapel Hill, NC
2008.8 – 2013.10	Research Assistant, Department of Computer Science, University of Kentucky, Lexington, KY
2013.11 – present	Research Scientist, Sage Bionetworks, Seattle, WA

C. Recent Peer-reviewd Publications

23. Yan Huang, Yin Hu, Corbin D. Jones, James N. MacLeod, Derek Chiang, Yufeng Liu, Jan F. Prins, and Jinze Liu. A Robust Method for Transcript Quantification with RNA-seq Data. Journal of Computational Biology, March 2013, 20(3): 167-187.
24. Yin Hu, Yan Huang, Ying Du, Christian Orellana, Darshan Singh, Amy Johnson, Anais Monroy, Pei-Fen Kuan, Scott Hammond, Liza Makowski, Scott Randell, Derek Chiang, David Hayes, Corbin Jones, Yufeng Liu, Jan Prins, Jinze Liu. DiffSplice: the Genome-Wide Detection of Differential Splicing Events with RNA-seq. Nucleic Acids Research, 2012, doi: 10.1093/nar/gks1026.
25. Yan Huang, Yin Hu, Corbin D. Jones, James N. MacLeod, Derek Chiang, Yufeng Liu, Jan F. Prins, and Jinze Liu. A Robust Method for Transcript Quantification with RNA-seq Data. Annual International Conference on Research in Computational Molecular Biology (RECOMB), 2012.
26. Darshan Singh, Christian F. Orellana, Yin Hu, Corbin D. Jones, Yufeng Liu, Derek Y. Chiang, Jinze Liu, and Jan F. Prins. FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. Bioinformatics, 2011, 27(19): 2633–2640.
27. Yin Hu, Kai Wang, Xiaping He, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. A Probabilistic Framework for Aligning Paired-end RNA-seq Data. Bioinformatics, 2010, 26:1950–1957.
28. Jizhou Gao, Yin Hu, Jinze Liu, and Ruigang Yang. Unsupervised Learning of High-order Structural Semantics from Images. International Conference on Computer Vision (ICCV), 2009.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Mutation and integrative subtyping based on kernalized tensor methods

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Larsson Omberg, Sage Bionetworks, TCGA PanCancer Analysis Working Group & UCSC GDAC

Adam Margolin, Sage Bionetworks ,TCGA PanCancer Analysis Working Group & UCSC GDAC

Name(s) & institute(s) of junior investigators

(Name no more than 2; append 1 page CV for each)

Rodrigo Dientsmann, Sage Bionetworks

Name(s) & institute(s) of non-ICGC collaborators

(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Cancer diagnosis, prognosis and treatment have traditionally been guided primarily by site of origin and pathology but stratification based on molecular phenotype is increasingly guiding clinicians as multiple subtypes for many cancer tissues have been discovered. The extent of the similarity of genomics and genetics across tumor subtypes from different tissues has however not been fully characterized as it is primarily driven by tissue of origin effects. The TCGA Pan-Cancer subtyping effort approached this question primarily using a technique of clusters of clusters and was not able extract consistent clusters from mutational data (submitted results). We have been working on a method using kernels and tensor decomposition for data integrative subtyping.

As preliminary work we have implemented a diffusion model that allows integration of external data such as protein interaction (PPI) networks and curated pathway databases with genetic features. Allowing for specific mutations, that are often rare, to be interpreted in the context of pathways or interacting genes. A similar method was recently shown to work well by Hofree, et al (Nature Methods, 2013) in patient stratification of TCGA data.

We have also developed a method for integrating multiple genomic scale datasets for integrative analysis based on the higher-order singular value decomposition (HOSVD) (Omberg et al. PNAS 2007) and the use of kernels.

Timelines & resources dedicated to project

Our analysis depends on the availability of the somatic variant calls across all tumors for mutation subtyping and copy-number alterations, corresponding mRNA profiling and/or methylation profiling for integrative subtyping.

Research proposal

Our preliminary work has been focused on the method development and theoretical applications. The ICGC-Pancancer proposed dataset provides a unique application by providing both multiple platforms of data for a subset of patients and also somatic mutations from whole genome sequencing. The latter will allow us to explore the effects of non-exonic mutations on pathway activity. We propose to perform integrative cross-tumor subtyping by:

A. **Compute a kernels for patient-patient similarity**
   Using the methodology described above for diffusing individual mutations on a scaffolding of curated pathways or ppi networks we can get a measure of patient specific mutated genes or pathways allowing us to compute a patient-patient distance matrix or kernel. Similarly for the other genomic datasets (expression, copy-number, etc) we can compute kernels to represent the distance between patients. By looking both at tumor specific and tumor integrative kernels in the next steps we can explore the effects of tissue on subtypes.
B. **Integrative subtyping based on tensor decomposition**
   By combining all the kernels for the different data types into a tensor with dimensions patient x patient x data type we can perform a decomposition into eigen-subspaces that capture groups of patients with similar genomic signatures
C. **Explore drivers of subtypes**
   Depending on the kernel used the drivers of the subtype can be extracted by projecting the eigen-substpaces defining the subtypes into the underlying genomic data giving the specific features contributing to the each subtype. Due to the non-linear nature of the proposed mutation kernel it will not be possible to project the specific mutational drivers but it should be possible to perform a post-hoc enrichment analysis to extract this information
D. **Explore clinical covariates of subtypes**
   We propose the association between subtypes and whatever clinical covariates are made available in the ICGC Pancancer dataset.

Legacy plans

Sage Bionetworks has at its core the value of repeatable and transparent research. Sage Bionetworks is a leader in openness in biomedical research, and has developed its Synapse software platform (www.synapse.org) to facilitate the sharing of data, code, and results among researchers and the wider public. For these reasons, the TCGA PanCancer consortium selected Synapse as its primary platform for organizing data and results among participating researchers. Similarly, all work proposed here will be shared and exposed within Synapse to the research community. All results from analyses proposed here will be available through an interactive and queryable online interface, and the complete provenance of all algorithms and data will be tracked.

Larsson Omberg, PhD
Senior Scientist, Computational Biology, Sage Bionetworks, Seattle, USA
Email: larsson.omberg@sagebase.org

A. Education

Royal Institute of Technology (KTH), Stockholm	M.SE.	12/99	Engineering Physics
University of Texas at Austin, TX	Ph.D.	12/07	Physics

B. Recent Professional Experience

1999	Center for Neurodynamics, Visiting researcher, University of Missouri, St Louis, MO
1999-00	Computer Science Dept., Instructor, Royal Institute of Technology, Stockholm, Sweden
2000	Jensen Education, Instructor, Stockholm, Sweden
2001-04	Physics Department, Graduate Research Assistant, University of Texas at Austin, Austin, TX
2001-02	Physics Department, Teaching Assistant, University of Texas at Austin, Austin, TX
2002-04	Physics Department, Assistant Instructor, University of Texas at Austin, Austin, TX
2004-07	Physics Department, Graduate Research Assistant, University of Texas at Austin, Austin, TX
2008	Dept. of Biomedical Engineering, Postdoctoral Researcher, University of Texas at Austin, Austin, TX
2008-11	Dept. of Biostatistics, Postdoctoral Researcher, Cornell University, Ithaca, NY

C. Selected Peer-reviewed Publications

1. L Omberg*, K Ellrott*, Y Yuan, K Kandoth, C Wong, The Cancer Genome Atlas Research Network, Friend, S.H., Stuart, J., Liang, H., Margolin, AA. (2013). Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. Nat Genet. Nat Genet. 2013;45(10):1121-6.
2. L Omberg, J Salit, N Hackett, J Fuller, R Matthew, L Chouchane, JL Rodriguez-Flores, et al. (2012). Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC genetics*, *13*(1), 49. doi:10.1186/1471-2156-13-49.
3. Kidd, J. M., et al. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *The American Journal of Human Genetics*, *91*(4), 660-671.
4. A. Brisbin, K. Bryc, J. Byrnes, F. Zakharia, L Omberg, J. Degenhardt, et al. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology*, *84*(4), 343-364.
5. N Hackett , MW Butler, R. Shaykhiev, J Salit ,L Omberg, J Rodriguez-Flores ,JG Mezey, et al. (2012). RNA-Seq Quantification of the Human Small Airway Epithelium Transcriptome. *BMC genomics*, *13*(1), 82. BioMed Central Ltd. doi:10.1186/1471-2164-13-82
6. JP Jarvis., L Scheinfeldt.*, S Soi*, C Lambert*, L Omberg*, B Ferwerda, A Froment, et al. (2012). Patterns of Ancestry , Signatures of Natural Selection , and Genetic Association with Stature in Western African Pygmies. *PLoS Genetics*, *8*(4), e1002641. doi:10.1371/journal.pgen.1002641
7. AE Tilley, T O'Connor, N Hackett, Y Strulovici-Barel, J Salit, X Zhou, T Raman, L Omberg, A Clark, JG Mezey, RG Crystal (2011). Biologic Phenotyping of the Human Small Airway Epithelial Response to Cigarette Smoking. (M. Königshoff, Ed.)PLoS *ONE, 6(7),* e22798.
8. MW Butler, NR Hackett, J Salit, Y Strulovici-Barel, L Omberg, JG Mezey, RG Crystal (2011). Glutathione S- transferase copy number variation alters lung gene expression. *The European respiratory journal : official Journal of the European Society for Clinical Respiratory Physiology*.
9. Y Strulovici-Barel*, L Omberg*, M O'Mahony, et al. (2010) Threshold of Biologic Responses of the Small Airway Epithelium to Low Levels of Tobacco Smoke. *American journal of respiratory and critical care medicine.* 2010; (646).
10. L Omberg, JR Meyerson, K Kobayashi, LS Drury, JF Diffley, O Alter (2009). Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Molecular systems biology.* 2009;5:312.
11. R-H Hübner, JD Schwartz, P De Bishnu, B Ferris, L Omberg, JG Mezey, HR Hackett, RG Crystal (2009). Coordinate control of expression of Nrf2-modulated genes in the human small airway epithelium is highly responsive to cigarette smoking. *Molecular medicine (Cambridge, Mass.).* 2009;15(7-8):203-19.
12. L Omberg, GH Golub and O Alter (2007). A Tensor Higher-Order Singular Value Decomposition for Integrative Analysis of DNA Microarray Data From Different Studies *Proc. Natl. Acad. Sci. USA* 104 (47), pp. 18371– 18376 (November 2007).
13. L Omberg, K Dolan, A Neiman and F Moss (2000). Detecting the onset of bifurcations and their precursors from noisy data. *Phys. Rev. E* 61 (5), pp. 4848–4853 (May 2000).

Adam Arne Margolin, PhD
Director, Computational Biology, Sage Bionetworks, Seattle, USA
Email: margolin@sagebase.org

A. Education

University of Pennsylvania, Philadelphia	B.S.	05/02	Information systems
University of Pennsylvania, Philadelphia	M.S.	12/02	Computer science
Columbia University, New York	M. Phil.	02/06	Biomedical informatics
Columbia University, New York	Ph.D.	01/08	Biomedical informatics

B. Recent Professional Experience

2001 Developer, EGenomics, Inc., New York, NY
2002 Gene Expression Omnibus, Developer, National Center for Biotechnology Information, Bethesda, MD
2001-03 Bioinformatics Application Developer, University of Pennsylvania, Abramson Cancer Research Institute, Philadelphia, PA
2005, 06 Functional Genomics and Systems Biology Group, Intern, IBM T.J. Watson Research Center, Yorktown Heights, NY
2003-08 Department of Biomedical Informatics, Ph.D. Student, Columbia University, New York, NY
2008-10 Cancer Program, Postdoctoral Associate, The Broad Institute of Harvard and MIT, Cambridge, MA
2010-11 Group Leader, Genotype-Specific Therapeutics Initiative, The Broad Institute of Harvard and MIT, Cambridge, MA
2011- Director, Computational Biology, Sage Bionetworks, Seattle, WA

C. Recent Peer-reviewd Publications

1. Jang, I. S., Neto, E. C., Friend, S. H., Margolin, AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput (in press).
2. Neto, E. C., Jang, I. S., Friend, S. H., Margolin, AA. The Stream Algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications in high throughput predictive modeling of cancer cell-line pharmacogenomic screens. Pac Symp Biocomput (in press).
3. Omberg, L., Ellrott, K., Yuan, Y., Kandoth, K., Wong, C., The Cancer Genome Atlas Research Network, Friend, S.H., Stuart, J., Liang, H., Margolin, AA. Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. Nat Genet. Nat Genet. 2013;45(10):1121-6.
4. Margolin AA*, Bilal E*, Huang E*, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Science translational medicine. 2013;5(181):181re1. Epub 2013/04/19.
5. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, Sauerwine BA, Shimoni Y, Moen Vollan HK, Mecham BH, Rueda OM, Tost J, Curtis C, Alvarez MJ, Kristensen VN, Aparicio S, Borresen-Dale AL, Caldas C, Califano A, Friend SH, Ideker T, Schadt EE, Stolovitzky GA, Margolin AA. Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. PLoS computational biology. 2013;9(5):e1003047. Epub 2013/05/15.
6. Barretina J.*, Caponigro G.*, Stransky N.*, Venkatesan K.*, Margolin A.A.*, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. Epub 2012/03/31.
7. Wei G.*, Margolin A.A.*, Haery L., Brown E., Cucolo L., Julian B., et al. Chemical genomics identifies small-molecule MCL1 repressors and BCL-xL as a predictor of MCL1 dependency. Cancer cell. 2012;21(4):547-62. Epub 2012/04/21.
8. Margolin, A. A., K. Wang, A. Califano, I. Nemenman, Multivariate dependence and genetic networks inference. IET Syst Biol, 2011. 4(6): 428-40.
9. Margolin, A.A., S.E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, and T.R. Golub, Empirical Bayes analysis of quantitative proteomics experiments. PLoS One, 2009. 4(10): p. e7454.
10. Margolin, A.A., T. Palomero, P. Sumazin, A. Califano, A.A. Ferrando, and G. Stolovitzky, ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc Natl Acad Sci U S A, 2009. 106(1): p. 244-9.
11. Margolin, A.A. and A. Califano, Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci, 2007. 1115: p. 51-72.
12. Margolin, A.A., K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, and A. Califano, Reverse engineering cellular networks. Nat Protoc, 2006. 1(2): p. 662-71.

Rodrigo Dienstmann, M.D.
Research Scientist, Sage Bionetworks, Seattle, USA
Email: rodrigo.dienstmann@sagebase.org

Education:
2001 - School of Medicine – Universidade Federal do Rio Grande do Sul, Brazil.

Professional Experience:
Jan 2002 – Dec 2003: Intership Internal Medicine, Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil
Jan 2004 – Jan 2006: Fellowship Medical Oncology, Brazilian National Cancer Institute, Rio de Janeiro, Brazil
Feb 2006 – Dic 2009: Clinical Research in Oncology, Brazilian National Cancer Institute, Rio de Janeiro, Brazil
Feb 2010 – Dec 2012: Molecular Therapeutic Research Unit/Phase 1 Unit, Vall d´Hebron Institute of Oncology, Barcelona, Spain
Jan 2013 – Sep 2013: Molecular Pathology Lab, Massachusetts General Hospital, Boston, USA

Recent Publications:

The genomic medicine frontier in human solid tumors: prospects and challenges
Dienstmann R, Rodon J, Barretina J, Tabernero J.
J Clin Oncol 2013;31(15):1874-84.

Biomarker-driven patient selection for early clinical trials
Dienstmann R, Rodon J, Barretina J, Tabernero J.
Curr Opin Oncol 2013;25(3):305-12.

Development of PI3K inhibitors: lessons learned from early clinical trials.
Rodon J, Dienstmann R, Serra V, Tabernero J.
Nat Rev Clin Oncol 2013;10(3):143-53.

Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors
Dienstmann R, Rodon J, Prat A, Perez-Garcia J, Adamo B, Felip E, Cortes J, Iafrate AJ, Nuciforo P, Tabernero J
Annals Oncol 2013 Nov 20. [Epub ahead of print]

Molecular prescreening to select patient population in early clinical trials
Rodón J, Saura C, Dienstmann R, Vivancos A, Cajal SR, Baselga J, Tabernero J.
Nat Rev Clin Oncol 2012;9(6):359-66

Molecular profiling of patients with colorectal cancer and matched targeted therapy in Phase 1 clinical trials
Dienstmann R, Serpico D, Rodon J, Saura C, Macarulla T, Elez ME, Alsina M, Capdevila J, Perez-Garcia J, Sánchez-Ollé G, Aura C, Prudkin L, Landolfi S, Hernández-Losa J, Vivancos A, Tabernero J.
Mol Cancer Ther 2012;11(9):2062-71.

Risk-benefit assessment of bevacizumab in the treatment of breast cancer
Dienstmann R, Ades F, Saini KS, Metzger-Filho O.
Drug Saf 2012;35(1):15-25.

Drug development to overcome resistance to EGFR inhibitors in lung and colorectal cancer
Dienstmann R, De Dosso S, Felip E, Tabernero J.
Mol Oncol 2012;6(1):15-26.

BRAF as a target for cancer therapy
Dienstmann R, Tabernero J.
Anticancer Agents Med Chem 2011;11(3):285-95.

Molecular predictors of response to chemotherapy in colorectal cancer
Dienstmann R, Vilar E, Tabernero J.
Cancer J 2011;17(2):114-26.

Personalizing therapy with targeted agents in non-small cell lung cancer
Dienstmann R, Martinez P, Felip E.
Oncotarget 2011;2(3):165-77.

Combined modality therapy of stage IIIC breast cancer
Dienstmann R, Branco LG, Rezende LM, Freitas LC, Lima CF, Rodrigues GJ, Noronha Filho H, Sarmento RMB, Small I, Bines J.
Breast J 2011;17(3):331-3.

Toxicity as biomarker of efficacy of molecular targeted therapies: focus on EGFR and VEGFR inhibiting anticancer agents
Dienstmann R, Braña I, Rodon J, Tabernero J.
Oncologist 2011;16(12):1729-40

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

Statistical Inference of Tumor Heterogeneity Using WGS Data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (No more than 2; 1 page CV for each)

Yuan Ji, NorthShore University HealthSystem, UChicago; Kevin White. UChicago

Name(s) & institute(s) of junior investigators	**Name(s) & institute(s) of non-ICGC collaborators**
Yitan Zhu, NorthShore University HealthSystem, UChicago	Mark Gerstein, Yale University

Background and preliminary data

Mutations acquired during a tumor's life history characterize subclones of cells. A hallmark of the dynamic evolution of cancer is the presence of superceding clonal expansions, driven by shifting selective pressures, mutational processes, and disrupted oncogenes or tumor suppressors. These processes mark the genome of cancer cells in a tumor in a way such that each tumor's life history is imprinted in the somatic mutations that can be found in that tumor. New somatic mutations give rise to new cellular subpopulations called subclones. In this way, the described process of somatic mutations induces the observed tumor heterogeneity (TH). We propose statistical models to decipher this narrative and infer the resulting subclonal diversification within and between tumors, i.e., intra- and inter-TH. There is a critical gap in the current literature. Despite the recognized importance of TH, there are, to our knowledge, no effective computational or experimental methods to accurately reveal the subclonal structure of a given tumor. But there have been some related developments. Specifically, some progress has been made in detecting the cellular heterogeneity in terms of copy number variants or single nucleotide variants (SNVs) across different subclones [*Shah et al., 2012; Jiao et al., 2013; Strino et al., 2013; Oesper et al., 2013*]. However, understanding the cellular variations of haplotype sequences for subclones is still an open research problem [*Landau et al., 2013*; *Serena et al., 2012*], including our own work [*Lee et al., 2013*]. In principle, such inference is feasible with the recent introduction of next-generation sequencing (NGS) and whole-genome sequencing (WGS) technologies (Fig. 1). In particular, different cells can be marked by the sequence variations, and we can describe tumor subclonalities by haplotypes defined as a unique set of sequences across multiple loci.

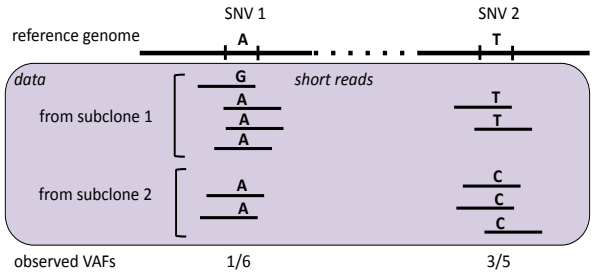


Fig. 1: Proposed inference on the estimation of tumor heterogeneity and mosaicism. The goal is to use WGS read-mapping information to infer the number and proportions of subclones in a tumor sample. The proposed method takes input as the observed variant allele fractions (VAFs) across multiple single nucleotide variants (SNVs) and outputs the subclonality and mosaicism for each sample.

In Lee *et al.* (2013a) we applied a Bayesian nonparametrics model (Lee et al., 2013b) to analyze simulated data as well as our own pilot DNA-Seq data. We use a finite Indian buffet process (IBP) model for a feature allocation matrix Z. The finite IBP serves as a prior, conditional on which we construct a sampling model for the WGS data. Specifically, for each SNV s in sample t, the observed data consist of two counts: N_{st}, the total number of short reads overlapping with the locus, and n_{st} the number of short reads among N_{st} that have a mutant sequence -- different from the reference sequence. For example, in Fig. 1 n_{st}= 1 and N_{st} = 6 for SNV 1. Given N_{st}, we assume a binomial sampling model, $n_{st} \sim \text{Bin}(N_{st}, p_{st})$. The parameter p_{st} refers to the mean VAFs which is the proportion of reads that possess a variant sequence relative to the reference sequence and is the key parameter in the model. We assume that the proportion p_{st} arises from the fact that sample t is a composition of (latent) subclones, some of which have the variant sequence at locus s. Let w_{tk} be the proportion of the (unknown) subclone k in sample t; we define that $\{z_{sk} = 1 \text{ or } 0\}$ the event that locus s of subclone k has a variant (reference) sequence. We propose $\boldsymbol{p_{st}} = \sum_{k=1}^{K} \boldsymbol{w_{tk} z_{sk}} + \boldsymbol{e_t}$. In other words, the proportion of variant reads are contributed from those subclones with the variant sequence, plus some

noise. We model noise $\boldsymbol{e_t} = \boldsymbol{w_{t0}p_0}$ to convey the notion that aside from the contribution from the subclones, a small proportion w_{t0} of reads could still possess the variant sequence due to systemic or experimental errors that occur with probability p_0. We assume that there are S SNVs and K subclones, and therefore $Z = [Z_{sk}]$ is a latent $(S \times K)$ binary matrix. We implement Markov chain Monte Carlo (MCMC) posterior simulation for fixed K, and then take a model selection perspective to select K and use cross validation to find an optimal K. Applying the model to WGS data for five cancer patients with pancreatic cancer and drastic different prognosis, we obtained some interesting preliminary results (Fig. 2). In summary, the figure seems indicate that while tumors are unique, there are subclones that do recur across different patients. This implies that SNV patterns observable in the genome might correlate with clinical data. This assumption is inherent in all tumor genome projects. The results also clearly show that each tumor is made of more than one subclone: usually two dominant subclones and other minor ones.

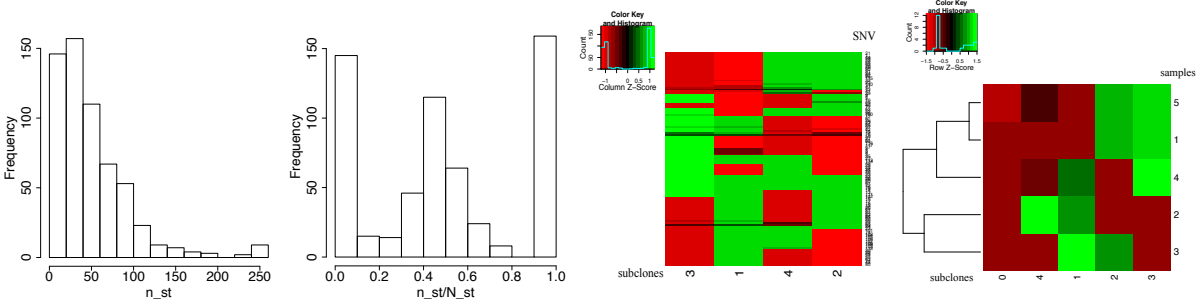


Fig2. Finite IBP Inference. The left two histograms summarize the in-house WGS read-count data in our analysis. The third plot (heatmap) shows the estimated binary matrix Z the columns of which are inferred subclones, and the last plot (heatmap) shows the estimated proportions of the subclones in five patient samples.

Timelines & resources dedicated to project

We will apply and extend the statistical models in Lee et al. (2013a) to analyze ICGC WGS data. The analysis will be completed in 2014. We have access to multiple high-performance computing clusters, located at The University of Chicago, NorthShore University HealthSystem, and Argonne National Laboratory, with over 20,000 CPUs, and petabytes of storage. We will publish all the analysis results and make available analysis tools.

Research proposal

We propose to deconvolute observed variant allele fractions (VAFs) using feature allocation models. We will estimate subclones in tumor samples and how samples are composed of different proportions of these subclones. The latter is known as mosaicism. The data are observed VAFs at each locus, defined as the proportion of short reads bearing a mutant sequence (Fig. 1, left). The proposed inference is based on feature allocation models [*Broderick et al., 2013*], and produces the number of subclones, defined as a set of unique haplotype sequences at SNVs. In Fig. 1, two SNVs are illustrated with two hypothetical sublcones. With more SNVs, more subclones might be inferred, describing the heterogeneity of the tumor sample. We use variations of the Indian buffet process (IBP) to define subclones as distinct sets of mutational incidences at these loci. In short, different subclones are the features and SNVs are the experimental units that select or do not select these features. A mixture of these hypothesized latent subclones is used to fit the observed VAFs. We develop models for TH to account for biological dependence and to exploit other data modalities, such as copy numbers. Besides simulation studies, we plan a cell line experiment for experimental verification of the proposed approach, collaborating with Dr. Gulukota Kamalakar at NorthShore University HealthSystem. We consider three more applications, including a study on intra-TH, a study with ICGC data, and a clinical study. For the latter we relate inferred TH with clinical characteristics to define patient subpopulations. Proposed work can be summarized in three steps given below:

- *Feature allocation models for inference on TH based on WGS data.*
- *Integration of multi-modality data (such as copy number variations) to enhance inference on TH.*
- *Validation and applications based on ICGC data and in-house pilot and large-scale cohort data at NorthShore University HealthSystem.*

Legacy plans

All generated results of the suggested studies will be published in peer-reviewed journals. Tools and software created during the course of the suggest studies will be publicly available. We have done this with extensive experiences before. A recent example can be found at link below for TCGA data.

http://health.bsd.uchicago.edu/yji/TCGA-Assembler.htm

References

Broderick, T., Jordan, M. I., and Pitman, J. (2013). Clusters and features from combinatorial stochastic processes. Statistical Science to appear.

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., Morris, Q. (2013). Inferring clonal evolution of tumors from single nucleotide somatic mutations. arXiv:1210.3384, http://arxiv.org/abs/1210.3384

Landau, D., Carter, S., Stojanov, P., Aaron, M., Stevenson, K., Lawrence, M., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S., Vartanov, A., Fernandes, S., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E., Neuberg, D., Brown, J., Getz, G., and Wu, C. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 152, 4, 714–26.

Lee, J., Mueller, P., Ji, Y., and Gulukota, K. (2013a). A Bayesian Feature Allocation Model for Tumor Heterogeneity. Tech. rep., UC Santa Cruz.

Lee, J., Mueller, P., Zhu, Y., and Ji, Y. (2013b). A nonparametric Bayesian model for local clustering with Application to Proteomics. *Journal of the American Statistical Association* **108**, 775–788.

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome biology 14, 7, R80.

Serena, N., Van Loo, P., Wedge, D., Alexandrov, L., Greenman, C., Lau, K., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S., Hinton, J., Menzies, A., Stebbings, L., Leroy, C., Jia, M., Rance, R., Mudie, L., Gamble, S., Stephens, P., Stuart, M., Tarpey, P., Papaemmanuil, E., Davies, H., Varela, I., David, M., Bignell, G., Leung, K., Butler, A., Teague, J., Martin, S., J¨onsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerd, A., Aparicio, S., Tutt, A., Sieuwerts, A., Borg, r., Thomas, G., Salomon, A., Richardson, A., Anne-Lise, B., Futreal, P., Stratton, M., Campbell, P., and of the International Cancer Genome Consortium, B. C. W. G. (2012). The life history of 21 breast cancers. Cell 149, 5, 994–991007.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L. M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J., Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S. K., Griffith, M., Moradian, A., Cheng, S. W., Morin, G. B., Watson, P., Gelmon, K., Chia, S., Chin, S. F., Curtis, C., Rueda, O. M., Pharoah, P. D., Damaraju, S., Mackey, J., Hoon, K., Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J. F., Meyer, I. M., Eaves, C. J., Wasserman, W. W., Jones, S., Huntsman, D., Hirst, M., Caldas, C., Marra, M. A., Aparicio, S. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature, 486(7403):395–9. doi: 10.1038/nature10933.

Strino, F., Parisi, F., Micsinai, M., and Kluger, Y. (2013). TrAp: a tree approach for fingerprinting subclonal tumor composition. Nucleic Acids Research, 41(17), Pp. e165.

Yuan Ji, PhD

Director of Biomedical Informatics
Center for Clinical and Research Informatics
NorthShore University HealthSystem
1001 University Place
Evanston, IL 60201
Phone: (224) 364-7312
Fax: (773) 834-4321

Associate Professor (part time)
Department of Health Studies
The University of Chicago
5841 S. Maryland Ave., MC
Chicago, IL 60637
Phone: (773) 834-0214
Email: yji@health.bsd.uchicago.edu
Web page: http://health.bsd.uchicago.edu/yji/

Positions and Employment

2003-2006	Assistant Professor, Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX
2006-2009	Assistant Professor, Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas M. D. Anderson Cancer Center, Houston, TX
2009-2012	Associate Professor, Department of Biostatistics, Division of Quantitative Sciences, The University of Texas M.D. Anderson Cancer Center, Houston, TX
2012-present	Director of Cancer Informatics, Center for Clinical and Research Informatics, NorthShore University HealthSystem, Evanston, IL
2013-present	Associate Professor (part time), Department of Health Studies, The University of Chicago, Chicago, IL

Selected Recent Peer-Reviewed Publications
(Publications selected from 63 peer-reviewed publications; * -- corresponding author)

1. Lee J, Mueller P, Zhu Y, **Ji Y***. A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association*. 108, 775-788, 2013.
2. Mitra R, Mueller P*, Liang S, Yue L, **Ji Y***. A Bayesian graphical model for ChIP-Seq data on histone modifications. *Journal of the American Statistical Association* 108(501):69-80, 2013.
3. Hu B, **Ji Y***, Xu Y, Ting AH. Screening for SNPs with allele-specific methylation based on next-generation sequencing data. *Stat. Biosci.* 5(1):179-197, 2013.
4. **Ji Y**, Wang SJ. A safer and more reliable method than the 3+3 design for practical phase I trials. *Journal of Clinical Oncology*. 31(14):1785-91, 2013.
5. Mitra R, Mueller P, Shoudan L, Xu Y, **Ji Y***. Towards breaking the histone code – Bayesian graphical models for histone modifications. *Circulation: Cardiovascular Genetics*. 6(4):419-426, 2013.
6. Telesca D*, Mueller P, Kornblau S, **Ji Y***. Modeling protein expression and protein signaling pathways. *Journal of the American Statistical Association* 107(500):1372-1384, 2012.
7. Xu Y, Zhang J, Yuan Y, Mitra R, Mueller P, **Ji Y***. A Bayesian Graphical Model for Integrative Analysis of TCGA Data. In *2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), 135-138, 2012*.
8. **Ji Y**, Xu Y, Zhang Q, Tsui KW, Yuan Y, Liang S, Liang H*. BM-map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics* 67(4):1215-1224, 2011.
9. **Ji Y***, Xu Y, Zhang Q, Tsui KW, Yuan Y, Liang S, Liang H*. BM-Map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics*. 67(4):1215-24, 2011
10. Baladandayuthapani V, **Ji Y**, Morris J, Talluri R, Nieto-Barajas L. Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data. *Journal of the American Statistical Association* 105(492):1358-1375, 2010

Current NIH Funding as PI (also PI for a private foundation grant)

1R01CA132897 Ji (PI) 9/15/2008-7/31/2014
NIH/NCI
Bayesian Models for Cancer Prognosis by Integrating Diverse Types of Data
The long-range goal of this application is to improve risk predication, treatment selection, and subtype classification in cancer prevention, diagnosis, and prognosis.

Kevin P. White

Education

Yale University, New Haven, B.S./M.S., Biology 1993
Stanford University, Stanford, CA, Ph.D., Developmental Biology 1998
Stanford Genome Technology Ctr, Palto Alto, CA, Postdoc, Biochemistry & Genomics, 1998-2000

Professional Positions

2006-present	Director, Joint Institute for Genomics & Systems Biology, The University of Chicago and Argonne National Laboratory
2006-present	James and Karen Frank Family Professor, Human Genetics and Ecology & Evolution, The University of Chicago
2004-2006	Associate Prof. of Ecology & Evolutionary Biology (joint appointment), Yale University
2004-2006	Associate Professor of Genetics, Yale University School of Medicine
2001-2004	Assistant Professor of Genetics, Yale University School of Medicine

Publications Selected from 97 peer-reviewed publications

1. Michelle N. Arbeitman, Eileen E. M. Furlong, Farhad Imam, Eric Johnson, Brian H. Null, Bruce S. Baker, Mark A. Krasnow, Matthew P. Scott, Ronald W. Davis and Kevin P. White. Gene Expression During the Life Cycle of Drosophila melanogaster. **Science,** 297: 2270-2275, **2002.**
2. Giot L, Bader JS, Brouwer C, Chaudhuri, et al. A genome-scale protein interaction map of Drosophila melanogaster. **Science**, 302: 1727-36, **2003.**
3. Viktor Stolc*, Zareen Gauhar*, Christopher Mason*, Gabor Halasz, Marinus F. van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Barbano, Harmen J. Bussemaker, and Kevin P White. A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster. **Science**, 306:655-60, **2004.**
4. Scott Rifkin, David Houle, Junhyong Kim and Kevin P. White. A mutation accumulation assay reveals extensive capacity for rapid gene expression evolution. **Nature,** 438:220-3, **2005.**
5. Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed and Kevin P. White. "Expression profiling in primates reveals a rapid evolution of human transcription factors." **Nature,** 440:242-5, **2006**.
6. Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Nègre N, Ludwig M, Stricker T, Al-Ahmadie HA, Tretiakova M, Camp RL, Perera-Alberto M, Rimm DL, Xu T, Rzhetsky A, White KP. Analysis of Drosophila Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer. **Science,** 323:1218-22, **2009**.
7. Hua SJ, Kittler R, and White KP. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. **Cell.** 137:1259-71, **2009**.
8. modENCODE Consortium, et, al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. **Science**. 330:1787-97. **2010**
9. Nègre N*, Brown CD*, Ma L*, Bristow CA*, Miller S*, Kheradpour P, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony R, Chen J, Hwang L, Wagner U, Auburn R, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Bild NA, Grundstad AJ, Hanley D, Mannervik M, Venken K, Bellen H, White R, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. A cis-regulatory map for the Drosophila genome. **Nature**. 471:527-31. **2011**.
10. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. **Nature.**489:57-74. 2012.:
11. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumors, **Nature.** 490:61-70. 2012
12. Xiaochun Ni, Yong E. Zhang, Nicolas Negre, Sidi Chen, Manyuan Long and Kevin P. White. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. **PLoS. Biology**. 10(11):e1001420. 2012.
13. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. **Blood**. 121: 975-83. 2013.
14. Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP. A comprehensive nuclear receptor network for breast cancer cells. **Cell Rep.** 3:538-51. 2013.
15. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. **Cell**. Sep 26;155:70-80. 2013.

Mark Gerstein

Education

Harvard College, AB Physics '89
Cambridge University, PhD Chemistry '93
Stanford University, postdoc '93-'96, Bioinformatics (advisor M Levitt)

Positions

2006- **AL Williams Prof. Biomedical Informatics, Yale**
2002- co-director Yale Computational Biology and Bioinformatics Program
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

Honors

'89-'93 Herchel-Smith Scholarship for PhD at Cambridge
'93-'96 Damon Runyon-Walter Winchell post-doctoral Fellowship
'09 AAAS Fellow

Consortia

Analysis co-chair: NHGRI modENCODE Project AWG ('07-), Brainspan Project ('09-), 1000 Genomes Functional Interpretation Group ('12-), ENCODE & Cancer Group ('13-) exRNA consortium ('13-)

Publications (senior author on all papers listed below, which are selected from a total of >460; H-index=116)

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, M Gerstein (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342:1235587

E Khurana, Y Fu, J Chen, M Gerstein (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comp Bio* 9:e1002886.

M Gerstein, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." *Nature* 489:91

A Abyzov, J Mariani... (16 authors)... M Gerstein, FM Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells." *Nature* 492:438

B Pei, C Sisu... (10 authors)... J Harrow, M Gerstein (2012). "The GENCODE pseudogene resource." *Genome Biol* 13:R51.

C Cheng, R Alexander... (16 authors)... M Gerstein (2012). " Understanding transcriptional regulation by integrative analysis of transcription factor binding data." *Genome Res* 22:1658.

DG MacArthur, S Balasubramanian... (50 authors)... M Gerstein, C Tyler-Smith (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." *Science* 335:823.

A Abyzov, AE Urban, M Snyder, M Gerstein (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21:974

A Sboner, L Habegger... (9 authors)... MA Rubin, M Gerstein (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." *Genome Biol* 11:R104.

HY Lam, XJ Mu, AM Stütz, A Tanzer, PD Cayting, M Snyder, PM Kim, JO Korbel, M Gerstein (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." *Nat Biotech* 28:47.

RP Alexander, G Fang, J Rozowsky, M Snyder, M Gerstein (2010). "Annotating non-coding regions of the genome." *Nat Rev Genet* 11:559.

KK Yan, G Fang, N Bhardwaj, RP Alexander, M Gerstein (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107:9186.

N Bhardwaj, KK Yan, M Gerstein (2010). "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels." *PNAS* 107:6841

M Gerstein, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330:1775.

Abstract of proposed research for WGS pan-cancer analysis

Title of abstract

Graphical Statistical Models for Integrating Multiple Genomics Characterizations Using ICGC Data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (No more than 2; 1 page CV for each)

Yuan Ji, NorthShore University HealthSystem**,** UChicago; Kevin White. UChicago

Name(s) & institute(s) of junior investigators	**Name(s) & institute(s) of non-ICGC collaborators**
Yitan Zhu, NorthShore University HealthSystem, UChicago	Mark Gerstein, Yale University

Background and preliminary data

Cancer is rarely caused by an abnormality in a single gene, but rather reflects perturbations to intracellular molecular interaction networks that attract cells to new malignant and carcinogenic states [*Barabasi et al., 2011*; *Creixell et al., 2012*]. Identifying the genomic interaction networks and understanding their behaviors in cancer conditions are critical to the elucidation of cancer molecular mechanisms and the development of cancer treatment [*Cancer Target Discovery, Development Network, 2010*].

ICGC provides comprehensive data measuring whole-genome genomics and epigenomics features of thousands of tumor samples for more than 20 types of cancer. The features recorded by ICGC include gene expression, DNA methylation, DNA copy number, and more. Such a wealth of information enables the study on interaction networks of these features for a systemic investigation of molecular compositions and dynamics of cancer. A core challenge is to develop effective tools that integrate the multi-platform and multi-disease data and systematically screen the analysis results in hopes of novel discoveries that would not have been available by only looking at data from a single feature.

Our group has extensive experience in integration of multi-platform genomic data and investigation on genomic regulation networks. We developed **Bayesian Graphical Models** (BGMs) to identify the dependence structure between different histone modification features [*Mitra et al., 2013a*] and to infer molecular interactions between multiple genomics and epigenomics features in cancer [*Xu et al., 2012; Mitra et al., 2013b*]. The **distinct features** of the proposed BGMs are twofold, which set our models apart from existing ones. First, our BGMs combine prior knowledge with data to conduct posterior inference on the network. This allows for efficient learning and inference on large networks. Second, our BGMs provide full probabilistic posterior inferences allowing for automatic adjustment of multiplicity. For example, the proposed BGMs generate posterior probabilities on the entire network, any sub-networks, and any edges in the network, with which false discovery rates can be easily estimated for any significance calculation. Our preliminary analysis using BGMs has shown promising results of characterizing molecular regulations and interactions between genomics characterizations in cancer. For example, gene pairs inferred to have a positive co-expression pattern are indeed statistically significantly enriched with genes that have known transcriptional activation relationship or belong to the same functional gene module (manuscript in preparation).

Timelines & resources dedicated to project

Data collection and preparation will be performed in Jan-Feb 2014. Computational analysis will be conducted in Feb-Oct 2014. Web portal and database will be constructed in Sept-Dec 2014. Manuscript preparation is planned in Jan-Feb 2015. We have access to multiple high-performance computing clusters, located at The University of Chicago, NorthShore University HealthSystem, and Argonne National Laboratory, with over 20,000 CPUs, and petabytes of storage.

Research proposal

We propose to develop **Zodiac**, a public web-based information system of cancer genomic

interactions. It will host all the significant interactions inferred by the proposed BGMs in at least 10 cancer types. Fig. 1 (top) summarizes the roadmap of building Zodiac, including the planned analyses, the software components, and web interface. For each cancer type, we plan to apply the BGMs to infer intragenic interactions within each gene and intergenic interactions between each pair of genes (Fig. 1

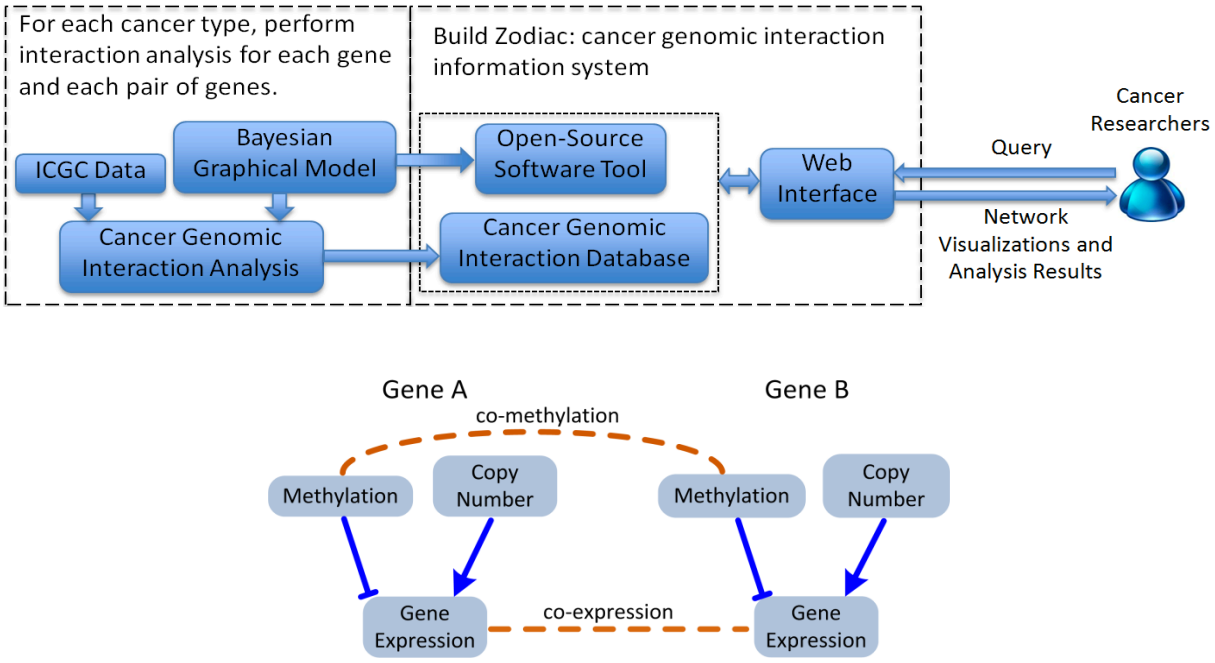


Fig. 1. Top. Roadmap of the proposed research project: Zodiac, a comprehensive depiction of genomic interactions in cancer. **Bottom.** An illustration of hypothetical genomic interactions within a gene and between a pair of genes that can be inferred by BGMs, shown as edges in the network. Blue solid lines indicate intragenic interactions; orange dashed lines indicate intergenic interactions.

bottom). The inferred interactions can serve as evidence and confirmation of various genomic mechanisms occurring in the specific cancer conditions, such as transcriptional activation by transcription factor, gene expression enhancing/repression by copy number variations or methylation variations, co-expression between genes, co-methylation between genes [*Akulenko et al., 2013*], etc. Specifically, we will perform the following analyses and steps:

- *Prepare a pipeline for retrieving and preparing sample-matched, multi-platform ICGC data of at least 10 cancer types. This pipeline will be developed in a similar fashion as our in-house software, TCGA-Assembler (*http://health.bsd.uchicago.edu/yji/TCGA-Assembler.htm)
- *For each gene in each cancer type, construct an intragenic interaction network including the multiple genomic features of the gene. For each pair of genes in each cancer type, use BGMs to model the interactions between different genomic features of the two genes (Fig. 1, right). Existing knowledge about genomic interactions, such as transcription factor binding from TRANSFAC [Wingender et al., 2000], will be used as probabilistic prior in the model.*
- *After inferring genomic interactions in each cancer type, we will compare the genomic interactions between cancer types to discover the similarity and dissimilarity of regulation mechanisms between cancer types.*
- *Construct a database to host all the identified genomic interactions in different cancer types and the changes of interactions over cancer types. Build a web portal for public access to the database.*

We will expand the analysis to include more than two genes in the future but will first test the feasibility of the proposal with two genes; the proposed analysis requires running a huge number of computational jobs in parallel, in the order of magnitude of $2\times10^{4*k}$, where k is the number of genes included in each network.

Legacy plans
All generated results of the suggested studies will be published in peer-reviewed journals. Tools and software created during the course of the suggest studies will be publicly available.

References

Akulenko, R., and Helms, V. (2013). DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. Human Molecular Genetics, Aug 1; 22(15):3016-22. doi: 10.1093.

Barabasi, A., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 12(1):56–68.

Cancer Target Discovery, Development Network (2010). Towards patient-based cancer therapeutics. Nature biotechnology, 28(9):904–6.

Creixell, P., Schoof, E., Erler, J., and Linding, R. (2012). Navigating cancer network attractors for tumorspecific therapy. Nature biotechnology, 30(9):842–8.

Mitra, R., Muller, P., Liang, S., Yue, L., and Ji, Y. (2013a). A bayesian graphical model for ChIP-Seq data on histone modifications. Journal of the American Statistical Association, 108:69–80.

Mitra, R., Mueller, P., Liang, S., Xu, Y., and Ji, Y. (2013b). Towards Breaking the Histone Code – Bayesian Graphical Models for Histone Modifications.Circulation: Cardiovascular Genetics. 2013 Aug;6(4):419-26. doi: 10.1161

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic acids research, 28(1):316–9.

Xu, Y., Zhang, J., Yuan, Y., Mitra, R., Muller, P., Ji, Y. (2012). A Bayesian graphical model for integrative analysis of TCGA data, IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), pp135–8, Washington, DC, 2-4 Dec.

Yuan Ji, PhD

Director of Biomedical Informatics
Center for Clinical and Research Informatics
NorthShore University HealthSystem
1001 University Place
Evanston, IL 60201
Phone: (224) 364-7312
Fax: (773) 834-4321

Associate Professor (part time)
Department of Health Studies
The University of Chicago
5841 S. Maryland Ave., MC
Chicago, IL 60637
Phone: (773) 834-0214
Email: yji@health.bsd.uchicago.edu
Web page: http://health.bsd.uchicago.edu/yji/

Positions and Employment

2003-2006	Assistant Professor, Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX
2006-2009	Assistant Professor, Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas M. D. Anderson Cancer Center, Houston, TX
2009-2012	Associate Professor, Department of Biostatistics, Division of Quantitative Sciences, The University of Texas M.D. Anderson Cancer Center, Houston, TX
2012-present	Director of Cancer Informatics, Center for Clinical and Research Informatics, NorthShore University HealthSystem, Evanston, IL
2013-present	Associate Professor (part time), Department of Health Studies, The University of Chicago, Chicago, IL

Selected Recent Peer-Reviewed Publications
(Publications selected from 63 peer-reviewed publications; * -- corresponding author)

1. Lee J, Mueller P, Zhu Y, **Ji Y***. A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association*. 108, 775-788, 2013.
2. Mitra R, Mueller P*, Liang S, Yue L, **Ji Y***. A Bayesian graphical model for ChIP-Seq data on histone modifications. *Journal of the American Statistical Association* 108(501):69-80, 2013.
3. Hu B, **Ji Y***, Xu Y, Ting AH. Screening for SNPs with allele-specific methylation based on next-generation sequencing data. *Stat. Biosci.* 5(1):179-197, 2013.
4. **Ji Y**, Wang SJ. A safer and more reliable method than the 3+3 design for practical phase I trials. *Journal of Clinical Oncology*. 31(14):1785-91, 2013.
5. Mitra R, Mueller P, Shoudan L, Xu Y, **Ji Y***. Towards breaking the histone code – Bayesian graphical models for histone modifications. *Circulation: Cardiovascular Genetics*. 6(4):419-426, 2013.
6. Telesca D*, Mueller P, Kornblau S, **Ji Y***. Modeling protein expression and protein signaling pathways. *Journal of the American Statistical Association* 107(500):1372-1384, 2012.
7. Xu Y, Zhang J, Yuan Y, Mitra R, Mueller P, **Ji Y***. A Bayesian Graphical Model for Integrative Analysis of TCGA Data. In *2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), 135-138, 2012*.
8. **Ji Y**, Xu Y, Zhang Q, Tsui KW, Yuan Y, Liang S, Liang H*. BM-map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics* 67(4):1215-1224, 2011.
9. **Ji Y***, Xu Y, Zhang Q, Tsui KW, Yuan Y, Liang S, Liang H*. BM-Map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics*. 67(4):1215-24, 2011
10. Baladandayuthapani V, **Ji Y**, Morris J, Talluri R, Nieto-Barajas L. Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data. *Journal of the American Statistical Association* 105(492):1358-1375, 2010

Current NIH Funding as PI (also PI for a private foundation grant)

1R01CA132897 Ji (PI) 9/15/2008-7/31/2014
NIH/NCI
Bayesian Models for Cancer Prognosis by Integrating Diverse Types of Data
The long-range goal of this application is to improve risk predication, treatment selection, and subtype classification in cancer prevention, diagnosis, and prognosis.

Kevin P. White

Education

Yale University, New Haven, B.S./M.S., Biology 1993
Stanford University, Stanford, CA, Ph.D., Developmental Biology 1998
Stanford Genome Technology Ctr, Palto Alto, CA, Postdoc, Biochemistry & Genomics, 1998-2000

Professional Positions

2006-present	Director, Joint Institute for Genomics & Systems Biology, The University of Chicago and Argonne National Laboratory
2006-present	James and Karen Frank Family Professor, Human Genetics and Ecology & Evolution, The University of Chicago
2004-2006	Associate Prof. of Ecology & Evolutionary Biology (joint appointment), Yale University
2004-2006	Associate Professor of Genetics, Yale University School of Medicine
2001-2004	Assistant Professor of Genetics, Yale University School of Medicine

Publications Selected from 97 peer-reviewed publications

1. Michelle N. Arbeitman, Eileen E. M. Furlong, Farhad Imam, Eric Johnson, Brian H. Null, Bruce S. Baker, Mark A. Krasnow, Matthew P. Scott, Ronald W. Davis and Kevin P. White. Gene Expression During the Life Cycle of Drosophila melanogaster. **Science,** 297: 2270-2275, **2002.**
2. Giot L, Bader JS, Brouwer C, Chaudhuri, et al. A genome-scale protein interaction map of Drosophila melanogaster. **Science**, 302: 1727-36, **2003.**
3. Viktor Stolc*, Zareen Gauhar*, Christopher Mason*, Gabor Halasz, Marinus F. van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Barbano, Harmen J. Bussemaker, and Kevin P White. A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster. **Science**, 306:655-60, **2004.**
4. Scott Rifkin, David Houle, Junhyong Kim and Kevin P. White. A mutation accumulation assay reveals extensive capacity for rapid gene expression evolution. **Nature,** 438:220-3, **2005.**
5. Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed and Kevin P. White. "Expression profiling in primates reveals a rapid evolution of human transcription factors." **Nature,** 440:242-5, **2006**.
6. Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Nègre N, Ludwig M, Stricker T, Al-Ahmadie HA, Tretiakova M, Camp RL, Perera-Alberto M, Rimm DL, Xu T, Rzhetsky A, White KP. Analysis of Drosophila Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer. **Science,** 323:1218-22, **2009**.
7. Hua SJ, Kittler R, and White KP. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. **Cell.** 137:1259-71, **2009**.
8. modENCODE Consortium, et, al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. **Science**. 330:1787-97. **2010**
9. Nègre N*, Brown CD*, Ma L*, Bristow CA*, Miller S*, Kheradpour P, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony R, Chen J, Hwang L, Wagner U, Auburn R, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Bild NA, Grundstad AJ, Hanley D, Mannervik M, Venken K, Bellen H, White R, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. A cis-regulatory map for the Drosophila genome. **Nature**. 471:527-31. **2011**.
10. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. **Nature.**489:57-74. 2012**.**:
11. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumors, **Nature.** 490:61-70. 2012
12. Xiaochun Ni, Yong E. Zhang, Nicolas Negre, Sidi Chen, Manyuan Long and Kevin P. White. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. **PLoS. Biology**. 10(11):e1001420. 2012.
13. McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. **Blood**. 121: 975-83. 2013.
14. Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP. A comprehensive nuclear receptor network for breast cancer cells. **Cell Rep.** 3:538-51. 2013.
15. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. **Cell**. Sep 26;155:70-80. 2013.

Mark Gerstein

Education

Harvard College, AB Physics '89
Cambridge University, PhD Chemistry '93
Stanford University, postdoc '93-'96, Bioinformatics (advisor M Levitt)

Positions

2006- **AL Williams Prof. Biomedical Informatics, Yale**
2002- co-director Yale Computational Biology and Bioinformatics Program
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

Honors

'89-'93 Herchel-Smith Scholarship for PhD at Cambridge
'93-'96 Damon Runyon-Walter Winchell post-doctoral Fellowship
'09 AAAS Fellow

Consortia

Analysis co-chair: NHGRI modENCODE Project AWG ('07-), Brainspan Project ('09-), 1000 Genomes Functional Interpretation Group ('12-), ENCODE & Cancer Group ('13-) exRNA consortium ('13-)

Publications (senior author on all papers listed below, which are selected from a total of >460; H-index=116)

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, M Gerstein (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342:1235587

E Khurana, Y Fu, J Chen, M Gerstein (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comp Bio* 9:e1002886.

M Gerstein, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." *Nature* 489:91

A Abyzov, J Mariani... (16 authors)... M Gerstein, FM Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells." *Nature* 492:438

B Pei, C Sisu... (10 authors)... J Harrow, M Gerstein (2012). "The GENCODE pseudogene resource." *Genome Biol* 13:R51.

C Cheng, R Alexander... (16 authors)... M Gerstein (2012). " Understanding transcriptional regulation by integrative analysis of transcription factor binding data." *Genome Res* 22:1658.

DG MacArthur, S Balasubramanian... (50 authors)... M Gerstein, C Tyler-Smith (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." *Science* 335:823.

A Abyzov, AE Urban, M Snyder, M Gerstein (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21:974

A Sboner, L Habegger... (9 authors)... MA Rubin, M Gerstein (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." *Genome Biol* 11:R104.

HY Lam, XJ Mu, AM Stütz, A Tanzer, PD Cayting, M Snyder, PM Kim, JO Korbel, M Gerstein (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." *Nat Biotech* 28:47.

RP Alexander, G Fang, J Rozowsky, M Snyder, M Gerstein (2010). "Annotating non-coding regions of the genome." *Nat Rev Genet* 11:559.

KK Yan, G Fang, N Bhardwaj, RP Alexander, M Gerstein (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107:9186.

N Bhardwaj, KK Yan, M Gerstein (2010). "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels." *PNAS* 107:6841

M Gerstein, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330:1775.

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
The impact of human retrotransposons on cancer	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Haig H. Kazazian, Jr.: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine Geoffrey J. Faulkner: Mater Medical Research Institute, South Brisbane, Australia	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Szilvia Solyom: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine Adam D. Ewing: University of California, Santa Cruz	
Background and preliminary data	

Somatic retrotransposon mobilization in the cancer genome has only recently been established as a new mutational phenomenon in a number of tissues including colon, liver, pancreas, and lung, discovered through projects associated with TGCA and our own work external to that effort. The predominant mutagens are apparently Long INterspersed Element-1 (L1) elements, which are autonomous mobile elements that comprise 17% of the human genome. We investigate the timing of insertional events and the extent of tumor heterogeneity conferred by these retrotransposons, as well as their impact on malignancy. Previously, we studied DNA from 4 colon cancer patients diagnosed with colonic polyps (3 adenomas and one hyperplastic), and DNA from 5 additional patients with colorectal dysplasia arising in inflammatory bowel disease (IBD), as well as from 7 patients with primary pancreatic cancer and metastases. In contrast to the paired polyp-cancer samples, the IBD cancers were immediately adjacent to, and likely originated from, their matched dysplasias. Two of the 4 patients with both colon adenomas and carcinomas also had metastases. After dissection of abnormal from normal tissue, next-generation L1-targeted resequencing (L1-seq) was carried out on DNA from these tissues. After PCR-validation and Sanger sequencing of putative somatic L1 insertions, we found for the first time that certain pre-cancerous lesions were mutagenized by somatic L1 insertions. We validated in total 74 somatic insertions in cancerous or pre-cancerous lesions of the colon, of which 31 occurred in pre-malignant lesions. We found only one insertion in normal colon. Surprisingly, 2 adenomas contained more than 10 insertions each. Two IBD dysplasias contained insertions as well, and 7 L1 insertions were present not only in these lesions, but also in their paired cancers. Finally, we validated 2 metastasis-specific insertions and 11 insertions found in both primary colon cancers and their matched metastases. Since 11 of 12 insertions in colon cancers were present in their matched metastases, and 7 of 11 insertions in IBD dysplasias were also present in their paired carcinomas, most insertions may be present in every cell of the cancer or the dysplasia. Regarding pancreatic cancer, we have validated 17 insertions, 9 of which were present in both sections of the primary cancer and in the metastasis, while 7 insertions were specific to the metastasis. These numbers are only representative examples and after extrapolation of the number of insertions from high and low stringency data sets, we expect that some tumors' somatic L1 burden will be in the hundreds. At the same time, the number of other retroelement insertions mobilized by L1s, such as Alus, SVAs, and processed pseudogenes in these tissues is unknown. Current results show clonal distribution of somatic L1s and therefore indicate that L1s may serve as novel biomarkers of neoplastic disease progression. Numerous genes – some with unknown function – were targeted by L1s in our samples and may play a role in malignancy. We have found both intronic and some exonic insertions. To conclude, somatic retrotransposition is involved early in the pathogenesis of some gastrointestinal cancers and may provide useful biomarkers of neoplastic progression. Data in this abstract are all new and do not include data on insertions in colon cancer published by Solyom et al. (Genome Research 2012).

The 2000 tumors included in the pan-cancer data would enable us to calculate the frequency with which L1s, Alus, SVAs, and processed pseudogenes mobilize in the cancer genome with unprecedented accuracy, providing a solid answer to the question of which tissues of origin support high levels of retrotransposition quantitatively. Additionally, although human endogenous retroviruses (HERVs) are considered to be immobile retrotransposons, we would also like to ascertain whether they may be able to mobilize in certain cancers. Such a large compendium of tumor-specific retrotransposition events along with integration of functional expression and epigenomic data available through TGCA and ICGC may indicate whether the somatic mobilization of L1s and the other retroelements is a cause or an effect of malignancy. Given our current results that indicate the majority of L1 insertions occur very early, they may not be simply cancer-specific passengers, especially because these insertions are much more disruptive than point mutations. Therefore, we wish to address the whole retrotransposon landscape of cancers and adjacent normal tissues both at the DNA and RNA level, to identify driver insertions, and to correlate insertion activity with the mutation and epimutation signatures of the cancers, as well as with clinical data. If resources permit, we would also like to conduct functional studies in patient-derived cell lines.

Timelines & resources dedicated to project

Aim 1: Mapping of new retrotransposon integration sites using WGS data available through TCGA/ICGC.

Aim 2: Study somatic retrotransposon insertion-induced tumor heterogeneity/clonality. Assess the timing of retrotransposition events with respect to molecular markers of tumor clonal evolution and clinical parameters.

Aim 3: Functional studies in cell lines.

Research proposal
Aim 1: Mapping of new L1, Alu, SVA, processed pseudogene, and HERV integration sites in primary cancers and paired normal tissue DNA and RNA. Samples with matched metastasis and/or potential preneoplastic lesions – if available – would be of high priority. Retroelement insertions have been detected only in epithelial cancers, but it is not known, if, for instance a non-epithelial cancer sends metastasis into an epithelial tissue, whether that tissue environment may facilitate retrotransposon mobilization. Furthermore, non-clonal, unique somatic retrotransposition events may occur in non-epithelial tumors. Therefore, in-depth analysis of putative subclonal insertions from all types of tumors from the pan-cancer data set would be an asset. We note that a number of computational tools are available for the detection of transposable element insertions but they have not been compared to one another to ascertain how much their predictions overlap and how many predictions are discordant among the various methods. A sub-aim of this proposal is to carry out such a comparison of methods and develop 'best-of-breed' transposable element detection software to comprehensively ascertain insertion sites in a patient-matched tumor/normal context. We have developed L1-seq and RC-seq in house to perform such a comparison. Aim 2: Study somatic retrotransposon insertion-induced tumor heterogeneity. Assess the timing of retrotransposition events with respect to stage of malignancy and other markers of tumor clonal evolution by (1) correlating retrotransposon activity with other types of mutations mapped by WGS; (2) with the transcript level of the retrotransposons themselves and that of all other genes based on RNA-seq where possible; (3) correlating retrotransposon activity with the epigenetic status of the retrotransposons themselves where appropriate data are available; and (4) identifying clinical correlates of transposon activity. For instance, we detected a correlation of L1 activity with age in colorectal cancer, but not in pancreatic cancer. The pan-cancer data set would potentially enable us to correlate retroelement activity with the chemo- or radiation therapy status of the patients. Ultimately, by comparing matched normal and tumor tissues to each other and to a large number of other cancers, we hope to identify a small subset of factors that may cause increased retrotransposon mobilization in cancers. Previously we also found all L1 insertions to be truncated in cancer and that many insertions integrated via a non-classical, so-called endonuclease-independent mechanism. Thus, we also hope to identify a small subset of factors and candidate genes that may cause retroelement truncation and alternative integration mechanism into the genome. These genes would be then tested for functionality (see Aim 3). Aim 3: If resources permit (fresh tissue or fresh-frozen tissues with excellent cell integrity available from patients), we may be able to make cell lines to investigate the cause and functional impact of retrotransposon insertions. We would utilize functional genetic assays to characterize the role of novel genes targeted by mobile element insertions and to confirm candidate genes' role in regulating transposon activity using our retrotransposon assay. If patient-derived tissues are not available, this Aim will be performed in already established human cell lines using the CRISPR/CAS9 or the TALEN system to introduce the observed insertions and mutations.
Legacy plans

We have significant experience developing software to detect transposable element and processed pseudogene insertions in a patient-matched tumor/normal context using TCGA data. A number of tools exist for this purpose, but have not been compared against one another to any significant extent. Thus, improvements in computational retroelement insertion detection methods and information regarding the overlap between results of different available methods will be available following the conclusion of our proposed study.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Haig Kazazian, M.D.	Professor
eRA COMMONS USER NAME	
kazazian	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Dartmouth College, Hanover, NH	A.B	1959	Medical Science
Dartmouth Medical School, Hanover, NH		1958 – 1960	Medicine
The Johns Hopkins Univ. School of Med., Baltimore, MD	MD	1962	Medicine
The Johns Hopkins Univ. School of Med., Baltimore, MD		1964 – 1966	Post Doc: Ped. Genetics

A. Positions and Honors

Post Graduate Training:

Intern and Resident in Pediatrics, Univ. of Minn. Hospitals, l962-64
Staff Associate, USPHS, NIAMD, Laboratory of Molecular Biology, l966-68
Resident in Pediatrics, The Johns Hopkins Hospital, l968-69
Royal Society of Medicine Foundation Fellow, MRC Laboratory of Molecular Biology, Cambridge, England, July-October, l970

Selected Appointments:

Prof. of Pediatrics, The Johns Hopkins Univ. Sch. of Med., l977-1994
Prof. of Biology, Ob. Gyn., Medicine, Johns Hopkins Univ. Sch. of Arts & Sciences, l979-1994
Dir., Ctr. for Medical Genetics, The Johns Hopkins Univ. Sch. of Med, 1989-1994
Chairman, Department of Genetics, University of Pennsylvania School of Medicine, 1994-2006
Seymour Gray Professor of Molecular Medicine in Genetic, Department of Genetics, University of Pennsylvania School of Medicine, 1994-2010
Professor, Institute for Genetic Medicine, Johns Hopkins Univ. Sch. of Med., 2010-present

Selected Honors and Awards:

Mead Johnson Award for Pediatric Research, Amer. Acad. of Peds., 1976
Member, Institute of Medicine, National Academy of Science, 1992
The Dr. Murray Thelin Award for Research in Hemophilia, The National Hemophilia Foundation, 1994
Fellow, American Academy of Arts and Sciences, 2007
William Allan Award, American Society of Human Genetics, 2008

B. Selected peer-reviewed publications (*from among 376 publications since 1965)*

Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG and Antonarakis SE: Haemophilia A resulting from de novo insertion of L1 represents a novel mechanism for mutation in man. *Nature* **332:** 164-6, 1988.

Cutting GR, Kasch LM, Rosenstein BJ, Zielenski J, Tsui LC, Antonarakis SE and Kazazian HH, Jr.: A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. *Nature* **346:** 366-9, 1990.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF and Kazazian HH, Jr.: Isolation of an active human transposable element. *Science* **254:** 1805-8, 1991.

Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD and Gabriel A: Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808-10, 1991.

Holmes SE, Dombroski BA, Krebs CM, Boehm CD and Kazazian HH, Jr.: A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat Genet* **7:** 143-8, 1994.

Feng Q, Moran JV, Kazazian HH, Jr. and Boeke JD: Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905-16, 1996.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD and Kazazian HH, Jr.: High frequency retrotransposition in cultured mammalian cells. *Cell* **87:** 917-27, 1996.

Associate Professor Geoffrey J. Faulkner

Current position Principal Research Fellow, Mater Medical Research Institute, Australia
Associate Professor, University of Queensland, Australia

Degrees PhD, University of Queensland, 2009
BSc (First Class Honours), University of Queensland, 2004

Brief summary of recent activity I am head of a research group focused on the discovery, mapping and functional characterisation of mobile genetic elements in the mammalian genome. My group has in recent years concentrated on the development of a high-throughput retrotransposon capture sequencing (RC-seq) method that can detect low frequency somatic mutations in DNA extracted from tissue (Baillie, et al., *Nature*, 2011). Very recently, we have shown that retrotransposition supplies driver mutations in liver cancer (Shukla et al., *Cell*, 2013). My lab has 11 members and is supported by >$4M in awarded funding.

Current funding (as chief investigator)

2014-2016 *Blocking mobile DNA activity in induced pluripotent stem cells.* NHMRC Project Grant ($685,378) Role: CIA

2014-2016 *Mobile DNA reveals new liver cancer risk factor genes.* NHMRC Project Grant ($612,562) Role: CIA

2013-2016 *Mobile DNA in human development and disease*. NHMRC Career Development Fellowship ($397,724) Role: CIA

2013-2015 *Mobile DNA as a contributor to human brain cancer.* NHMRC Project Grant ($643,847) Role: CIA

2013-2015 *Genetic consequences of mobile DNA activation in abnormal neurodevelopment.* NHMRC Project Grant ($403,390) Role: CIA

2010-2015 *Systems biology of liver cancer: an integrative genomic-epigenomic approach.* EU FP7 Large Cooperation Grant ($22,530,000) Role: one of 12 co-CIAs

Current supervision 7 postdoctoral research fellows, 2 research assistants, 1 PhD student

Awards, prizes and other recognition

2013- TRI Caucus member

2013- NHMRC Grant Review Panel member

2011 US National Institute of Mental Health joint “No. 1 research advance of 2011”

2011 FEBS Anniversary Prize

2010- Referee for Wellcome Trust fellowships, Czech Science Foundation, EU FP7 and UK MRC research grants and NHMRC grants.

2010- Invited member of Functional Annotation of Mouse (FANTOM5) consortium

2009- Reviewer for *Nature*, *Science*, *Cell*, *PNAS* and a broad range of genomics journals (>30 reviews per annum).

2009 ASMR Queensland Premier's Award

Highlighted publications (30 publications total, >4,400 citations, h-index: 21)

Shukla, R., Upton, K.R., others, **Faulkner, G.J.*** *Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma.* **Cell** 153, 101-111 (2013)

Baillie, J.K., Barnett, M.W., Upton, K.R., others, **Faulkner, G.J.*** *Somatic retrotransposition alters the genetic landscape of the human brain.* **Nature** 479, 534-537 (2011)

Mattick, J.S., Taft, R.T., **Faulkner, G.J.*** *A global view of genomic information – moving beyond the gene and the master regulator*. **Trends in Genetics** 26, 21-8 (2010)

Faulkner, G.J., et al. *The regulated retrotransposon transcriptome of mammalian cells.* **Nature Genetics** 41, 563-71 (2009)

Conference abstracts (25 invited/plenary international talks in past 5 yrs + various others)

BIOGRAPHICAL SKETCH

NAME Szilvia Solyom, PhD	POSITION TITLE Postdoctoral fellow
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Eotvos Lorand University, Faculty of Science (Budapest, Hungary)	MSc	06/2004	Genetics
Biocenter Oulu / University of Oulu (Oulu, Finland)	PhD	04/2011	Cancer Genetics
Johns Hopkins University School of Medicine (Baltimore, USA)	Postdoctoral	05/2011	Genetics

A. Personal Statement

My long-term interest is to understand the genetic background of cancer, with a particular emphasis on how mutations contribute to malignancy. As a MSc and PhD student, I studied how classical mutations and large genomic rearrangements predispose to hereditary colon and breast cancer. As a postdoctoral fellow, my goal is to assess the impact of human retrotransposons ("jumping genes") on malignancy.

B. Positions and Honors

2011 – : Postdoctoral fellow, Kazazian Laboratory, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

C. Selected Peer-reviewed Publications

1. **Solyom S***, Ewing AD*, Rahrmann EP, Doucet D, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, Wheelan S, Upton KR, Shukla R, Faulkner GJ, Largaespada DA, and Kazazian HH Jr: Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 2012 Dec;22(12):2328-2338.* equal contribution

2. **Solyom S***, Ewing AD*, Hancks DC*, Takeshima Y, Awano H, Matsuo M, Kazazian HH Jr: Pathogenic orphan transduction created by a non-reference LINE-1 retrotransposon. Hum Mutat. 2012 Feb;33(2):369-371. * equal contribution

3. **Solyom S*,** Aressy B*, Pylkäs K*, Patterson-Fortin J*, Hartikainen JM, Kallioniemi A, Kauppila S, Nikkilä J, Kosma VM, Mannermaa A, Greenberg RA, Winqvist R: Breast cancer-associated Abraxas mutation disrupts nuclear localization and DNA damage response functions. Sci Transl Med. 2012 Feb 22;4(122):122ra23. * equal contribution

4. **Solyom S**, Winqvist R, Nikkilä J, Rapakko K, Hirvikoski P, Kokkonen H, Pylkäs K: Screening for large-size genomic rearrangements in the FANCA gene reveals extensive deletion in a Finnish breast cancer family. Cancer Lett. 2011 Mar;302(2):113-118.

Funding
Recipient of the 2013 AACR Basic Cancer Research Fellowship
Project Title: The impact of human retrotransposons on gastrointestinal cancers; amount: $45,000; term: November 1st, 2013-October 31st, 2014

Adam D. Ewing, Ph.D.
Center for Biomolecular Science and Engineering
University of California at Santa Cruz
ewingad@soe.ucsc.edu

Education

Postdoctoral Scholar, Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Sept. 2011 - present. Mentor: David Haussler, Ph.D.

Postdoctoral Fellow, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Oct. 2010 – Aug 2011. Mentor: Haig H. Kazazian, Jr., M.D.

Ph.D., Genomics and Computational Biology, University of Pennsylvania School of Medicine, Aug. 2005 – Sept. 2010. Mentor: Haig H. Kazazian, Jr., M.D.

B.A. Biology and Computer Science, Hiram College, Aug. 2001 – May 2005, *Cum Laude*, Dept. Hons. Biology, Dept. Hons. Computer Science

Selected Publications

Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L, Wilson RK, Haussler D. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* 14:R22.

Solyom S, **Ewing AD**, Rahrmann EP, Doucet TT, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, Wheelan S, Upton KR, Shukla R, Faulkner GJ, Largaespada DA, Kazazian HH Jr. 2012. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* Dec 22: 2328-38.

Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* 21: 985-990.

Ewing AD, Kazazian HH Jr. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* 20: 1262-1270. (Highlighted in *Nature*, *Nature Reviews Genetics*, *Nature Methods*, and *Cell*).

Hancks DC*, **Ewing AD***, Chen JE, Tokunaga K, Kazazian HH Jr. 2009. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res.* 19: 1983-1991.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

CNV, Structural aberrations and mitochondrial genome analysis from whole genome sequencing

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Raju Kucherlapati, Harvard Medical School, TCGA
Peter Park, Harvard Medical School, TCGA

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Angela Hadjipanayis, Harvard Medical School
Lixing Yang, Harvard Medical School

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Lynda Chin, M.D. Anderson Cancer Center

Background and preliminary data

We have been a part of the TCGA effort since its inception. Raju Kucherlapati is the PI of a TCGA Genome Characterization Center. The goals of our Center is generate whole genome sequence data from tumor/normal samples and analyze these data for copy number changes, structural aberrations, mitochondrial genomes and pathogens. Our group has contributed to all of the TCGA projects. In addition, our group also developed a number of new algorithms for detection of copy number variation from whole genome analysis (BicSeq), structural aberration analysis (Meerkat). We have also published on mitochondrial genomes in several different cancers, participated in the analysis of HPV sequences in head and neck cancer and in analyzing transposition events in many different cancers. Here are some publications from our group on this subject.

1. Cancer Genome Atlas Research Network. (2011) Nature. 474(7353):609-15,
2. Xi etal (2011) Proc Natl Acad Sci U S A. 108(46):E1128-E1136,
3. Lee etal (2012) Science. 337(6097):967-71.,
4. Cancer Genome Atlas Network. (2012) Nature. 487(7407):330-7.,
5. Larman etal (2012) Proc Natl Acad Sci U S A. 109(35):14087-14091.
6. Cancer Genome Atlas Research Network. (2012) Nature. 27;489(7417):519-25.
7. Cancer Genome Atlas Network. (2012) Nature. 490(7418):61-70.
8. Cancer Genome Atlas Research Network, (2013) Nature. 497(7447):67-73.
9. Yang et al (2013) Cell. 153(4):919-29.
10. Cancer Genome Atlas Research Network. (2013) Nature. 499(7456):43-9.

Timelines & resources dedicated to project

Personnel: We have highly experienced personnel in Boston and in Houston who have extensive experience in all aspects of the proposed data analysis (see publication list above)

Computational cluster. We have access to computational clusters at Harvard and at M.D. Anderson and another in South Korea at the Samsung Genome Center
Biological expertise: Our team is composed of not only computational and Informatics personnel but highly trained molecular and cancer biologists, who would be able to discern the biological significance of any findings.
Timeline: We will devote as much time is necessary for the analysis of the samples.

Research proposal

1. **Copy number variation (CNV) analysis. We will examine the whole genome sequence data for copy number variation using our BicSeq algorithm. This method, divides the genomes into segments or bins of desirable size (usually 1 kb) and counts the number of reads in that bin from the tumor and its corresponding normal samples. Based on the ratio of the number of reads from each pair of samples, it is possible to deduce all CNVs in the genome. We have also developed another algorithm, called Integer that uses the copy number information to deduce the purity and ploidy of the tumor samples. Based on this information, it is possible to precisely estimate the changes in copy number in each sample.**
2. **Structural aberration analysis. We currently use two methods to initially deduce somatic structural rearrangements in tumors. We use BreakDancer a publicly available method and Meerkat, a method that we have developed for these analyses. BreakDancer deduces structural aberrations based on discordant reads of the two ends of sequence obtained from Illumina sequencing. Meerkat also uses discordant paired-ends but also requires the detection of a chimeric reads that span the rearrangement breakpoint. Using two different methods and requiring reads from the rearrangement junctions provides greater level of accuracy of the results. We also attempt to confirm all productive rearrangements by examining RNA-Seq data from the tumors.**
3. **Mitochondrial Genomes. In each cell for each nuclear genome, there are several hundreds to thousands of mitochondria each carrying a single mitochondrial DNA genome. Therefore, whole genome sequencing of tumor/normal pairs provide an abundance of coverage of the mitochondrial genome. We have developed methods to analyze the mitochondrial genome to determine (a) the number of mitochondria, (b) mitochondrial mutations in the form of single nucleotide variants and insertion/deletions and (c) the degree of heteroplasmy of mitochondria. It is emerging that mitochondrial genome changes could play an important role in cellular metabolism and this study will help identify critical changes in this organelle**
4. **Pathogens. Analysis of whole genome sequencing is first accomplished by aligning the sequence data to that of the standard human genome. All of the mitochondrial DNA and any pathogen DNA associated with the tumor does not align to the human genome. We have developed a simple method to search the non-aligned sequences for mitochondrial and many of the known human pathogens. Using this approach we were able to establish that some head and neck tumors have HPV sequences in them and in many cases these sequences are integrated into the human genome. Other groups have shown**

the involvement of bacterial sequences in colorectal and other tumors and our analysis will reveal such information as well.

Legacy plans

We will work on this project and deposit the results of the analysis in the appropriate database. We will work on this project until the completion of the analyses and publication of the results. All of the tools we have developed are available for licensing from Harvard Medical School. For academic investigators, these tools are provided at no cost and several members of the TCGA community are already using these tools.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Raju Kucherlapati	Professor of Genetics and Medicine
eRA COMMONS USER NAME (credential, e.g., agency login) raj123	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
P.R. College, Kakinada, India	B.S.	06/1960	Biology
Andhra University, India	M.S.	06/1962	Genetics
University of Illinois, Urbana, IL	Ph.D.	03/1972	Genetics
Yale University, New Haven, CT	Fellow	03/1972-1975	Genetics

A. Personal Statement

I will be a co-mentor for Dr. Yang

B. Positions and Honors

Research and Professional Experience:

1967-72 Graduate Student, University of Illinois, Urbana, IL
1972-75 Postdoctoral Fellow, Yale University, New Haven, CT
1975-82 Assistant Professor, Department of Biochemical Sciences, Princeton University, Princeton, NJ
1982-89 Professor, Dept. of Genetics, University of Illinois College of Medicine, Chicago, IL
1989-2001 Professor and Chairman, Dept. of Molecular Genetics, Albert Einstein College of Medicine, NY
2001-2008 Scientific Director, Harvard-Partners Center for Genetics and Genomics, Brigham & Women's Hospital
2001- Paul C. Cabot Professor of Genetics and Professor of Medicine, Harvard Medical School, Boston, MA

Honors:

1962 Pushpa Rangaswamy Iyengar Memorial Prize for standing first in class of 1962
1. Damon Runyon Memorial Cancer Research Fellowship
1974-1975 NIH Postdoctoral Fellowship
1981-1984 Member, NIH Biomedical Sciences Study Section
1985-1989 Member, NIH Mammalian Genetics Study Section
1988 Co-organizer, Cold Spring Harbor Symposium: Intermediates in Genetic Recombination
1989 Chairman, Gordon Research Conference on Molecular Genetics
1989 Co-organizer, NIH Workshop: Applications of Homologous Recombination to Human Genetics
1990-1995 Member, Genome Research Review Committee, NCHGR, NIH
1992-1997 Co-organizer, Chromosome 12 Workshops #1, 2, 3 and 4

1995-1997 Member, Mental Retardation Review Committee, NICHD, NIH
1997-1999 Chairman, Mental Retardation Review Committee, NICHD, NIH
1998-2003 Member, National Advisory Council for Human Genome Research, NHGRI, NIH
1999 Vice President and President-Elect, The Harvey Society, 1998-1999
2000 President, The Harvey Society, 1999-2000
2002 & 2005 Co-Chair, AACR Conference on Colon Cancer
2001-2005 Co-Chair, NCI Mouse Models for Human Cancer Consortium Steering committee
2006-2007 Co-Chair, AACR Centennial Meeting Organizing Committee
2006 Elected as Fellow of the American Association for the Advancement of Science (AAAS)
2008 Elected to the Institute of Medicine of the National Academy of Sciences 2008

C. Selected Peer-reviewed Publications

1. Krauter K, Montgomery K, Yoon S-J, LeBlanc-Straceski J, Renault B, Marondel I, Herdman V, Cupelli L, Banks A, Lieman J, Menninger J, Bray-Ward P, Nadkarni P, Weissenbach J, Chumakov I, Cohen D, Miller P, Ward D, Kucherlapati R. A second generation YAC contig map of human chromosome 12. Nature 1995;377:321-333.
2. Edelmann W, Cohen PE, Kane M, Lau K, Morrow B, Bennett S, Umar A, Kunkel T, Cattoretti G, Chaganti R, Pollard JW, Kolodner RD, Kucherlapati R. Meiotic pachytene arrest in MLH-1-deficient mice. Cell 1996;85:1125-1134.
3. Edelmann W, Yang K, Umar A, Heyer J, Lau K, Fan K, Liedtke W, Cohen PE, Kane MF, Lipford JR, Yu N, Crouse GF, Pollard JW, Kunkel T, Lipkin M, Kolodner R, Kucherlapati R. Mutation in the mismatch repair gene *Msh6* causes cancer susceptibility. Cell 1997;91:467-477.
4. Merscher S, Funke B, Epstein JA, Heyer J, Puech A, Lu MM, Xavier RJ, Demay MB, Russell RG, Factor S, Tokooya K, St. Jore B, Lopez M, Pandita RK, Lia M, Carrion D, Schorle H, Kobler JB, Scambler P, Wynshaw-Boris A, Skoultchi AI, Morrow BE, Kucherlapati R. *TBX1* is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. Cell 2001;104:619-629.
5. Lander ES, Kucherlapati R, et al. "International Human Genome Sequencing Consortium". Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
6. Montgomery KT, Lee E, Miller A, Lau S, Shim C, Decker J, Chiu D, Emerling S, Sekhon M, Kim R, Lenz J, Han J, Ioshikhes I, Renault B, Marondel I, Yoon S-J K, Song K, Murty VVVS, Scherer S, Yonescu R, Kirsch IR, Ried T, McPherson J, Gibbs R and Kucherlapati R. A high-resolution map of human chromosome 12. Nature 2001;409:945-946.
7. Velcich A. Yang W, Heyer J, Fragale A, Nicholas C, Viani S, Kucherlapati R, Lipkin M, Yang K, Augenlicht L. Colorectal cancer in mice genetically deficient in the mucin Muc2. Science 2002;295:1726-1729.
8. Waterston RH, Kucherlapati R, Montgomery K, Lander ES. Initial sequencing and comparitive analysis of the mouse genome. Nature 2002;420:520-562.
9. Takahashi C, Bronson RT, Socolovsky M, Contreras B, Lee KY, Jacks T, Noda M, Kucherlapati R, Ewen ME. Rb and N-ras function together to control differentiation in the mouse. Mol Cell Biol 2003;23:5256-68. **PMCID: PMC165732**

10. Lin Q, Clark AB, McCulloch SD, Yuan T, Bronson RT, Kunkel TA, Kucherlapati R. Increased susceptibility to UV-induced skin carcinogenesis in polymerase eta-deficient mice. Cancer Res 2006;66:87-94.
11. Scherer SE, Nelson D, Kucherlapati R, Weinstock G, Gibbs RA; Baylor College of Medicine Human Genome Sequencing Center Sequence Production Team. The finished DNA sequence of human chromosome 12. Nature 2006;440:346-351.
12. Roberts AE, Araki T, Swanson KD, Montgomery KT, Schiripo TA, Joshi VA, Li L, Yassin Y, Tamburino AM, Neel BG, Kucherlapati RS. Germline gain-of-function mutations in SOS1 cause Noonan syndrome. Nat Genet. 2007;39(1):70-74.
13. The Cancer Genome Atlas Research Network Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature 2008;*455(7216):1061-8. **PMCID: PMC2671642**
14. Vassilopoulos S, Esk C, Hoshino S, Funke BH, Chen CY, Plocik AM, Wright WE, Kucherlapati R, Brodsky FM. A role for the CHC22 clathrin heavy-chain isoform in human glucose metabolism. *Science 2009;*324(5931):1192-6. **PMCID: PMC2975026**
15. Kucherlapati MH, Esfahani S, Habibollahi P, Wang J, Still ER, Bronson RT, Mahmood U, Kucherlapati RS. (2013) Genotype directed therapy in murine mismatch repair deficient tumors. PLoS One. 8(7):e68817.

D. Research Support

Ongoing Research Support

5 U01 CA84301-10 (Kucherlapati)
09/30/1999 – 08/31/2014
NIH/NCI
Mouse Models for Human Cancer
The goal of the project is to generate mouse models for gastrointestinal cancer.
Role: Principal Investigator

1U24CA144025-0110
09/29/2009 – 07/31/2014
NIH/NCI
Harvard Genome Characterization Center
The goal of the proposed effort is to analyze 2,000-2,500 tumor samples each year over a five-year period of time and identify a set of genes that can be resequenced by the members of The Cancer Genome Atlas (TCGA).
Role: Principal Investigator

Completed Research Support

5 P50 CA127003-02 (Fuchs)
08/31/2007 – 06/30/2012
NIH
Molecular Fluorescent Imaging of Colorectal Neoplasms (Project 2)

The goal of this project is the use of mouse models for imaging tumors and development of therapeutic strategies.
Role: Project Director

3U24CA144025-02S2
09/1/2010 – 08/31/2012
NIH/NCI
Harvard Genome Characterization Center (ARRA)
The goal of the proposed effort is to initiate a pilot effort to sequence mouse tumors
Role: Principal Investigator

3U24CA144025-02S1
09/1/2010 - 08/31/2011
NIH/NCI
Harvard Genome Characterization Center (ARRA)
The goal of the proposed effort is to accelerate transition from array based CGH to sequence based determination of CNV in human tumor samples.
Role: Principal Investigator

5 U24 CA126554-03 (Kucherlapati)
09/28/2006 – 08/31/2011
NIH/NCI
Cancer Genomics Center
The goal of this project is to identify regions of the caner genome that show copy number changes.
Role: Principal Investigator

1R01 DE016140-01 (Kucherlapati)
07/01/2004 – 06/30/2008
NIH/NIDCR
Clinical, Genetic and Morphometric Analysis of VCFS/DGS
The goal of the project is to conduct clinical research on the facial features of VCFS/DGS patients using morphometric methods
Role: Principal Investigator

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Peter J. Park	Associate Professor of Pediatrics
eRA COMMONS USER NAME (credential, e.g., agency login) PPARK1	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Harvard University, Cambridge, MA	AB/SM	9/90-6/94	Applied Mathematics
California Institute of Technology, Pasadena, CA	PhD	9/94-8/99	Applied Mathematics
Harvard School of Public Health, Boston, MA	SM	9/99-6/00	Biostatistics
Harvard School of Public Health, Boston, MA	Postdoctoral	7/00-6/01	Biostatistics

A. Personal Statement

My area of focus is development and application of computational methods for analysis of high-throughput genomic data to better understand epigenetic mechanisms and cancer genetics. In epigenetics, we specialize in analysis of ChIP-seq data for histone modifications and transcription factors (the algorithm we have developed is one of the most popular ones today) as well as integrative analysis using multiple data types. In cancer genomics, we specialize in identification and analysis of copy number and structural variations from whole-genome sequencing data and their relevance in tumorigenesis. We have made major contributions to NIH consortium projects including The Cancer Genome Atas (TCGA) and Encyclopedia of DNA Elements (DNA) projects.

B. Positions and Honors.

2001-2005 Instructor, Harvard Medical School, Boston, MA
2006-2010 Assistant Professor of Pediatrics, Harvard Medical School, Boston, MA
2010- Associate Professor of Pediatrics, Harvard Medical School, Boston, MA

2002-2006 Instructor, Department of Biostatistics, Harvard School of Public Health, Boston, MA
2003-2008 Associate Director of Bioinformatics, Harvard-Partners Center for Genetics and Genomics
2004- Affiliated faculty, Harvard-MIT Health Sciences and Technology
2007- Member, Dana-Farber/Harvard Cancer Center
2007- Member, Biological and Biomedical Sciences Program at Harvard Medical School
2010- Member, Division of Genetics, Brigham and Women's Hospital
2011- Affiliate faculty, Children's Hospital Stem Cell Program
2012- Affiliate faculty, Harvard Stem Cell Institute
2012- Co-director, Bioinformatics and Integrative Genomics (BIG) Pre-doctoral Training Program

Honors: National Merit Scholarship (1990), John Harvard Scholarship (1993), Alfred P. Sloan Research Fellowship (2010), Harvard Medical School Young Mentor Award (2012)

C. Selected peer-reviewed publications.

Most relevant to the current application (*equal contribution; **co-corresponding authors)

1. Kim TM, Laird PW, **Park PJ**. (2013) The landscape of microsatellite instability in colorectal and endometrial cancer genomes, *Cell,* 155:858-868, 2013.
2. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh C, Zhang C, Ren X, Protopopov A, Chin L, Kucherlapati R, Lee C, **Park PJ**. (2013) Diverse mechanisms of somatic structural variations in human cancer genomes, *Cell*, 153:919-29. PMCID: PMC3704973.
3. Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, **Park PJ** (2013) Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research,* 23:217-27. PMCID: PMC3561863.
4. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV**, **Park PJ****, and The Cancer Genome Atlas Research Network (2012) Landscape of somatic retrotransposition in human cancers, *Science,* 337:967-71. PMCID: PMC3656569.
5. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, Kucherlapati R, **Park PJ** (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci USA*, 108:E1128-3. PMCID: PMC3219132.

Additional recent publications

1. Tolstorukov MY*, Sansam CG*, Lu P*, Koellhoffer EC, Helming KC, Alver BH, Tillman EJ, Evans JA, Wilson BG, **Park PJ****, Roberts CWM**. (2013) The Swi/Snf tumor suppressor complex sculpts the nucleosome landscape at promoters on a genome scale, *Proc Natl Acad Sci USA*, 110:10165-70.
2. Apostolou E, Ferrari F, Walsh RM, Bar-Nur O, Stadtfeld M, Cheloufi S, Stuart HT, Polo JM, Ohsumi TK, Borowsky ML, Kharchenko PV, **Park PJ***, Hochedlinger K*. (2013) Genome-wide Chromatin Interactions of the Nanog Locus in Pluripotency, Differentiation, and Reprogramming, *Cell Stem Cell*, 12:699-712.
3. Ferrari F*, Jung YL*, Kharchenko PV, Plachetka A, Alekseyenko AA, Kuroda M**, **Park PJ**** (2013) Comment on "Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters", *Science*, 340:273, 2013. PMCID: PMC3665607.
4. Cancer Genome Atlas Research Network. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499:43-9.
5. Cancer Genome Atlas Network. (2013) Integrated genomic characterization of endometrial carcinoma, *Nature*, 497:67-73.
6. Evrony GD*, Cai X*, Lee E, Hills LB, Elhosary PC, Parker JJ, Atabay KD, Lehmann HS, Gilmore EC, Poduri A, **Park PJ**, Walsh CA (2012) Single neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, 151:483-96.

7. ENCODE consortium (594 authors including **Park PJ**) (2012) An integrated encyclopedia of DNA elements in the human genome, *Nature,* 489:57-74. PMCID: PMC3439153.
8. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TP, Sandstrom R, Thurman RE, MacAlpine DM, Stamatoyannopoulos JA, Kellis M, Elgin SCR, Kuroda MI, Pirrotta V, Karpen GH*, **Park PJ***. (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*, *Nature*, 471:480-5. PMCID: PMC3109908.
9. Tolstorukov MY, Volfovsky N, Stephens RM, **Park PJ** (2011) Impact of chromatin structure on sequence variability in the human genome, *Nature Structural & Molecular Biolology*, 18:510-5. PMCID: PMC3188321.
10. **Park PJ.** (2009) ChIP-seq: advantages and challenges of a maturing technology, *Nature Reviews Genetics*, 10:669-80. PMCID: PMC3191340.

D. Research Support

Ongoing research support

U24 CA143845 (Chin, Lynda) 9/29/09-07/31/14
The Cancer Genome Atlas Data Analysis Center
Goal: Development of a data analysis pipeline for the Cancer Genome Atlas, to analyze multiple data types from microarray and next-generation sequencing platforms for 20-25 tumors types. Role: co-investigator

U24 CA144025 (Kucherlapati, Raju) 9/29/09-7/31/14
Harvard Genome Characterization Center
Goal: Identification of genomic aberrations in tumor samples by array comparative genomic hybridization and next-generation sequencing, as part of the Cancer Genome Atlas project. Role: co-investigator

U54 LM008748 (Kohane, Isaac S) 9/15/10-9/14/14

National Center for Biomedical Computing: Informatics for Integrating Biology and the Bedside
Goal: Developing statistical methods for quantitative linkage and association analysis and predictive
models for clinical outcomes. Role: co-investigator

R01 HL080494 (Seidman, Jonathan) 4/01/11-4/30/15
Defining Genetic Architecture and Pathways of DCM
Goal: To identify novel causal genes and mutations for dilated cardiomyopathy and characterize the RNA changes. Role: co-investigator

R01 CA113794 (Roberts, Charles) 6/01/11-5/31/15

The Function of Snf5, an Epigenetic Tumor Suppressor
Goal: To provide insight into normal Swi/Snf function, define a mechanism by which disruption of this complex causes the rapid onset of aggressive, lethal cancers and identify novel targets for therapeutic intervention. Role: co-investigator

1U19 HD077671 (Green, Robert) 9/5/13-8/31/18

Genome sequence-based screening for childhood risk and newborn illness
Goal: To test the feasibility and impact on physicians and parents of genomic sequencing in the newborn to assess future risk of childhood onset disease and to guide diagnosis and treatment of sick newborns

Completed research support

RL1 DE019021 (Maas, Richard) 9/30/07-9/29/12
Systems-based Consortium for Organ Design and Engineering
Goal: Study of the regeneration process using tooth, pancreatic islets, and stem cells as model systems
Role: co-investigator

R01 GM082798 (Park, Peter) 9/25/07-8/31/12
Integrative Analysis of Multiple Genomic Data Sets
Goal: To develop computational methods for meta-analysis of gene expression data. Role: PI

RC2 HG005639 (Kellis, Manolis) 9/30/09-8/31/12
A Data Analysis Center for integration of fly and worm modENCODE datasets
Goal: Analysis of genomic data from multiple platforms to characterize functional elements in the genomes. Role: PI on subcontract

RC1 HG005482 (Park, Peter) 9/22/09-6/30/12
Statistical methods for estimation of copy number from next-generation sequencing
Goal: To develop algorithms for segmentation of copy number profiles based on next-generation sequencing data. Role: PI

U01 HG004258 (Karpen, Gary) 5/04/07-3/31/12
Genome-Wide Mapping of Chromosomal Proteins in Drosophila
Goal: Comprehensive identification of the DNA regions bound by chromosomal proteins or affected by histone modification using chromatin-immunoprecipitation on tiling arrays (ChIP-chip). Role: PI on subcontract

RC2 HL102815 (Daley, George) 9/30/09-2/29/12
Comparative phenotypic, functional, and molecular analysis of ESC and iPSC
Goal: Application of genomic technologies to characterize the differences between embryonic stem cells and induced pluripotent stem cells, if any. Role: co-PI

BIOGRAPHICAL SKETCH

NAME Hadjipanayis, Angela	POSITION TITLE Postdoctoral Fellow
eRA COMMONS USER NAME (credential, e.g., agency login) anghadji23	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Maryland	BSc.	1999	Biochemistry
Georgetown University	MSc.	2001	Biochemistry
University of Florida	Ph.D	2008	Genetics
Brigham and Women's Hospital/Harvard Medical School	Postodoctoral	present	Cancer Genomics

A. Personal Statement

I will be a Post-doctoral Fellow aiding in the analysis of whole genome sequences.

B. Positions and Honors

Postions and Employment

1998-2002 Research Technician at USAMRIID, Department of Defense, Fort Detrick MD.

Honors

2010-2011 Ruth L. Kirschstein NRSA Postdoctoral Award, Harvard Medical School Training Grant (5 T32 GM 7748-33).
2009 American Association of Cancer Research (AACR) AFLAC Scholar-in-Training Award, Denver, Colorado. Competitive travel award based on the qualitative rating of the abstract and the letter of recommendation.

2006 – 2008 Children's Tumor Foundation Young Investigator Award. Competitive fellowship application for predoctoral and postdoctoral fellows.
2008

Travel Award, Children's Tumor Foundation, to attend and present work at the NF Conference, Bonita Springs, FL .
2007

Travel Award, University of Florida College of Medicine Graduate Program. Competitive award to present work at the American Society of Human Genetics.
2001

Certificate of Achievement for testing anthrax samples during the October 2001 attacks, Army Secretary of State.

Other Experience and Professional Memberships

2005 – 2006 Peer Reviewer, Medical Guild graduate student grants, UF College of Medicine
2007-present American Society of Human Genetics member
2009-present American Association for Cancer Research member
2011-present American Association for the Advancement of Sciences member

C. Selected Peer-reviewed Publications

TCGA Research Network (2013). The Cancer Genome Atlas Pan Cancer analysis project. Nat Genet 45 (10): 1113-20.

TCGA Research Network (2013). Comprehensive molecular characterization of urothelial carcinoma of the bladder. Nature (in press)

Brennan CW, Verhaak RGW, Mckenna A, Campos B, Noushmehr H et al (2013). Somatic Genomic Landscape of Glioblastoma. Cell. 155(2):462-77.

TCGA Research Network (2013). Integrated genomic characterization of endometrial carcinoma. Nature 497 (7447): 67-73.

TCGA Research Network (2013). Comprehensive genomic characterization of squamous cell lung cancers. Nature 489 (7417): 519-25.

Larman TC, DePalma SR, **Hadjipanayis** AG; Cancer Genome Atlas Research Network, Protopopov A, Zhang J, Gabriel SB, Chin L, Seidman CE, Kucherlapati R, Seidman JG (2012). Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A. 109 (35): 14087-91.

TCGA Research Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature. 487 (7407): 330-7.

Xi R, **Hadjipanayis** AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, Kucherlapati R, Park PJ (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. 108 (46): E1128-36.

D. Research Support

None

Program Director/Principal Investigator (Last, First, Middle): Chin, Lynda

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Lynda Chin, MD	Professor and Chair, Dept of Genomic Medicine Scientific Director, Institute for Applied Cancer Science
eRA COMMONS USER NAME (credential, e.g., agency login) LYNDA_CHIN	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Brown University, Providence, RI	BS	09/84-06/88	Neuroscience
Albert Einstein College of Medicine, Bronx, NY	MD	09/89-06/93	Medicine
Columbia Presbyterian Medical Center, NY, NY	Internship	07/93-06/94	Internal Medicine
Albert Einstein College of Medicine, Bronx, NY	Residency	07/94-06/97	Dermatology
Albert Einstein College of Medicine, Bronx, NY	Postdoctoral	07/93-06/97	Molecular Genetics

A. Personal Statement

As a board-certified physician, my research activities have been shaped by my desire and instinct to impact on patient survival. At the same time, I firmly believe that effective translation to clinic endpoints begins with strong cutting-edged basic science. Thus, I have built a research program that spans genetically engineered mouse models (GEMM), basic cancer genetics, and cancer biology, as well as cutting-edge cancer genomics and computational biology. My research program focuses on mining and translating complex multi-dimensional cancer genomic data using a systems approach that employs comparative oncogenomics across species and develops computational network modeling of complex genomic data. My research integrates these findings with a variety of high-throughput functional genomics to prioritize most likely cancer-relevant genetic elements of interest for downstream biological and mechanistic studies.

B. Positions and Honors

Positions

1996 – 1997 Chief Resident, Dermatology, Albert Einstein College of Medicine (AECOM), NY
1998 – 2004 Assistant Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute (DFCI), Boston, MA
1999 – 2004 Scientific Director, Arthur & Rochelle Belfer Cancer Genomics Center, DFCI, Boston, MA
2005 – 2009 Associate Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2008 – Member, scientific steering committee, International Cancer Genome Consortium (ICGC).
2009 – 2011 Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2009 – Associate Member, the Broad Institute of MIT and Harvard, Boston, MA
2009 – 2011 Co-director, Melanoma Program, Dana-Farber/Harvard Cancer Center, Boston, MA
2009 – 2011 Scientific Director, the Belfer Institute for Applied Cancer Science, DFCI, Boston, MA
2009 – Member, Executive Subcommittee, The Cancer Genome Atlas (TCGA), USA
2011 – Professor and Chair, Department of Genomic Medicine, UTMDACC, Houston, TX
2011 – Scientific Director, Institute for Applied Cancer Science, UTMDACC, Houston, TX

Awards and Honors

2000 The Wilson Stone Memorial Award for Research Excellence
2001 The BASF Bioresearch Corporation Award
2001 Charles E. Culpeper Medical Scholar Award
2002 Election to American Society for Clinical Investigators (ASCI)
2003 The James S. McDonnell Foundation 21st Century Research Award

2004 The Claire and Richard Morse Research Award
2006 Election to the Council of the American Society for Clinical Investigators (ASCI)
2009 The Milstein Innovation Award, American Skin Association
2010 The Latta Lecturer, UCLA, CA
2012 Appointed, M.G. & Lillie A. Johnson Chair for Cancer Treatment and Research
2012 Elected, Institute of Medicine of National Academies (IOM)

Meeting Organization
2009 Translational Cancer Genomics (AACR)
The biology and genetics of brain cancers (AACR)
2011 Changing tumor landscape (Keystone)
Cancer Genomics (EMBO/EMBL)
2012 Translational Cancer Genomics (AACR)

C. Selected Peer-reviewed Publications (*selected from a total of 129 peer-reviewed publications*)

Most relevant to the current application

1. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L; TCGA Research Network. The somatic genomic landscape of glioblastoma. ***Cell***. 2013 Oct 10;155(2):462-77.
2. Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. ***Nat Rev Genet***. 2013 Oct;14(10):703-18.
3. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, Lawrence MS, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio RC, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton DL, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury JE, Davies MA, Gershenwald JE, Wagner SN, Hoon DS, Schadendorf D, Lander ES, Gabriel SB, Getz G, Garraway LA, Chin L. A landscape of driver mutations in melanoma. ***Cell*** 2012; 150(2): 251-263. PMCID: PMC3600117
4. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. ***Science*** 2013; 339(6122): 957-9
5. Hu J, Ho AL, Yuan L, Hu B, Hua S, Hwang SS, Zhang J, Hu T, Zheng H, Gan B, Wu G, Wang YA, Chin L, DePinho RA. From the Cover: Neutralization of terminal differentiation in gliomagenesis. ***Proc Natl Acad Sci U S A***. 2013 Sep 3;110(36):14520-7. PMCID: PMC3767545

Additional publications of importance to the field (in chronological order)

1. Kim MJ, Gans J, Nogueira CN, Wang A, Paik JH, Feng, B, Brennan C, Hahn W, Cordon-Cardo C, Wagner SN, Flotte T, Duncan L, Granter SR and Chin L. Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene. ***Cell,*** 2006; 125(7): 1269-1281.
2. Chin L[#], Gray JW. Translating insights from the cancer genome into clinical practice. ***Nature***, 2008; 452(7187): 553-563. (*[#]corresponding author*) PMC2730524
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblasotma genes and core pathways. ***Nature***, 2008; 455(7216): 1061-1068. Chin L[#] and Meyerson M[#]. (*[#]co-corresponding authors*) PMC2671642

4. Scott KL, Kabbarah O, Liang M-C, Ivanova E, Anagnostou V, Wu J, Dhakal S, Wu M, Chen S, Feinberg T, Huang J, Saci A, Widlund HR, Fisher DE, Xiao YH, Rimm DL, Protopopov A, Wong KK, Chin L. GOLPH3 modulates mTOR signaling and sensitivity to rapamycin in cancer. ***Nature***, 2009; 459: 1085-1090. PMC2753613
5. Hu J, Hwang SS, Liesa M, Gan B, Sahin E, Jaskelioff M, Ding Z, Ying H, Boutin AT, Zhang H, Johnson S, Ivanova E, Kost-Alimova M, Protopopov A, Wang YA, Shirihai OS, Chin L, Depinho RA. Antitelomerase Therapy Provokes ALT and Mitochondrial Adaptive Mechanisms in Cancer. ***Cell***. 2012; 148(4): 651-663. PMC3286017
6. Ding Z, Wu CJ, Jaskelioff M, Ivanova E, Kost-Alimova M, Protopopov A, Chu GC, Wang G, Lu X, Labrot ES, Hu J, Wang W, Xiao Y, Zhang H, Zhang J, Zhang J, Gan B, Perry SR, Jiang S, Li L, Horner JW, Wang YA, Chin L, Depinho RA. Telomerase reactivation following telomere dysfunction yields murine prostate tumors with bone metastases. ***Cell***. 2012; 148(5): 896-907.
7. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, Zhang H, Zeid R, Ren X, Cibulskis K, Sivachenko AY, Wagle N, Sucker A, Sougnez C, Onofrio R, Ambrogio L, Auclair D, Fennell T, Carter SL, Drier Y, Stojanov P, Singer MA, Voet D, Jing R, Saksena G, Barretina J, Ramos AH, Pugh TJ, Stransky N, Parkin M, Winckler W, Mahan S, Ardlie K, Baldwin J, Wargo J, Schadendorf D, Meyerson M, Gabriel SB, Golub TR, Wagner SN, Lander ES, Getz G, Chin L, Garraway LA. Melanoma genome sequencing reveals frequent PREX2 mutations. ***Nature*** 2012; 485(7399): 502-506. PMC3367798
8. Genovese G, Ergun A, Shukla SA, Campos B, Hanna J, Ghosh P, Quayle SN, Rai K, Colla S, Ying H, Wu CJ, Sarkar S, Xiao Y, Zhang J, Zhang H, Kwong L, Dunn K, Wiedemeyer WR, Brennan C, Zheng H, Rimm DL, Collins JJ, Chin L. microRNA regulatory network inference identifies miR-34a as a novel regulator of TGF-β signaling in glioblastoma. ***Cancer Discov*** 2012; 2(8): 736-774.
9. The Cancer Genome Atlas Research Network, et al. Comprehensive genomic characterization of squamous cell lung cancers. ***Nature*** 2012; 489(7417): 519-525. PMC3466113
10. Kwong LN, Costello JC, Liu H, Jiang S, Helms TL, Langsdorf AE, Jakubosky D, Genovese G, Muller FL, Jeong JH, Bender RP, Chu GC, Flaherty KT, Wargo JA, Collins JJ, Chin L. Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma. ***Nat Med*** 2012; 18(10): 1503-1510.

D. Research Support

Ongoing Research Support

1P01 CA163222 Fisher (PI) 12/01/11-11/30/16
NIH/NCI
Overcoming Resistance to BRAF (V600E) Targeted Therapies in Melanoma
The goal is to identify genetic events conferring resistance to BRAF in vivo.
Role: Project PI

1U01 CA168394 Mills (PI) 05/01/12-04/31/17
NIH/NCI
Biological Annotation of TCGA Data
The goal of this project is to facilitate the translation of newfound genomic knowledge into cancer therapeutics and diagnostics.
Role: Project PI

R1204 Chin (PI) 12/16/11-12/31/16
Cancer Prevention & Research Institute of Texas Recruitment Award
The goal of this project is to accelerate the translation of genomics to medicine and enabling evidence-based early disease management.
Role: PI

7P01 CA117969 DePinho (PI) 04/15/06-12/31/15
NIH/NCI
Genetics and Biology of Pancreatic Ductal Adenocarcinoma

The goal of this P01 is to further elucidate the genetics and biology of PDAC to a level that will guide the rational development of effective targeted agents, alone and in combination.
Role: Project PI

7U01 CA141508 Chin (PI) 08/01/09-07/31/14
NIH/NCI
Uses of GEM Models for Translational Cancer Research
The goal of this study is to generate genome instability melanoma and prostate cancer animal models. Comparative oncogenomics will be performed to identify driver events in metastasis as well as functional genetic creens to identify metastatic determinants.
Role: Co-PI

5U24 CA143845 Getz (PI) 08/01/09-07/31/14
NIH/NCI
The Cancer Genome Atlas Data Analysis Center
The goals of this project are to define caBIG compliant data format, design analysis modules to consolidate data from all components of TCGA and to perform integrative analyses, and to implement this high-throughput analysis pipeline in an industrial-level production mode.
Role: Co-PI

U24 CA144025 Kucherlapati (PI) 08/01/09-07/31/14
NIH/NCI
Harvard Genome Characterization Center in the Cancer Genome Atlas
TCGA goals are to utilize these high-throughput technologies to 1) Generate a complete genome, transcriptome, and epigenome characterization of 2000 cancer/normal pairs per year and make the data publicly available;
Role: Co-PI

5U54 CA163125 Chin (PI) 08/01/09-07/31/14
NIH/NCI
Role of Tumor Stroma in Therapeutic Response and Resistance
Global unbriased profiling will be performed to catalogue transcriptomic, epigenomic and proteomicalterations in BRAF mutant human melanomas and derivative cells at baseline, post-treatment and upon relapse on selective BRAF inhibitor.
Role: Co-PI

Completed Research Support

NIH U54 CA126505 Chin (PI) 09/25/06-08/31/11
TMEN Genomics & Bioinformatics Core
Role: PI

NIH 1RC2 CA148268 Hahn/ Chin (PI) 09/01/09-08/31/11
Functional Annotation of Cancer Genomes: TCGA, Glioblastoma and Ovarian Cancer

NIH P50 CA12703 Fuchs/ DePinho/Chin (PI) 07/01/07-08/31/11
National Institutes of Health: Gastrointestinal Cancer: Genomics and Bioinformatics Core

NIH 7P01 CA95616 DePinho (PI) 03/01/08-02/28/13
Genetics and Biology of Malignant Glioma
This PO1 endeavors to improve understanding of the pathogenesis of glioblastoma multiforme (GBM). Project 3's goal focuses on identification of novel GBM genes through large-scale functional genomics.
Role: Project PI

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Comparative analysis of mutational patterns in Mitochondrial DNA	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Han Liang, The University of Texas MD Anderson Cancer Center; Member of TCGA Endometrial, Lung, Gastric and Pan-Cancer Working Groups; Chair, TCGA Pan-Cancer Clinical/Predictor Working Group	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Leng Han, UT MD Anderson Cancer Center Jun Li, UT MD Anderson Cancer Center	

Background and preliminary data

Mitochondria play an important role in cellular energy metabolism, free radical generation, and apoptosis. Mitochondrial DNA (mtDNA) is a maternally-inherited 16,569-bp closed-circle genome that encodes two rRNAs, 22 tRNAs, and 10 polypeptides. Dysfunctions in mitochondrial function are an important cause of many neurological diseases and drug toxicities and may contribute to carcinogenesis and tumor progression. Furthermore, the mitochondrial genome is a fundamental tool for human population genetics and has played a critical role in mapping the migration of humanity across the globe. Despite that functional importance, mtDNA is not targeted in any of the currently used genome/exome sequencing methods, and often ignored in DNA-seq analysis. However, mtDNA sequences are available in whole genome sequencing data. One major challenge is to distinguish the mtDNA from nuclear copies of the mitochondrial genomes (nuMTS). Our group has developed a software BM-MAP, which can efficiently deal with this multi-mapping issue.

Timelines & resources dedicated to project

a. Building software – January to June 2014
b. mtDNA data analysis – July 2014 to December 2014
c. Manuscript preparation – January to February 2015
d. Manuscript submission – 20th March 2015.

Two postdoc fellows will work on this project. In additional to the computational resources provided by ICGC, we have computing resources such as MD Anderson High Performance Clusters.

Research proposal

We aim to develop an open-source software tool that can reliably and easily extract mitochondrial genome information from exome and/or whole genome sequencing data. This tool will be able to evaluate mitochondrial genome alignment quality, estimate relative mitochondrial copy numbers and detect heteroplasmy, somatic mutations and other structural variants of the mitochondrial genome. Considering the large number of samples in ICGC & TCGA, our software will be set up to run in parallel or serial on large DNA sequencing datasets.

Furthermore, an important complication in aligning DNA reads to the mitochondrial genome is the presence of nuclear copies of the mitochondrial genomes (nuMTS). The nuMTS can cause ambiguity about whether a read should map to the nuclear or the mitochondrial genome. Our experience in developing the efficient mapping software, BM-Map will facilitate this critical step in the analysis.

We will systematically characterize the mutations and copy number variations in the mitochondrial genome in the ICGC tumor samples. Then we will perform the comparative analyses on mutation and copy number variation patterns between the mitochondrial genome and the nuclear genome as well as their potential functional connections. Moreover, we will perform the comparative analysis across different tumor types.

Legacy plans

All related data, such as executable code, documentation will be uploaded to the synapse. The software will be open-source.

Han Liang, Ph.D.

Assistant Professor
Department of Bioinformatics and Computational Biology
The University of Texas MD Anderson Cancer Center
1400 Pressler Street, Houston, TX 77030, USA
Lab webpage: http://odin.mdacc.tmc.edu/~hliang1
E-mail: hliang1@mdanderson.org; Telephone: 1-713-745-9815; Fax: 1-713-563-4242

EDUCATION

Ph.D. Quantitative and Computational Biology, **Princeton University**, Princeton, NJ, USA 09/2001–03/2006
B.S. Chemistry, **Peking University**, China 09/1997–07/2001

POSITION

- Faculty Member, Graduate Program in Structural & Computational Biology & Molecular Biophysics **Baylor College of Medicine**, Houston, TX, 03/2011–
- Regulator Member, **The University of Texas Graduate School of Biomedical Sciences at Houston,** 09/2010–
- Assistant Professor, Department of Bioinformatics and Computational Biology, **The University of Texas M. D. Anderson Cancer Center**, Houston, TX
- Postdoctoral Research Scholar, Department of Ecology and Evolution, **The University of Chicago**, Chicago, IL, 05/2006–06/2009. Advisor: Wen-Hsiung Li

SELECTED PUBLICATIONS (as an Assistant Professor; 41 publications in career; *corresponding author)

1. Yang Y, Han L, Yuan Y, Li J, Hei N, **Liang H***. (2013) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. **Nat Commun** (in press)
2. Piao H, Yuan Y, Wang M, Sun Y, **Liang H**, Ma L. (2013) α-catenin suppresses tumorigenesis by inhibiting NF-κB signaling in E-cadherin-negative basal-like breast cancer cells. **Nat Cell Biol** (in press)
3. Cancer Genome Atlas Research Network (including **Liang H**), Weinstein JN, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. (2013) **Nat Genet** 45: 1113-1120
4. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Friend S, Stuart J, **Liang H**, Margolin AA. Enabling Transparent and Collaborative Analysis of 12 tumor types within The Cancer Genome Atlas. (2013) **Nat Genet** 45: 1121-1126
5. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RGW, Kane DW, Wakefield C, Weinstein JN, Mills GB*, **Liang H***. (2013) TCPA: A Resource for Cancer Functional Proteomics Data. **Nat Methods** 10 (11): 1046-47
6. The Cancer Genomics Research Network (including **Liang H** as a key contributor) (2013) Integrated Genomic Characterization of Endometrial Carcinoma. **Nature** 497:67-73
7. Li Y, Zhang L, Ball RL, Liang X, Li J, **Liang H***. (2012) Comparative Analysis on Somatic Copy-Number Alterations Across Different Types of Human Cancer Reveals Two Distinct Classes of Breakpoint Hotspots. **Hum Mol Genet** 21(22):4957-65.
8. Chen D, Sun Y, Wei Y, Zhang P, Rezaeian AH, Teruya-Feldstein J, Gupta S, **Liang H**, Lin H-K, Hung M-C, Ma L. (2012) LIFR is a Breast Cancer Metastasis Suppressor Upstream of the Hippo-YAP Pathway and a Prognostic Marker. **Nat Med** 18(10):1511-17
9. **Liang H***, Cheung LWT, Li J, Ju Z, Yu S, Stemke-Hale K, Dogruluk T, Lu Y, Liu X, Gu C, Guo W, Scherer SE, Carter H, Westin SN, Dyer MD, Verhaak RGW, Zhang F, Karchin R, Liu GC, Lu KH, Broaddus RR, Scott KL, Hennessy BT, Mills GB. (2012) Whole-exome Sequencing Combined with Functional Genomics Reveals Novel Candidate Driver Cancer Genes in Endometrial Cancer. **Genome Res** 22(11): 2120-29
10. Li J, Roebuck P, Grünewald S, **Liang H***. (2012) SurvNet: a Web Server for Identifying Network-based Biomarkers that Most Correlate with Patient Survival Data. **Nucleic Acids Res** 40 (W): W123-126
11. Kim YH*, **Liang H***, Liu X, Lee J, Cho JY, Cheong JH, Kim H, Li M, Downey TJ, Sun Y, Sun J, Dyer MD, Beasley EM, Noh SH, Weinstein JN, Liu CG, Powis G. (2012) AMPKα Modulation in Cancer Progression: Multilayer Integrative Transcriptome Analysis in Asian Gastric Cancer. **Cancer Res** 72(10): 2512-21
12. Yuan Y, Xu Y, Xu J, Ball RL, **Liang H*** (2012) Predicting Lethal Phenotype of Knockout Mouse by Integrating Comprehensive Genomic Data. **Bioinformatics** 28 (9): 1246-1252

LENG HAN, Ph.D.

Department of Bioinformatics and Computational Biology
MD Anderson Cancer Center, University of Texas
E-mail: lhan1@mdanderson.org

EDUCATION:
Ph.D., Genetics & Bioinformatics, 2010
Chinese Academy of Sciences, Kunming Institute of Zoology, Kunming, China
Advisors: Drs. Zhongming Zhao and Bing Su

B.S., Biotechnology, 2005
Wuhan University, Wuhan, China
Advisor: Dr. Zhongming Zhao

RESEARCH EXPERIENCES:
MD Anderson Cancer Center, University of Texas 2/2012-Present
Post-doctoral Fellow, Department of Bioinformatics and Computational Biology,
Advisor: Dr. Han Liang

Stanford University 8/2010-2/2012
Post-doctoral Fellow, Department of Radiology, School of Medicine,
Advisor: Dr. Joseph C. Wu

Vanderbilt University & Virginia Commonwealth University 3/2008-8/2010
Bioinformatics Engineer, Genomics Sciences Resource (GSR) & Bioinformatics Resource Center (BRC)
Visiting Scholar, Virginia Institute for Psychiatric and Behavioral Genetics
Advisor: Dr. Zhongming Zhao

SELECTED PUBLICATIONS (*Equal contribution)**:**

1. The Cancer Genome Atlas Research Network (including Han L): The Cancer Genome Atlas Pan-Cancer analysis project. **Nature Genetics** 2013, 45: 1113-1120
2. Dey D*, Han L*, Oikonomopoulos A, Sanada F, Hosoda T, Unno K, Almeida PD, Leri A, Wu JC: Dissecting the Molecular Relationship Between Cardiac-Derived and Bone Marrow-Derived Progenitor Cells. **Circulation Research** 2013,112:1253-1262
3. Lan F*, Lee AS*, Liang P*, Sanchez-Freire V, Nguyen PK, Wang L, Han L, Yen M, Wang Y, Sun N *et al*: Abnormal Calcium Handling Properties Underlie Familial Hypertrophic Cardiomyopathy Pathology in Patient-Specific Induced Pluripotent Stem Cells. **Cell Stem Cell** 2013,12:101-113
4. Xia J*, Han L*, Zhao Z: Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. **BMC Genomics**, 2012, 13: S7
5. Du X*, Han L*, Guo A, Zhao Z: Features of Methylation and Gene Expression in the Promoter-Associated CpG Islands Using Human Methylome Data. **Comparative and Functional Genomics** 2012, 598987.
6. Han L, Zhao Z: CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? **BMC Bioinformatics** 2009, 10:65.
7. Han L, Zhao Z: Contrast features of CpG islands in the promoter and other regions in the dog genome. **Genomics** 2009, 94:117-124.
8. Han L, Zhao Z: Comparative analysis of CpG islands in four fish genomes. **Comparative and Functional Genomics** 2008:565631.
9. Han L, Su B, Li WH, Zhao Z: CpG island density and its correlations with genomic features in mammalian genomes. **Genome Biology** 2008, 9:R79.
10. Jiang C*, Han L*, Su B, Li WH, Zhao Z: Features and trend of loss of promoter-associated CpG islands in the human and mouse genomes. **Molecular Biology and Evolution** 2007, 24:1991-2000.

Jun Li
Department of Bioinformatics and Computational Biology
The University of Texas MD Anderson Cancer Center
1400 Pressler Street, Houston, Texas 77030, U.S.A.
Email: jli14@mdanderson.org

EDUCATION

- Ph.D. in Computational Biology, CAS-MPG Partner Institute for Computational Biology (PICB), Key Laboratory of Computational Biology, Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai, China, 2008-2013.
- B.S. in Mathematics, Wuhan University, Wuhan, China, 2004-2008.

POSITION

- Postdoctoral Fellow, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, 2013 -

PROFESSIONAL ACTIVITIES

- Editorial Board, Journal of Bioinformatics and Comparative genomics, 2013 -

SELECTED AWARDS AND HONORS

- Huirui Fellowship, Shanghai Institutes for Biological Sciences, 2012
- Trainee Research Day GSTPC Poster Contest Finalist, University of Texas MD Anderson Cancer Center, 2012
- Outstanding Student Honor of CAS-MPG Partner Institute for Computational Biology, 2010
- Outstanding Student Scholarship of Wuhan University, 2006
- Outstanding Student Scholarship of Wuhan University, 2005

Selected PUBLICATIONS

1. The Cancer Genome Atlas Research Network (including **Li, J.**), Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, L., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. **Nature Genetics** 45, 1113-1120.
2. Yang, Y., Han, L., Yuan, Y., **Li, J.**, Hei, N. and Liang, H. (2013) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. **Nature Communications** (in press)
3. **Li, J.***, Lu, Y.*, Akbani, R., Ju, Z., Roebuck, P., Liu, W., Yang, J.Y., Broom, B., Verhaak, R., Kane, D., Wake_eld, C., Weinstein, J., Mills, G. and Liang, H. (2013) TCPA: A Resource for Cancer Functional Proteomics Data. **Nature Methods** 10: 1046-1047 (*co-first authors)
4. Yang, J.*, **Li, J.***, Grunewald, S. and Wan, X. (2013) BinAligner: a heuristic method to align biological networks. **BMC Bioinformatics** 14(Suppl 14):S8 (*co-first authors)
5. **Li, J.**, Roebuck, P., Grunewald, S. and Liang, H. (2012) SurvNet: A bioinformatics tool for identifying network-based biomarkers that most correlate with patient survival data. **Nucleic Acids Research** 40(W): W123-126
6. Yang, J., **Li, J.**, Dong, L. and Grunewald, S. (2011) Analysis on the reconstruction accuracy of the Fitch method for inferring ancestral states. **BMC Bioinformatics** 12:18
7. Yang, J., **Li, J.**, Dong, L. and Grunewald, S. (2011) A heuristic Algorithm to Align Protein Interaction Networks. **Journal of Biomathematics** 26(3):569-575
8. Li J, Yang J, Li F, Hu K, Dong L and Grunewald S. (2010) Pairwise Alignment of Protein-Protein Interaction by Linear Programming. **Acta Biophysica Sinica**, 29(1)

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Systematic Characterization of RNA Editing Patterns in Human Cancer	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Han Liang, The University of Texas MD Anderson Cancer Center; Member of TCGA Endometrial, Lung, Gastric and Pan-Cancer Working Groups; Chair, TCGA Pan-Cancer Clinical/Predictor Working Group	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Leng Han, UT MD Anderson Cancer Center Yuan Yuan, Baylor College of Medicine	

Background and preliminary data

As an important epigenetic control, RNA editing is a widespread post-transcriptional mechanism that confers specific and reproducible nucleotide changes in selected RNA transcripts. In contrast to somatic mutations, "mutations" at the RNA level caused by RNA editing have received little attention in cancer in general. RNA editing events can result in recurrent changes in protein sequences that can function as oncogenic driver events (e.g., Chen et al. Nature Medicine 2013) and in alternative splicing patterns. Thus, the analysis of RNA editing represents a new paradigm with tremendous potential for discovering biomarkers and therapeutic targets.

We aim to systematically characterize the RNA editing patterns and identify clinically relevant RNA editing events. Through the analysis of TCGA genomic data, my group has accumulated considerable experience in analyzing RNA editing events in a large number of patient samples across tumor types. But due to the limitations of exome-seq data, our analyses have focused on protein-coding regions. With the availability of ICGC WGS data, we plan to extend this analysis to the whole-genome level.

Timelines & resources dedicated to project

a. RNA editing calling – January to June 2014
b. Data analysis – July 2014 to December 2014
c. Manuscript preparation – January to February 2015
d. Manuscript submission – 20th March 2015.

One postdoc fellow and one Ph.D student will work on this project. In additional to the computational resources provided by ICGC, we have computing resources such as MD Anderson High Performance Clusters. We also have experimental collaborations to functionally investigate the RNA editing candidates of interest.

Research proposal

We propose to characterize RNA editing patterns in different cancer types using ICGC sequencing data. We will develop a computational pipeline to identify the RNA editing across tumor samples by comparing the RNA-seq and whole-genome sequencing data. We will filter SNPs from dbSNP and 1000 Genome project, and mutations from TCGA and COSMIC. Moreover, from the sequencing data, we will also be able to quantify the RNA editing level. We will detect and quantify RNA editing events in all ICGC tumor samples in a systematic way.

We will compare the RNA editing patterns in tumor samples and matched normal samples as well as across different cancer types. Importantly, we will systematically identify clinically relevant RNA editing sites, for example, those sites with a differential RNA editing activity among tumor subtypes or correlate with patient survival. We will elucidate the impact of different RNA editing enzymes on the editing events (e.g., for A-to-I events, ADAR1 and ADAR2). We will prioritize the RNA editing events through cross-tumor bioinformatics analyses and identify RNA editing events with a potential functional effect in protein sequence, splicing patterns and gene regulation.

For the RNA editing events of particular interest, we will perform further functional investigation. We will assess their effect on tumor growth and proliferation through highly sensitive cell viability assays and elucidate their affected signaling pathways through functional proteomics and cell line studies. The knowledge gained from the proposed research study will not only fundamentally advance our understanding of the role of RNA editing in tumorigenesis, but also facilitate the development and implementation of novel biomarkers and targets.

Legacy plans

All related data, such as executable code, documentation will be uploaded to the synapse. We will also build a comprehensive database to include all RNA editing sites, especially functionally important RNA editing sites.

Han Liang, Ph.D.

Assistant Professor
Department of Bioinformatics and Computational Biology
The University of Texas MD Anderson Cancer Center
1400 Pressler Street, Houston, TX 77030, USA
Lab webpage: http://odin.mdacc.tmc.edu/~hliang1
E-mail: hliang1@mdanderson.org; Telephone: 1-713-745-9815; Fax: 1-713-563-4242

EDUCATION

Ph.D. Quantitative and Computational Biology, **Princeton University**, Princeton, NJ, USA 09/2001–03/2006
B.S. Chemistry, **Peking University**, China 09/1997–07/2001

POSITION

- Faculty Member, Graduate Program in Structural & Computational Biology & Molecular Biophysics **Baylor College of Medicine**, Houston, TX, 03/2011–
- Regulator Member, **The University of Texas Graduate School of Biomedical Sciences at Houston,** 09/2010–
- Assistant Professor, Department of Bioinformatics and Computational Biology, **The University of Texas M. D. Anderson Cancer Center**, Houston, TX
- Postdoctoral Research Scholar, Department of Ecology and Evolution, **The University of Chicago**, Chicago, IL, 05/2006–06/2009. Advisor: Wen-Hsiung Li

SELECTED PUBLICATIONS (as an Assistant Professor; 41 publications in career; *corresponding author)

1. Yang Y, Han L, Yuan Y, Li J, Hei N, **Liang H***. (2013) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. **Nat Commun** (in press)
2. Piao H, Yuan Y, Wang M, Sun Y, **Liang H**, Ma L. (2013) α-catenin suppresses tumorigenesis by inhibiting NF-κB signaling in E-cadherin-negative basal-like breast cancer cells. **Nat Cell Biol** (in press)
3. Cancer Genome Atlas Research Network (including **Liang H**), Weinstein JN, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. (2013) **Nat Genet** 45: 1113-1120
4. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Friend S, Stuart J, **Liang H**, Margolin AA. Enabling Transparent and Collaborative Analysis of 12 tumor types within The Cancer Genome Atlas. (2013) **Nat Genet** 45: 1121-1126
5. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RGW, Kane DW, Wakefield C, Weinstein JN, Mills GB*, **Liang H***. (2013) TCPA: A Resource for Cancer Functional Proteomics Data. **Nat Methods** 10 (11): 1046-47
6. The Cancer Genomics Research Network (including **Liang H** as a key contributor) (2013) Integrated Genomic Characterization of Endometrial Carcinoma. **Nature** 497:67-73
7. Li Y, Zhang L, Ball RL, Liang X, Li J, **Liang H***. (2012) Comparative Analysis on Somatic Copy-Number Alterations Across Different Types of Human Cancer Reveals Two Distinct Classes of Breakpoint Hotspots. **Hum Mol Genet** 21(22):4957-65.
8. Chen D, Sun Y, Wei Y, Zhang P, Rezaeian AH, Teruya-Feldstein J, Gupta S, **Liang H**, Lin H-K, Hung M-C, Ma L. (2012) LIFR is a Breast Cancer Metastasis Suppressor Upstream of the Hippo-YAP Pathway and a Prognostic Marker. **Nat Med** 18(10):1511-17
9. **Liang H***, Cheung LWT, Li J, Ju Z, Yu S, Stemke-Hale K, Dogruluk T, Lu Y, Liu X, Gu C, Guo W, Scherer SE, Carter H, Westin SN, Dyer MD, Verhaak RGW, Zhang F, Karchin R, Liu GC, Lu KH, Broaddus RR, Scott KL, Hennessy BT, Mills GB. (2012) Whole-exome Sequencing Combined with Functional Genomics Reveals Novel Candidate Driver Cancer Genes in Endometrial Cancer. **Genome Res** 22(11): 2120-29
10. Li J, Roebuck P, Grünewald S, **Liang H***. (2012) SurvNet: a Web Server for Identifying Network-based Biomarkers that Most Correlate with Patient Survival Data. **Nucleic Acids Res** 40 (W): W123-126
11. Kim YH*, **Liang H***, Liu X, Lee J, Cho JY, Cheong JH, Kim H, Li M, Downey TJ, Sun Y, Sun J, Dyer MD, Beasley EM, Noh SH, Weinstein JN, Liu CG, Powis G. (2012) AMPKα Modulation in Cancer Progression: Multilayer Integrative Transcriptome Analysis in Asian Gastric Cancer. **Cancer Res** 72(10): 2512-21
12. Yuan Y, Xu Y, Xu J, Ball RL, **Liang H*** (2012) Predicting Lethal Phenotype of Knockout Mouse by Integrating Comprehensive Genomic Data. **Bioinformatics** 28 (9): 1246-1252

LENG HAN, Ph.D.

Department of Bioinformatics and Computational Biology
MD Anderson Cancer Center, University of Texas
E-mail: lhan1@mdanderson.org

EDUCATION:
Ph.D., Genetics & Bioinformatics, 2010
Chinese Academy of Sciences, Kunming Institute of Zoology, Kunming, China
Advisors: Drs. Zhongming Zhao and Bing Su

B.S., Biotechnology, 2005
Wuhan University, Wuhan, China
Advisor: Dr. Zhongming Zhao

RESEARCH EXPERIENCES:
MD Anderson Cancer Center, University of Texas 2/2012-Present
Post-doctoral Fellow, Department of Bioinformatics and Computational Biology,
Advisor: Dr. Han Liang

Stanford University 8/2010-2/2012
Post-doctoral Fellow, Department of Radiology, School of Medicine,
Advisor: Dr. Joseph C. Wu

Vanderbilt University & Virginia Commonwealth University 3/2008-8/2010
Bioinformatics Engineer, Genomics Sciences Resource (GSR) & Bioinformatics Resource Center (BRC)
Visiting Scholar, Virginia Institute for Psychiatric and Behavioral Genetics
Advisor: Dr. Zhongming Zhao

SELECTED PUBLICATIONS (*Equal contribution)**:**

1. The Cancer Genome Atlas Research Network (including Han L): The Cancer Genome Atlas Pan-Cancer analysis project. **Nature Genetics** 2013, 45: 1113-1120
2. Dey D*, Han L*, Oikonomopoulos A, Sanada F, Hosoda T, Unno K, Almeida PD, Leri A, Wu JC: Dissecting the Molecular Relationship Between Cardiac-Derived and Bone Marrow-Derived Progenitor Cells. **Circulation Research** 2013,112:1253-1262
3. Lan F*, Lee AS*, Liang P*, Sanchez-Freire V, Nguyen PK, Wang L, Han L, Yen M, Wang Y, Sun N *et al*: Abnormal Calcium Handling Properties Underlie Familial Hypertrophic Cardiomyopathy Pathology in Patient-Specific Induced Pluripotent Stem Cells. **Cell Stem Cell** 2013,12:101-113
4. Xia J*, Han L*, Zhao Z: Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. **BMC Genomics**, 2012, 13: S7
5. Du X*, Han L*, Guo A, Zhao Z: Features of Methylation and Gene Expression in the Promoter-Associated CpG Islands Using Human Methylome Data. **Comparative and Functional Genomics** 2012, 598987.
6. Han L, Zhao Z: CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? **BMC Bioinformatics** 2009, 10:65.
7. Han L, Zhao Z: Contrast features of CpG islands in the promoter and other regions in the dog genome. **Genomics** 2009, 94:117-124.
8. Han L, Zhao Z: Comparative analysis of CpG islands in four fish genomes. **Comparative and Functional Genomics** 2008:565631.
9. Han L, Su B, Li WH, Zhao Z: CpG island density and its correlations with genomic features in mammalian genomes. **Genome Biology** 2008, 9:R79.
10. Jiang C*, Han L*, Su B, Li WH, Zhao Z: Features and trend of loss of promoter-associated CpG islands in the human and mouse genomes. **Molecular Biology and Evolution** 2007, 24:1991-2000.

Yuan Yuan

7675 Phoenix Dr, Apt 446
Houston, TX, 77030

Tel: 806-252-3924
Email: yy2@bcm.edu

EDUCATION

Ph.D. Candidate in Computational Biology (SCBMB) 08/2011-present
Baylor College of Medicine, Houston, TX
Overall GPA: 4.0/4.0

Master of Science in Mathematics 06/2008-08/2010
Texas Tech University, Lubbock, TX
Overall GPA: 4.0/4.0

Master of Science in Biology 08/2006-05/2009
Texas Tech University, Lubbock, TX
Overall GPA: 4.0/4.0

Bachelor of Science in Life Sciences/ Chu Kochen Honors College 10/2002-06/2006
Zhejiang University, Hangzhou, Zhejiang, China
Overall GPA: 3.7/4.0 Major GPA: 4.0/4.0 Rank: Top 5%

HONORS & AWARDS

- WiML 2013 Travel Award 12/2013
- Professor John J. Trentin Scholarship Award, BCM 10/2012
- Texas Tech Mathematics Graduate Scholarship 09/2009
- Scholarship for Academic Excellence, Zhejiang University 10/2005
- Scholarship for Academic Excellence, Zhejiang University 10/2004

SELECTED PUBLICATIONS

- Piao H, **Yuan Y**, Wang M, Sun Y, Liang H, Ma L. α-catenin suppresses tumorigenesis by inhibiting NF-κB signaling in E-cadherin-negative basal-like breast cancer cells. ***Nature Cell Biology*** (in press)
- Yang Y, Han L, **Yuan Y**, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. ***Nature Communications*** (in press)
- Omberg L, Ellrott K, **Yuan Y**, Kandoth C, The Cancer Genome Atlas Research Network, Friend S, Stuart J, Liang H, Margolin AA, Transparent and collaborative analysis of 12 tumor types within The Cancer Genome Atlas. ***Nature Genetics***. 2013, 45 (10):1113-1120
- The Cancer Genome Atlas Research Network (including **Yuan Y**), Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart J, The Cancer Genome Atlas Pan-Cancer analysis project. ***Nature Genetics***. 2013, 45: 1113-1120
- **Yuan Y**, Norris C, Xu Y, Tsui K-W, Ji Y and Liang H, BM-Map: an efficient software package for accurately allocating multireads of RNA-sequencing data. ***BMC Genomics***. 2012, 13(Suppl 8):S9
- **Yuan Y**, Xu Y, Xu J, Ball R and Liang H. Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. ***Bioinformatics***. 2012 May 1; 28(9):1246-52

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Decoding the role of mutations in regulatory elements: systematic eQTL analysis across tumor types

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Han Liang, The University of Texas MD Anderson Cancer Center; Member of TCGA Endometrial, Lung, Gastric and Pan-Cancer Working Groups; Chair, TCGA Pan-Cancer Clinical/Predictor Working Group

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yang Yang, University of Texas MD Anderson Cancer Center& UT School of Public Health	Peng Wei, University of Texas School of Public Health

Background and preliminary data

One central question in cancer genomics is to understand the functional effects of somatic mutations in the context of tumor cells. Previous studies have focused on mutations in protein-coding regions. In contrast, the roles of mutations in non-coding regions remain largely unknown. In particular, mutations in the regulatory elements of a gene may directly modify the expression of the related gene transcripts, thereby imposing a critical effect on the cancer phenotype. The combination of whole-genome genetic association studies and the measurement of global gene expression, eQTL (expression quantitative trait loci) analysis, allows the systematic identification of such mutations. Integrating the ICGC somatic mutations (both point mutations and indels) and gene expression (RNA-seq and microarray data), we aim to systematically identify eQTL across tumor types. Our team has considerable experience in Pan-Cancer and eQLT analyses. In particular, we have developed cutting-edge statistical and computational methods to detect multiple mutation-multiple phenotype associations. Our novel methods hinge on the powerful and versatile generalized estimation equations (GEE) and adaptive sum of powered score test (aSPU). We have successfully applied our developed methods to discover associations between rare germline mutations and multiple blood lipids using WGS data.

Timelines & resources dedicated to project

a. Methodology development – January to June 2014
b. Data analysis – July 2014 to December 2014
c. Manuscript preparation – January to February 2015
d. Manuscript submission – 20th March 2015.

One postdoc fellow and one Ph.D student will work on this project. In additional to the computational resources provided by ICGC, we have computing resources such as MD Anderson High Performance Clusters.

Research proposal

We will develop a novel statistical method for eQTL analysis for Pan-Cancer genomic/transcriptomic data. Since the frequency of somatic mutations is usually low, our method will consider multiple point mutations, as defined by, for example, the regulatory region of a gene, simultaneously to detect multiple mutation-multiple phenotype associations. We propose to capitalize on the ENCODE project resources to group point mutations and indels into functional units and test their effects on multiple genes' expression profiles. Our novel method hinge on the powerful and versatile generalized estimation equations (GEE) and adaptive sum of powered score test (aSPU). Moreover, the new method will consider the sample heterogeneity across tumor types. Through the simulation, we will compare the performance of our methods with alternative methods.

We will apply our methods to ICGC WGS data (including both single nucleotide variations and indels) and gene expression data (including both RNA-seq and microarray data) across ICGC tumor types. We propose to capitalize on the ENCODE project resources to group point mutations and indels into functional units and test their effects on multiple genes' expression profiles defined by, for example, biological pathways, or the same gene's expression across multiple cancer types.

Given the high-confidence candidates we identified, we will perform higher-level functional analysis such as pathway or module/network analysis. We will perform the comparative analysis to estimate the relative abundance of eQTL sites and investigate the functional consequences in different tumor contexts. For the candidates of particular interest, we will validate their functional roles through experimental investigation.

Legacy plans

All related data, such as executable code, documentation will be uploaded to the synapse.
The scripts or related software package will be open-source and publically available.

Han Liang, Ph.D.

Assistant Professor
Department of Bioinformatics and Computational Biology
The University of Texas MD Anderson Cancer Center
1400 Pressler Street, Houston, TX 77030, USA
Lab webpage: http://odin.mdacc.tmc.edu/~hliang1
E-mail: hliang1@mdanderson.org; Telephone: 1-713-745-9815; Fax: 1-713-563-4242

EDUCATION

Ph.D. Quantitative and Computational Biology, **Princeton University**, Princeton, NJ, USA 09/2001–03/2006
B.S. Chemistry, **Peking University**, China 09/1997–07/2001

POSITION

- Faculty Member, Graduate Program in Structural & Computational Biology & Molecular Biophysics **Baylor College of Medicine**, Houston, TX, 03/2011–
- Regulator Member, **The University of Texas Graduate School of Biomedical Sciences at Houston,** 09/2010–
- Assistant Professor, Department of Bioinformatics and Computational Biology, **The University of Texas M. D. Anderson Cancer Center**, Houston, TX
- Postdoctoral Research Scholar, Department of Ecology and Evolution, **The University of Chicago**, Chicago, IL, 05/2006–06/2009. Advisor: Wen-Hsiung Li

SELECTED PUBLICATIONS (as an Assistant Professor; 41 publications in career; *corresponding author)

1. Yang Y, Han L, Yuan Y, Li J, Hei N, **Liang H***. (2013) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. **Nat Commun** (in press)
2. Piao H, Yuan Y, Wang M, Sun Y, **Liang H**, Ma L. (2013) α-catenin suppresses tumorigenesis by inhibiting NF-κB signaling in E-cadherin-negative basal-like breast cancer cells. **Nat Cell Biol** (in press)
3. Cancer Genome Atlas Research Network (including **Liang H**), Weinstein JN, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. (2013) **Nat Genet** 45: 1113-1120
4. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Friend S, Stuart J, **Liang H**, Margolin AA. Enabling Transparent and Collaborative Analysis of 12 tumor types within The Cancer Genome Atlas. (2013) **Nat Genet** 45: 1121-1126
5. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RGW, Kane DW, Wakefield C, Weinstein JN, Mills GB*, **Liang H***. (2013) TCPA: A Resource for Cancer Functional Proteomics Data. **Nat Methods** 10 (11): 1046-47
6. The Cancer Genomics Research Network (including **Liang H** as a key contributor) (2013) Integrated Genomic Characterization of Endometrial Carcinoma. **Nature** 497:67-73
7. Li Y, Zhang L, Ball RL, Liang X, Li J, **Liang H***. (2012) Comparative Analysis on Somatic Copy-Number Alterations Across Different Types of Human Cancer Reveals Two Distinct Classes of Breakpoint Hotspots. **Hum Mol Genet** 21(22):4957-65.
8. Chen D, Sun Y, Wei Y, Zhang P, Rezaeian AH, Teruya-Feldstein J, Gupta S, **Liang H**, Lin H-K, Hung M-C, Ma L. (2012) LIFR is a Breast Cancer Metastasis Suppressor Upstream of the Hippo-YAP Pathway and a Prognostic Marker. **Nat Med** 18(10):1511-17
9. **Liang H***, Cheung LWT, Li J, Ju Z, Yu S, Stemke-Hale K, Dogruluk T, Lu Y, Liu X, Gu C, Guo W, Scherer SE, Carter H, Westin SN, Dyer MD, Verhaak RGW, Zhang F, Karchin R, Liu GC, Lu KH, Broaddus RR, Scott KL, Hennessy BT, Mills GB. (2012) Whole-exome Sequencing Combined with Functional Genomics Reveals Novel Candidate Driver Cancer Genes in Endometrial Cancer. **Genome Res** 22(11): 2120-29
10. Li J, Roebuck P, Grünewald S, **Liang H***. (2012) SurvNet: a Web Server for Identifying Network-based Biomarkers that Most Correlate with Patient Survival Data. **Nucleic Acids Res** 40 (W): W123-126
11. Kim YH*, **Liang H***, Liu X, Lee J, Cho JY, Cheong JH, Kim H, Li M, Downey TJ, Sun Y, Sun J, Dyer MD, Beasley EM, Noh SH, Weinstein JN, Liu CG, Powis G. (2012) AMPKα Modulation in Cancer Progression: Multilayer Integrative Transcriptome Analysis in Asian Gastric Cancer. **Cancer Res** 72(10): 2512-21
12. Yuan Y, Xu Y, Xu J, Ball RL, **Liang H*** (2012) Predicting Lethal Phenotype of Knockout Mouse by Integrating Comprehensive Genomic Data. **Bioinformatics** 28 (9): 1246-1252

Yang Yang
7900 Cambridge ST, APT.19-2D, Houston, TX, 77054, USA
xyy2006@msn.com; 765-602-9740

EDUCATION BACKGROUND

12/2014: PhD student in Biostatistics (4th year, qualification exam passed), School of Public Health, University of Texas Health Science Center at Houston, Houston, TX
GPA: 3.9

06/2010: M.S. Bioinformatics, School of Informatics, Indiana University-Purdue University Indianapolis, IN
GPA: 3.5
M.S Thesis: "Effects of estrogen on large intergenic non-coding RNA in breast cancer cells."

07/2008: M.A. Theoretic Veterinary Medicine, College of Veterinary Medicine, Nanjing Agricultural University, Nanjing, CN (No degree received)
Average score: 87.31/100

07/2006: B.S. National Life Science and Technique Talent-training Base, Nanjing Agricultural University, Nanjing, CN
Average score: 80/100
B.S Thesis: Fiber Analysis of Lateral Line Nerve in Silurus Asotus

PUBLICATIONS/MANUSCRIPTS UNDER REVISION

1. **Yang, Y**.*, Wei, P.*, Guo, X., Zhou, D., Assassi, S., Zhou, X. (2013) Impact of age and autoantibody status on the gene expression of scleroderma fibroblasts in response to silica stimulation. *European Journal of Inflammation* Vol.11, no.3, 631-639 (*contributed equally).
2. Wei, P.*, **Yang, Y**.*, Guo, X., Lai, S., Assassi, S., Tan, F., Zhou, X. (2013) Integrative studies of genetic and environmental factors in primary human fibroblasts in scleroderma. *Human Molecular Genetics* (in minor revision) (*contributed equally).
3. **Yang, Y.**, Han, L., Yuan, Y., Li, J., Hei, N., Liang, H. (2013) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications* (in press).
4. Leng Han, Yuan Yuan, Siyuan Zheng, Mary E. Edgerton, **Yang Yang**, Lixia Diao, Jun Li, Yanxun Xu, Roeland G.W. Verhaak, The Cancer Genome Atlas Research Network, Han Liang. (2013) Pseudogene Expression Defines Biologically and Clinically Relevant Cancer Subtypes. *Nature Communications* (in major revision).
5. The Cancer Genome Atlas Research Network (including **Yang Yang**), John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander & Joshua M Stuart. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45, 1113–1120, doi:10.1038/ng.2764.
6. Hei Nainan, Yang Ping, **Yang Yang**, Liu Jinxiong, Bao Huijun, Liu Haili, Zhang Hui and Chen Qiusheng (2010). Fine structural observation on the oogenesis and vitellogenesis of the Chinese soft-shelled turtle (Pelodiseus sinensis). Zygote, 18, pp 109-120. doi:10.1017/S0967199409990116.

REFERENCES AVAILABLE UPON REQUEST

CURRICULUM VITAE

Contact information:
Peng Wei, Ph.D.
1200 Pressler Street, RAS E-817, Houston, TX 77030
Phone: (713) 500- 9565
Email: Peng.Wei@uth.tmc.edu

EDUCATION

2009 **PhD, Biostatistics,** University of Minnesota, Minneapolis, MN, USA
2004 **BS, Statistics,** School of Mathematical Sciences, Peking University, China

PROFESSIONAL EXPERIENCE

Assistant Professor **July 2009 – Present**
Division of Biostatistics and Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston

SELECTED PEER REVIEWED PUBLICATIONS

- Cao, Y.*, **Wei, P.***, Bailey, M., Kauwe, J.S.K., Maxwell, T.J. (2014) A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology*, 38: 51-59. (*contributed equally).
- Tang, H.**, **Wei, P.**, Duell, E.J., Risch, H.A., Olson, S.H., Bueno-de-Mesquita, H.B., *et al* (2013) Genes-environment interactions in obesity- and diabetes-associated pancreatic cancer risk: A GWAS data analysis. *Cancer Epidemiology, Biomarkers & Prevention* (in press) (** graduate student of Wei).
- **Wei, P.,** Pan, W. (2012) Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. *The Annals of Applied Statistics*, **6**:334-355.
- **Wei, P.,** Liu, X. Fu,Y.X. (2011) Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study. *BMC Proceedings*, **5**:S20.
- Chen, G.K.*, **Wei, P**.*, DeStefano, A.L. (2011) Incorporating biological information into association studies of sequencing data. *Genetic Epidemiology*, **35**:S29-S34 (*contributed equally).
- **Wei, P.**, Pan, W. (2010) Network-based genomic discovery: application and comparison of Markov random field models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:105-125.
- **Wei, P.**, Pan, W. (2008) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, **24**(3):404-411.
- **Wei, P.**, Pan, W. (2008) Incorporating gene functions into regression analysis of DNA-protein binding data and gene expression data to construct transcriptional networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5**(3):401-415.

SELECTED GRANT SUPPORT

- Title: Genetic Susceptibility and Risk Model for Pancreatic Cancer
  Funding Source: NIH/NCI (R01CA169122)
  Role: Principal Investigator
  Duration: 9/17/2013 – 5/31/2017

- Title: Association Analysis of Rare Variants with Sequencing Data
  Funding Source: NIH/NHLBI (R01HL116720)
  Role: Principal Investigator
  Duration: 9/1/2013 – 5/31/2016

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

The landscape of RNA splicing alterations in human cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Matthew Meyerson, Dana-Farber Cancer Institute/Broad Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Angela Brooks, Dana-Farber Cancer Institute/Broad Institute	

Background and preliminary data

Recent findings from whole-exome sequencing studies have shown that somatic mutations frequently occur in splicing factors across multiple cancer types, supporting the need to systematically and globally characterize splicing alterations across human cancers. Alterations in splicing can arise through somatic mutation in *trans*-acting splicing factors or through mutations in *cis*-acting splice sites or splicing regulatory sequences that are encoded in pre-mRNA. In this work, we will integrate whole-exome, whole-genome, and RNA sequencing to 1) perform a pan-cancer survey of splicing alterations in human cancer, and 2) further our understanding of coding and non-coding somatic mutations that affect RNA splicing.

Towards the goal of characterizing splicing alterations in human cancers, we have developed a computational analysis pipeline called JuncBASE (Brooks *et al.,* Genome Research 2011) to identify and quantify alternative splicing in cancer RNA-Seq data. Using JuncBASE, we performed a pan-cancer analysis of splicing alterations associated with mutations in the splicing factor *U2AF1* and found 438 significantly altered splicing events in lung adenocarcinoma and acute myeloid leukemias (Brooks *et al.* accepted). Through cell line transfection experiments and qPCR validation, we further showed that *U2AF1* mutation causes splicing alterations found in patient samples, including an alternative splicing event in *CTNNB1*. As part of the TCGA lung adenocarcinoma working group, we analyzed RNA-Seq data to identify samples showing skipping of exon 14 in the *MET* oncogene (TCGA Network, under review). Exon 14 of *MET* was previously known to be somatically mutated through splice site mutations and deletions. These mutations lead to in-frame skipping of exon 14, which removes a negative regulatory CBL binding site in the resulting protein, thus causing gene activation. Using JuncBASE, we found *MET* exon 14 skipping in 4% of cases. One of these cases harbored a nonsense mutation within exon 14—a somatic mutation thought to cause loss-of-function and not previously associated with changes in splicing. Somatic mutations of *MET* exon 14 challenges our assumptions of the loss-of-function effects of splice sites and nonsense mutations, as these lead to an activation of the *MET* oncogene.

By performing a comprehensive survey of alternative splicing across multiple human cancers we will identify additional examples of gene alterations through affected RNA splicing. We will also move toward the further functional annotation of coding and non-coding somatic mutations that affect RNA splicing.

Timelines & resources dedicated to project

We plan to complete our pan-cancer computational analysis of splicing alterations in the 12 TCGA Pan-Cancer cancer types in early 2014, followed by a write-up of the results. This project would be particularly dependent on raw RNA-sequencing data as well as somatic variant calls from both whole-exome and whole-genome data.

Research proposal

Through the integration of mRNA, whole-exome and whole genome sequencing we propose to identify RNA splicing alterations in human cancers and investigate the underlying somatic mutations that cause splicing alterations. We propose to do so in the following three ways:

1**) Perform a pan-cancer identification and quantification of alternative splicing observed across human cancers using JuncBASE.**
We are currently analyzing TCGA RNA-Seq data using JuncBASE to identify and quantify alternative splicing. JuncBASE is able to identify known and novel alternative splicing from RNA-Seq data, which is beneficial for looking at splicing alterations which may not be annotated. JuncBASE also classifies the type of alternative splicing (e.g., skipped exon, alternative 5' splice site, intron retention). Finally, JuncBASE quantifies splicing of individual exons by calculating a "percent spliced in" (PSI) value – the proportion of all isoforms that splices in an alternative exon. As part of the TCGA Pan-Cancer Analysis Working group, we have already made available alternative splicing quantification for lung adenocarcinoma and acute myeloid leukemias in Synapse (https://www.synapse.org/#!Synapse:syn1701264) and are committed to sharing these splicing profiles to other members of the consortium. Splicing alterations in genes that are known to be significantly mutated across cancers will be of particular interest for follow up.

2) **Using known models of splice site and splicing regulatory sequences, associate transcriptome changes in splicing and gene expression with whole-exome and whole-genome somatic mutations calls.**
Currently, splice site mutation analysis has been limited to the first and last dinucleotide of introns; however, splice site recognition extends into neighboring intronic and exonic regions. We will identify somatic mutations that occur within these extended splice site sequences as well as within putative splicing regulatory sequences. Splicing regulatory sequences are known to lie in exons; therefore, synonymous and non-synonymous mutations may disrupt these regulatory sequences. Intronic mutations available through whole-genome sequencing will allow us to integrate intronic splicing enhancer and silencer mutations that exist in regions not captured by exome sequencing. We will characterize the effects of these somatic mutations on the transcriptome by associating gene expression changes and/or splicing changes.

3) **Use outlier detection and mutual exclusivity to identify significantly altered splicing events**. Splicing alterations may be missed by using known models of splice sites and splicing regulatory sequences; therefore, we will identify recurrent outlier splicing alterations from RNA-Seq data and look for correlated somatic mutations at the DNA level in *cis* or in *trans*. In collaboration with Benjamin Raphael's group, we will also identify putative driver splicing alterations that are mutually exclusive with somatic mutation in driver genes and pathways through modifications of their Dendrix algorithm (see additional abstract).

This work will have a significant impact on our understanding of the role of splicing in tumorigenesis by identifying additional somatic mutations that affect splicing and characterizing the functional consequence of these mutations on the transcriptome. In addition, computational tools will be further developed and will be available to the scientific community to 1) examine splicing alterations in tumor transcriptome RNA-Seq and 2) to identify somatic mutations in regulatory sequences and splice sites that alter splicing of genes, thus, predicting the impact of mutations beyond protein coding changes.

Legacy plans

A beta version of our tool to analyze alternative splicing in RNA-Seq data, JuncBASE, is currently available at: http://www.broadinstitute.org/cancer/cga/juncbase. The software package will be continuously updated and made available.

Quantification of alternative splicing events will be made available to the research community. Currently, quantification of alternative splicing observed in TCGA lung adenocarcinomas and acute myeloid leukemias is available through the TCGA Pan-Cancer group's Synapse site: https://www.synapse.org/#!Synapse:syn1701264

Curriculum vitae for Matthew Meyerson, M.D., Ph.D. Education and Training

1985 A.B., Chemistry and Physics, Harvard College
1993 M.D., Harvard Medical School
Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996 Resident, Clinical Pathology, Massachusetts General Hospital
Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006 Associate Member, Broad Institute of Harvard and MIT
2005- Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006- Senior Associate Member, Broad Institute of Harvard and MIT
2009- Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

ards and Honors

1999 Pew Scholar in the Biomedical Sciences
2004 Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005 Clinical Investigator Award, American Lung Association
2009 Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010 Team Science Award, American Association for Cancer Research
2011 Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012 Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …, Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-13795.

NAME: Angela N. Brooks	POSITION TITLE Postdoctoral researcher

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
University of California, San Diego	B.S.	2000-2005	Biology w/ Spec. in Bioinformatics
University of California, Berkeley	Ph.D.	2005-2011	Molecular and Cell Biology
Dana-Farber Cancer Institute/Broad Institute	Postdoc	2011-present	Cancer Genomics

A. Personal Statement

My research interest is in identifying genomic alterations that contribute to tumorigenesis. For my postdoctoral work, I have studied the effects of somatic mutations in splicing factors and splicing regulatory elements on alternatively spliced transcripts through analysis of TCGA RNA-Seq data.

B. Positions and Honors

Research Positions:

2011 Postdoctoral Fellow, Steven Brenner Research Group, University of California, Berkeley
2011-present Postdoctoral Fellow, Lab of Matthew Meyerson, Dana-Farber Cancer Institute/Broad Institute

Awards and Honors:

2012-present Merck Fellow of the Damon Runyon Cancer Research Foundation
2007-2010 National Science Foundation Graduate Research Fellowship
2005-2007 Chancellor's Fellowship for Graduate Study at UC Berkeley

C. Selected Peer-reviewed Publications

1. Schattner P, **Brooks AN**, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Research. 2005;33(Web Server issue): W686-9
2. Macrae IJ, Zhou K, Li F, Repic A, **Brooks AN**, Cande WZ, Adams PD, Doudna JA. Structural basis for double-stranded RNA processing by Dicer. Science. 2006; 311(5758):195-8
3. Blanchette M, Green RE, MacArthur S, **Brooks AN**, Brenner SE, Eisen MB, Rio DC. Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the *Drosophila* hnRNP A/B family members. Molecular Cell. 2009;33(4):438-49
4. **Brooks AN***, Yang L*, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, and Graveley BR. Conservation of an RNA Regulatory Map between *Drosophila* and Mammals. Genome Research. 2011; 21:193-202 *equal contribution
5. Graveley BR*, **Brooks AN***, Carlson JW*, Duff MO*, Landolin J*, Yang L*, …, Andrews J, Brenner SE, Brent M, Cherbas P, Gingeras TR, Hoskins RA, Kaufman T, Oliver B, Celniker SE. The Developmental Transcriptome of *Drosophila melanogaster*. Nature. 2011; Mar 24;471(7339):473-9 *equal contribution
6. **Brooks AN***, Aspden JL*, Podgornaia AI, Rio DC, Brenner SE. Identification and experimental validation of splicing regulatory elements in *Drosophila melanogaster* reveals functionally conserved splicing enhancers in metazoans. RNA. 2011; Aug 24 *equal contribution
7. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, Sougnez C,…,**Brooks A**,…,Meyerson M. Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. Cell 2012;150, 1107–1120.
8. **Brooks AN**, Choi PS, Waal L, Sharifnia T,…,Meyerson M. A pan-cancer analysis of transcriptome changes associated with somatic mutations in *U2AF1* reveals commonly altered splicing events. PLOS ONE. *accepted*

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca**by ~~27th November~~31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Relationships between pathogenic infection and genomic alterations in human cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Matthew Meyerson, Dana-Farber Cancer Institute / Broad Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s)& institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Akinyemi Ojesina, DFCI/Broad Institute Chandra Sekhar Pedamallu, DFCI/Broad Institute	

Background and preliminary data

About 15-20% of cancers worldwide are thought to be associated with pathogens. Our overarching premise is that the identification of pathogens in cancer, as well as the elucidation of how somatic alterations synergize with infection in cancer may facilitate the development of novel diagnostic, therapeutic and preventive strategies in cancer. For example, vaccines have been developed based on the association between human papillomaviruses (HPV) and cervical cancer. **Therefore, we have developed a computational pipeline,** PathSeq, for the identification and analyses of microbial sequences in high-throughput sequencing data in a scalable manner. PathSeq has been successfully used in identification of microbial sequences in transcriptome, exome and whole genome sequencing data derived from several large-scale studies including TCGA projects on head and neck cancer, stomach cancer, thyroid cancer, melanoma, etc. We have also recently completed whole exome, whole genome and transcriptome analyses of cervical cancer from Norway and Mexico (Ojesina*et al*. Nature DOI: 10.1038/nature12881). In addition to the discovery of novel somatic mutations in *HLA-B*, *MAPK1* and *ERBB2*, we observed that HPV integration sites were often coincident **with loci with genomic amplification. In addition, we observed elevated expression levels of genes in close proximity to HPV integration sites.However, our investigations were limited** by small numbers of whole genome sequencing data for analysis**. In addition, comprehensive** analyses are yet to be performed across different tumor types.
In this project, wewill test the hypothesis that similar investigations using whole genome data amassed from the combined dataset of TCGA and ICGC projects will lead to new insights into how human and pathogenic genomes and transcriptomes synergize to cause cancer. Therefore we will (i) characterize the spectrum of bacterial and viral sequences in whole genome sequencing data across the TCGA/ICGC dataset, and (ii) determine the relationships between the pathogen abundance/integration, and genomic alterations/gene expression signatures across several human cancers.

Timelines & resources dedicated to project

Tentative timeline:

- October 2014: Completion of PathSeq runs for WGS and corresponding RNASeq data for ~2000 tumors
- December 2014: Completion of integrative analysis
- January 2015: Manuscript preparation/submission

Resources required:

- WGS, RNASeq BAMs from tumor and normal (wherever applicable) for PathSeq runs.
- Mutation calls, copy number data, expression data for integrative analysis.
- Large scale computing cluster for PathSeq runs.

Research proposal

1. Identification of pathogen sequences and viral integration sites by computational subtraction
Using PathSeq, input whole genome sequence reads will be subtractively aligned by BWA to human reference sequences, and the remaining reads will be aligned to microbial reference sequences (viral, fungal, bacterial, archaeal). The residual reads will also be*de novo* assembled to facilitate the identification of new or novel microbes in the diseased tissue. Furthermore, chimeric human-pathogen reads or read pairs will serve as landmarks for the identification of viral integration sites.

These analyses will be done in 2 phases, the first focusing on cancers with a strong index of suspicion for the presence of pathogens, e.g. carcinomas of the head and neck, oral cavity, esophagus, stomach, liver, colon/rectum, bladder, cervix, as well as lymphomas. The second phase will focus on all other tumor types.

2. Integrative analyses of viral integration and somatic genomic alterations in human cancers
We will either download available data or liaise with other investigators focusing on the analyses of somatic genomic alterations (e.g. sSNVs, sCNAs), transcriptomic profiles and methylation patterns across the various cancers. Thereafter, integrative analyses of pathogen abundance/integration data with these complimentary datasets will be performed.

Tumors positive for specific pathogens or pathogen families, e.g. HPVs, EBV, etc will be analyzed as a group regardless of the tissue type of origin to identify genomic characteristics that are common to tumors positive for specific pathogens.

We will compare genomic alterations between tumors with integration and those without integration (within the subset of tumors with specific viral infection). We will also compare genomic alterations between tumors with coincident viral integration and human genomic amplification, and tumors with only viral integration.

Furthermore, we will compare viral integration site data obtained from whole genome and transcriptome sequencing data in order to identify the patterns of integration associated with viral gene expression

Legacy plans

1. The PathSeq algorithm used for making microbial sequence calls, is available on LSF and Amazon Cloud.
2. Mutation calls, copy number values and gene expression data will be downloaded from other companion projects.
3. Viral detection and integration data will be made publicly available.

Curriculum vitae for Matthew Meyerson, M.D., Ph.D.

Education and Training

1985	A.B., Chemistry and Physics, Harvard College
1993	M.D., Harvard Medical School
1994	Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996	Resident, Clinical Pathology, Massachusetts General Hospital
1995-1998	Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

1998-2006	Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006	Associate Member, Broad Institute of Harvard and MIT
2005-	Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
2005-2009	Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006-	Senior Associate Member, Broad Institute of Harvard and MIT
2009-	Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

Awards and Honors

1999	Pew Scholar in the Biomedical Sciences
2004	Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005	Clinical Investigator Award, American Lung Association
2009	Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010	Team Science Award, American Association for Cancer Research
2011	Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012	Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobiumenterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …,Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers.Nature. 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. ProcNatlAcadSci U S A. 2001;98(24):13790-13795.

NAME	POSITION TITLE
Akinyemi I. Ojesina	Research Fellow in Medical Oncology Dana-Farber Cancer Institute
eRA COMMONS USER NAME (credential, e.g., agency login) AIOJESINA	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE	MM/YY	FIELD OF STUDY
University of Ibadan, Ibadan, Nigeria	M.B.,B.S.	1998	Medicine and Surgery
Harvard University, Cambridge, MA, USA	Ph.D.	2007	Biological Sciences in Public Health
Dana-Farber Cancer Institute, Boston, MA, USA	Postdoc	2008-14	Cancer Genomics

A. Personal Statement.

I am interested in the genomic analysis of infection-associated malignancies, pathogen discovery, and host-pathogen dynamics in tumor progression. We have developed a computational pathogen discovery pipeline in the Meyerson laboratory. I have also recently concluded the analysis of whole exome sequencing of 115 cervical carcinoma-normal paired samples, RNA sequencing of 79 cases and whole genome sequencing of 14 tumor-normal pairs.

B. Positions and Honors.

Positions and Employment:

1998-1999 House Officer, University College Hospital (UCH), Ibadan, Nigeria
1999-2000 Community Health Physician, National Youth Service Corps, Lagos, Nigeria
2000-2008 Research Fellow in Medical Genetics, College of Medicine, University of Ibadan
2001-2007 Doctoral Candidate, Biological Sciences/Public Health, Harvard University, Cambridge, MA
2008- Research Fellow in Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2008- Postdoctoral Scholar, Broad Institute of MIT and Harvard, Cambridge, MA

Awards and Honors:

1993 Distinction in Biochemistry, University of Ibadan
2001 UICC International Cancer Technology Transfer Fellowship
2001-2006 Gates Foundation-funded AIDS Prevention Initiative in Nigeria (APIN) Fellowship
2006-2007 NIH Fogarty AIDS International Research Fellowship
2008-2011 Rebecca Ridley Kry Fellowship of the Damon Runyon Cancer Research Foundation
2011 AACR-GlaxoSmithKline Outstanding Clinical Scholar Award
2013 AACR-GlaxoSmithKline Outstanding Clinical Scholar Award

C. Selected Relevant Publications

1. Kostic AD, **Ojesina AI**, Pedamallu C, Jung J, Verhaak RGW, Getz G, Meyerson M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature Biotechnology 2011; 29(5):393-396.
2. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, **Ojesina AI**, et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. Genome Research 2012; 22(2):292-298.
3. Lohr JG, Stojanov P, ……. **Ojesina AI**, ……. Shipp MA, Getz G, Golub TR. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci U S A. 2012; 109(10):3879-3884.
4. Bhatt AS, Freeman SS, Herrera AF, Pedamallu CS, ……. **Ojesina AI**,……. Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013; 369(6):517-28.
5. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics 2013; 45(10):1113-20.
6. Francis J, Kiezun A, Ramos AH, ……. **Ojesina AI**, ……. Meyerson M. Somatic mutations of *CDKN1B* in small intestine neuroendocrine tumors. Nature Genetics 2013; 45(12):1483-6.
7. **Ojesina AI***, Lichtenstein L*, et al. Landscape of genomic alterations in cervical carcinomas. Nature 2013; Dec 25. DOI: 10.1038/nature12881. [Epub ahead of print]. *Equal contribution.

D. Previous Research Support

DRG-1998-08 (PI: Ojesina)
07/01/08 – 06/30/11
Damon Runyon Cancer Research Foundation
Pathogen Discovery and Genomic Characterization of Infection-Associated Malignancies

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Chandra Sekhar Pedamallu, Ph.D	**Sr. Computational Biologist, Meyerson Lab, Dana-Farber Cancer Institute, Boston, MA** **Visiting Scientist, The Broad Institute, Cambridge, MA**
CONTACT INFORMATION chandra@broadinstitute.org	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Nagarjuna University, India	B.Tech	1999	Mechanical Engineering
Indian Institute of Technology Madras, India	M.Tech	2001	Industrial Management
Nanyang Technological University, Singapore	Ph.D	2007	Systems Engineering
New England Biolabs, Ipswich, MA	Postdoc	2007-2010	Bioinformatics

A. Personal Statement

My research focus is to find pathogens that may cause human cancers or autoimmune diseases using next generation sequencing. In particular, I am developing cutting edge high-throughput computational methods to identify microbial reads from next generation sequencing data.

B. Positions and Honors

02/2001-08/2001 Software Trainee, Nextset Software Pvt. Ltd, India
08/2001-02/2003 Team Leader and System Analyst, Indonet Global Limited (Subsidiary of SOI Inc, USA), India
02/2003-02/2006 Research scholar, Nanyang Technological University, Singapore
07/2005-12/2005 Guest Ph.D. Student, Institute of Informatics, University of Szeged, Hungary
03/2006-02/2007 Visiting Scientist, New England Biolabs, USA
02/2007-04/2010 Postdoctoral Research Fellow, New England Biolabs, USA
05/2010-present Visiting Scientist,The Broad Institute, Cambridge, MA
05/2010-present Sr. Computational Biologist, Meyerson Lab, Medical Oncology, DFCI, Boston, MA

Honors: • Postdoctoral travel grant support for Sixth International Wolbachia Conference 2010; • Marquis Who's Who in the world (26th Edition); • Postdoctoral research fellowship at New England Biolabs (2007 - 2010); • Nanyang Technological University research scholarship (2003 - 2006); • Indian Institute of Technology Madras half time teaching assistance scholarship (1999 - 2001).

C. Recent relevant publications (From 39 peer-reviewed publications; includes Book chapters)

1. A. D. Kostic, A. I. Ojesina, **C. S. Pedamallu**,….. Meyerson M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. **Nature Biotechnology**, 29, 2011.
2. A. D. Kostic, D. Gevers, **C. S. Pedamallu**,…..,Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. **Genome Research**, 22(2): 292–298, 2012
3. A. S. Bhatt, S. S. Freeman, A. F. Herrera, **C. S. Pedamallu**,…..,Meyerson M. Sequence-based discovery of Bradyrhizobium enterica within cord colitis syndrome. **New England Journal of Medicine**. 2013
4. Cancer Genome Atlas Research Network: **C. S. Pedamllu**, member of analysis group. **Nature Genetics**, 45: 1113–1120, 2013
5. J. M. Francis*, A. Kiezun*, A. H. Ramos*, S. Serra, **C. S. Pedamallu**,…. Meyerson M. Somatic mutation of CDKN1B in small intestine neuroendocrine tumors. **Nature Genetics**, 45: 1483–1486, 2013
6. A. I. Ojesina*, L. Lichtenstein*, S. S. Freeman, **C. S. Pedamallu…**Meyerson M. Landscape of Genomic Alterations in Cervical Carcinomas. **Nature**, 2013.
7. The Cancer Genome Atlas Research Network: **C. S. Pedamallu**, member of the disease working group and analysis working group. Comprehensive molecular characterization of urothelial carcinoma of the bladder. **Nature**, 2013.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrated analysis of copy number and rearrangement

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Matthew Meyerson, DFCI / Broad Institute
Gad Getz, MGH / Broad Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Marcin Imielinski, Broad Institute / MGH / DFCI, Jeremiah Wala, Broad Institute / HMS / DFCI	

Background and preliminary data

Rearrangement and copy number states of cancer genome are usually inferred separately; however, they represent different aspects of the same tumor genome structure. The tumor copy number state is usually inferred through statistical segmentation of a local abundance signal (e.g. tumor / normal read depth ratio), yielding intervals of constant copy number. Downstream processing of this segmented signal is used to identify integer copy states and find genomic regions that statistically enriched in gains and losses. Rearrangements are normally identified through detection of paired-end and clipped-read signatures of DNA breaks and fine-mapped through re-alignment and local-assembly. Applied to WGS data, this fractured analysis approach often yields mutually inconsistent copy number and rearrangement reconstructions, over-segmented copy number profiles, and rearrangement adjacencies that are unassigned to a copy state.

We have developed an integrated structural variant analysis pipeline that combines these two data modalities to infer integer copy states on both genomic segments and their aberrant and reference connections. This pipeline employs both germline and somatic rearrangement data to achieve a consistent reconstruction of both germline cell and ancestral tumor clone as an annotated interval graph. Downstream analysis of this interval graph can be used to nominate sub-clonal somatic alterations (structural variants and point mutations), complex amplicons, multi-gene fusions, and neo-telomeres. Our preliminary analysis of TCGA WGS data shows that this inference is robust to aneuploidy, stromal admixture, hyper-segmentation, and missing rearrangement data.

Timelines & resources dedicated to project

Timeline: Initial interval-graph annotation of 2000 genomes by Sept 2014; Recurrence, germline vs somatic and neo-telomere analyses by December 2014; FISH and PCR validation of sub-clonal alterations, complex amplicons, and neo-telomeres by March 2015; Manuscript preparation / submission, May 2015

Resources: tumor and normal whole genome read data, copy number segmentation, rearrangement adjacency calls for tumor and normal, RNA-seq data, tissue sections for FISH and single cell multiplex PCR,

Research proposal

The proposed research will leverage the power of the ICGC/TCGA WGS dataset and integrated structural variant analysis to address several key cancer genomics questions:

(1) What genomic loci are recurrently amplified and/or deleted?
Standard analyses of copy number recurrence in human cancer rely on reference-centric segmentations of the human genome into intervals of constant copy number. Using our mixed-integer programming based integrated structural variant analysis pipeline (karyoMIP), we will generalize this recurrence analysis to reflect the altered tumor genome structure induced by both rearrangements and somatic copy number alterations (SCNAs). We predict that integrated analysis will enable nomination of novel recurrently altered gene targets (not previously nominated through array based copy number analysis) through cleaner copy number segmentation, increased sensitivity for focal copy number changes, and the ability to detect high-copy fusions.

(2) What is the interplay of copy number gain and rearrangement?
Our integrated structural variant analysis associates both aberrant edges and intervals with an integer copy state. Aberrant adjacencies present at more than a single copy are consistent with a rearrangement *followed* by a copy gain. Aberrant adjacencies linking intervals at different copy states are consistent with a rearrangement *preceded* by a copy gain. We will analyze the incidence of such edges across the ICGC/TCGA WGS data to determine the frequency of "late" vs "early" rearrangements and determine a background model for rearrangement gain. Additional timing information will be inferred through multiplicity analysis of somatic single nucleotide variants (sSNVs) in amplified intervals. We will validate predictions of rearrangement copy states using multiplex FISH.

(3) What is the landscape of sub-clonal somatic substitutions and copy number alterations?
The abundant supply of sSNVs in whole genome sequence data, particularly those from highly-mutated tumor types, enables the characterization of sub-clonal tumor cell populations. A confounding factor to this analysis is aneuploidy, since intervals harboring copy gains or losses will have altered allele fractions. Accurate assignment of integer copy states to intervals is critical for fully harnessing the clonality information stored in WGS sequences, particularly for tumor types with heavily rearranged and mutated genomes (e.g. lung cancer). We predict that the increased accuracy of copy number inference obtained by our integrated structural variant analysis pipeline will enable improved inference of the sub-clonal structure of whole genome sequenced tumors. In addition to nominating subclonal sSNVs, we will be able to identify sub-clonal SCNA's by analyzing intervals whose abundance significantly deviates from their fitted integer state. Such intervals will be candidates for sub-clonal gains and losses. Analytic predictions will be validated using FISH and single cell PCR.

(4) What alterations and sequences underly "loose ends" in tumor genome reconstructions?
Our integrated structural variant analysis framework nominates "loose-ends" in the tumor genome. Though many of these unmapped rearrangements reflect the limited sensitivity for standard-insert WGS to detect structural variants, others may correspond to the creation of novel chromosome ends or insertion of foreign DNA sequence. We will analyze "loose ends" in our analytic reconstruction for enrichment with repetitive sequences of various classes. This will include analysis of telomeric or sub-telomeric sequences, tandem and interspersed repeats, and other classes of unmappable sequences. We will validate neo-telomere predictions using multiplex FISH on tumor tissue sections.

(5) How does the germline structural variant state predispose to somatic rearrangements?
Somatic rearrangements occur in the setting of a genome already altered by germline rearrangements and copy number changes. Germline rearrangements introduce novel adjacencies and contiguities that may predispose to the formation of certain somatic rearrangements in *cis*. We will apply our integrated structural variant analysis framework (which takes into account both germline and somatic aberrant adjacencies) to identify germline structural alterations that locally increase the risk for somatic rearrangement events.

Legacy plans

We will provide karyoMIP software and annotated interval graphs for both tumor and normal WGS specimens to the community. We will also provide "loose-end" classifications and tables of recurrently deleted and amplified loci.

Curriculum vitae for Matthew Meyerson, M.D., Ph.D.

Education and Training

1985	A.B., Chemistry and Physics, Harvard College
1993	M.D., Harvard Medical School
1994	Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996	Resident, Clinical Pathology, Massachusetts General Hospital
1995-1998	Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

1998-2005	Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006	Associate Member, Broad Institute of Harvard and MIT
2005-	Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
2005-2009	Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006-	Senior Associate Member, Broad Institute of Harvard and MIT
2009-	Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

Awards and Honors

1999	Pew Scholar in the Biomedical Sciences
2004	Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005	Clinical Investigator Award, American Lung Association
2009	Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010	Team Science Award, American Association for Cancer Research
2011	Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012	Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …, Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-13795.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

A. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science

2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Imielinski, Marcin	Research Fellow in Pathology
EMAIL ADDRESS marcin@broadinstitute.org	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Rutgers College	B.S., B.A.	5/00	Computer Science, Biological Sciences
University of Pennsylvania School of Medicine (UPenn)	M.D., Ph.D.	5/08	Medicine, Genomics and Computational Biology
Massachusetts General Hospital (MGH), Harvard Medical School (HMS)	Resident	6/11	Clinical Pathology
Brigham and Women's Hospital (BWH), MGH, HMS	Fellow	6/12	Molecular Genetic Pathology

A. Personal Statement

I am an M.D. with clinical training in molecular genetic pathology and a Ph.D. computational biologist with broad experience in genomics and systems biology. I'm fascinated by the potential of integrated 'omics and big data analytics to transform cancer medicine and reveal fundamental features of tumor biology.

B. Positions and Honors

2000-2008 M.D. / Ph.D. Student, Genomics and Computational Biology Program, UPenn
2007-2010 Research Associate, Center for Applied Genomics, Children's Hospital of Philadelphia
2008-2011 Resident in Pathology, MGH / HMS
2011-2012 Clinical Fellow in Molecular Genetic Pathology, BWH / MGH / HMS
2010-Present Postdoctoral fellow in Dr. Matthew Meyerson lab, DFCI / Broad Institute
2012-Present Research Fellow, Department of Pathology, MGH / HMS

Honors: National Merit Scholar (1995), Presidential Scholar, Rutgers College (1995), Henry Rutgers Scholar, Rutgers College (1999), Best Student Poster Award, 5th International Conference for Systems Biology, Heidelberg, Germany (2004), BioAdvance Fellowship in Bioinformatics (2004), Best Paper Award in Session, 26th American Control Conference, New York, NY (2007), Trainee Research Award, American Society for Human Genetics (2009), Best Abstract, Pathology, MGH Clinical Research Day (2010), Most downloaded article in July 2010 for journal "Chaos" (2010), AACR Scholar-in-Training Award (2012), Best Poster in Anatomic Pathology, HMS Pathology Retreat (2013), Top 5 Abstract, Dana-Farber / Harvard Cancer Center Lung Cancer Research Symposium (2013)

C. Selected Peer-reviewed Publications (Selected from 36 peer-reviewed publications)

1. Imielinski, M. et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. *J Clin Invest* (2013), in press.
2. Berger, A.H. et al. Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* (2013), in press.
3. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150, 1107–1120 (2012).
4. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* 150, 251–263 (2012).
5. Imielinski, M. et al. Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol Cell Proteomics* 11, M111.014910 (2012)
6. Imielinski, M. & Belta, C. Deep epistasis in human metabolism. *Chaos* 20, 026104 (2010).
7. Imielinski, M. et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* 41, 1335–1340 (2009).

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Wala, Jeremiah	Graduate Student, Harvard University
eRA COMMONS USER NAME (credential, e.g., agency login)	
jeremiahwala	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Cornell University	B.S.	2009	Engineering Physics
Cornell University	M.S.	2010	Applied Physics
Harvard University	M.D.	2018 (exp)	
Harvard University	Ph.D.	2016 (exp)	Cancer Genomics

Research Experience

2006-2009 Electrospinning and microfluidics, Cornell University, Ithaca, NY
2009-2010 Medical computer vision (thoracic CT), Cornell University, Ithaca, NY
2011-2012 Radiotherapy treatment planning optimization, Massachusetts General Hospital, Boston, MA
2012- Cancer genomics, Broad Institute, Cambridge, MA

Honors

2005 Cornell Tradition Fellowship, Cornell University
2009 *Magna cum laude*, Cornell University
2009 Hartman Prize in Experimental Physics, Cornell University
2010 Cuykendall Memorial Teaching Award, Cornell University

Peer-reviewed Publications

1. Craft D, McQuaid D, **Wala J**, Chen W, Salari T, Bortfeld T. Multicriteria VMAT optimization. Medical Physics. 2012; 57(17), 686-696. PMID: 22320778
2. Salari E, **Wala J**, Craft D. Exploring trade-offs between VMAT dose quality and delivery efficieny using a network optimization approach. Physics in Medicine and Biology. 2012; 57(17), 5587-5600. PMID:
3. **Wala J**, E Salari, W Chen, D Craft. Optimal partial-arcs in VMAT treatment planning. 2012. Physics in Medicine and Biology. 2012; 57:5861-5874
4. **Wala J**, D Craft, J Paly, A Zietman, J Efstathiou. Maximizing dosimetric benefits of IMRT in the treatment of localized prostate cancer through multicriteria optimization planning. 2013. Medical Dosimetry. 2013; 38(3):298-303.
5. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, **Wala J**, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013; 45:1134-40. NIHMSID: 517488.
6. Berger A, Imielinski M, Duke F, **Wala J**, Kaplan N, Shi G, Andres D, Meyerson M. Oncogenic RIT1 mutations in lung adenocarcinoma. 2014. Oncogene. In Press

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Mining the epigenomic consequences of non-coding structural genomic alterations	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Matthew Meyerson, DFCI / Broad Institute Gad Getz, MGH / Broad Institute	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Marcin Imielinski, Broad Institute / MGH / DFCI, Cheng-Zhong Zhang, Broad Institute / DFCI	

Background and preliminary data

DNA rearrangements bring previously distant genomic loci into proximity, potentially impacting local histone modifications, DNA methylation, 3D genomic structure, and gene expression. Epigenomic signals (e.g. CHIP-seq, Hi-C, Chia-PET, methylation sequencing) are typically interpreted in reference coordinates, which ignores the tangled structure of tumor genomes. Re-interpretation of epigenomic signals through the lens of the novel adjacencies and contiguities induced by rearrangements may shed light into functional roles of non-coding rearrangements. For example, the apposition of a highly active enhancer or promoter with a cancer gene may induce over-expression, as is the case with IGH-MYC translocations in Burkitt's lymphoma. More generally, complex rearrangements may bring together several activating (or suppressive) epigenetic marks that induce (or suppress) the expression of downstream "driver" oncogenes (or TSGs). Such driver mechanisms may be particularly sensitive to therapy that modulates the activity of histone-modifying enzymes.

Uncovering the epigenomic consequences of rearrangements in tumor specimens will require combining integrated structural variant analysis with the growing universe of epigenomic tracks, both those obtained from the tumor of interest (e.g. transcriptome sequencing, methylation arrays) and reference cell lines / tissue types (e.g. ENCODE, NIH Roadmap). Part of the challenge includes researching novel paradigms for visualization of epigenomic data in rearranged cancer genomes, i.e. departing from the "reference-centric" and linear visualizations used in standard genome browsers (e.g. UCSC, IGV). Our preliminary work has been to build a curated resource of relevant epigenomic datasets and R based visualization toolkit to facilitate inspection of epigenomic signal patterns in complex loci from TCGA WGSed tumors. In the course of this project, we plan to extend this analytic and visualization framework to connect the dots between genomic alterations, histone modifications, methylation patterns, and gene expression, and nominate "driver" loci.

Timelines & resources dedicated to project

Timeline: Curated reference epigenomic track resource by May 2014, interval-graph annotation of 2000 genomes by Sept 2014, joint analysis with epigenomic tracks, methylation, and expression, "driver" locus nomination by December 2014, validation of predicted epigenomic changes in select tumors (via ChIP-Seq, 4C / Hi-C / ChiaPET) by March 2015, Manuscript preparation / submission, May 2015

Resources: tumor and normal whole genome read data, copy number segmentation, rearrangement adjacency calls for tumor and normal, published epigenomic reference datasets, tumor tissue for downstream epigenomic profiling experiments.

Research proposal

Tumor genome sequences are tangled networks of genomic intervals connected through rearrangement-induced adjacencies. Epigenomic data consist of numeric signals (e.g. TF binding intensity) and intervals (e.g. functional elements, chromatin-state annotations on the reference genome, or pairs of such signals / intervals (e.g. 4C, 5C, Hi-C, Chia-PET). By crossing the output of our integrated structural variant analysis framework (karyoMIP) with reference epigenomic annotations and sample-specific epigenomic profiles (DNA methylation, RNA-seq), we propose to identify novel driver loci targeted by rearrangements. This analysis will comprise several components:

(1) Are novel contiguities of activating (or repressive) reference epigenetic marks induced by cancer rearrangements? Using epigenomic profiles from an appropriate primary tissue or immortalized cell line dataset, we will identify walks on the annotated tumor interval-graph that correspond to novel contiguities of activating (or repressive) chromatin marks and target genes. The simplest such "walk" would involve two-intervals, e.g. resulting from a promoter-gene fusion. This analysis will rest on several assumptions: (a) that the reference sample is a relevant model for the "baseline" epigenomic state of the sequenced tumor (b) that the genome of the reference sample is minimally rearranged and (c) that reference epigenomic signals are unchanged in the tumor other than with respect to their relative location. Though some of these assumptions may not in fact be true (tumor epigenomes can be altered through non-rearrangement based mechanisms), we predict that the size of the ICGC/TCGA dataset will allow us to detect signal through the noise. To maximize power, we will formulate a statistical background model for the formation of such walks and nominate "drivers" as outliers from this model.

(2) Do any novel contiguities result in altered local gene expression and DNA methylation patterns? Since many of the ICGC/TCGA specimens contain transcriptome and DNA methylation data, we will be able to inspect the profiles associated with walks nominated in step (1) for concordant changes. We predict that rearrangement induced contiguities will induce local consistency in methylation profiles in the associated intervals. Walks nominated to be "drivers" will be predicted to be associated with profound methylation changes and altered expression of a target gene. To increase the specificity of the transcriptomic analysis we will employ enhancer / promoter gene relationships using proximity-ligation data (e.g. Hi-C, Chia-PET) to filter candidates. Since these relationships will themselves be altered in tumor genomes, we will consider several approaches for extrapolating such results.

(3) Do nominated loci harbor altered chromatin marks and 3D epigenetic structure in the tumor? To validate candidate loci obtained from analysis of reference epigenomic annotations, we will epigenomically profile the corresponding tumors determine whether rearrangements induced silencing or activation of a target gene(s). This will include ChIP-Seq for accessibility (DNAse-seq) and key histone marks (H3K9me3, H3K27ac, H3K27me3) and proximity ligation approaches (eg 5C, Chia-PET, Hi-C) on tumor and matched normal to validate whether rearrangements resulted in the apposition of novel loci. We will also assay tumor and normal samples for expression of the target gene(s).

In addition to probing the epigenomic consequences of cancer rearrangements, this work will also provide novel analytic and visualization tools, including modeling the background rate of novel epigenetic mark contiguities and plotting epigenomic profiles of highly rearranged loci.

Legacy plans

We will maintain and curate a resource of relevant reference epigenomic annotations and their mapping to tumor samples. We will publish tables annotating candidate and validated loci across the patient set. We will provide novel statistical modeling and visualization tools to the community.

Curriculum vitae for Matthew Meyerson, M.D., Ph.D.

Education and Training

1985	A.B., Chemistry and Physics, Harvard College
1993	M.D., Harvard Medical School
1994	Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996	Resident, Clinical Pathology, Massachusetts General Hospital
1995-1998	Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

1998-2005	Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006	Associate Member, Broad Institute of Harvard and MIT
2005-	Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
2005-2009	Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006-	Senior Associate Member, Broad Institute of Harvard and MIT
2009-	Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

Awards and Honors

1999	Pew Scholar in the Biomedical Sciences
2004	Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005	Clinical Investigator Award, American Lung Association
2009	Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010	Team Science Award, American Association for Cancer Research
2011	Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012	Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …, Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-13795.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

A. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science

2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Imielinski, Marcin	Research Fellow in Pathology
EMAIL ADDRESS marcin@broadinstitute.org	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Rutgers College	B.S., B.A.	5/00	Computer Science, Biological Sciences
University of Pennsylvania School of Medicine (UPenn)	M.D., Ph.D.	5/08	Medicine, Genomics and Computational Biology
Massachusetts General Hospital (MGH), Harvard Medical School (HMS)	Resident	6/11	Clinical Pathology
Brigham and Women's Hospital (BWH), MGH, HMS	Fellow	6/12	Molecular Genetic Pathology

A. Personal Statement

I am an M.D. with clinical training in molecular genetic pathology and a Ph.D. computational biologist with broad experience in genomics and systems biology. I'm fascinated by the potential of integrated 'omics and big data analytics to transform cancer medicine and reveal fundamental features of tumor biology.

B. Positions and Honors

2000-2008 M.D. / Ph.D. Student, Genomics and Computational Biology Program, UPenn
2007-2010 Research Associate, Center for Applied Genomics, Children's Hospital of Philadelphia
2008-2011 Resident in Pathology, MGH / HMS
2011-2012 Clinical Fellow in Molecular Genetic Pathology, BWH / MGH / HMS
2010-Present Postdoctoral fellow in Dr. Matthew Meyerson lab, DFCI / Broad Institute
2012-Present Research Fellow, Department of Pathology, MGH / HMS

Honors: National Merit Scholar (1995), Presidential Scholar, Rutgers College (1995), Henry Rutgers Scholar, Rutgers College (1999), Best Student Poster Award, 5th International Conference for Systems Biology, Heidelberg, Germany (2004), BioAdvance Fellowship in Bioinformatics (2004), Best Paper Award in Session, 26th American Control Conference, New York, NY (2007), Trainee Research Award, American Society for Human Genetics (2009), Best Abstract, Pathology, MGH Clinical Research Day (2010), Most downloaded article in July 2010 for journal "Chaos" (2010), AACR Scholar-in-Training Award (2012), Best Poster in Anatomic Pathology, HMS Pathology Retreat (2013), Top 5 Abstract, Dana-Farber / Harvard Cancer Center Lung Cancer Research Symposium (2013)

C. Selected Peer-reviewed Publications (Selected from 36 peer-reviewed publications)

1. Imielinski, M. et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. *J Clin Invest* (2013), in press.
2. Berger, A.H. et al. Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* (2013), in press.
3. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150, 1107–1120 (2012).
4. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* 150, 251–263 (2012).
5. Imielinski, M. et al. Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol Cell Proteomics* 11, M111.014910 (2012)
6. Imielinski, M. & Belta, C. Deep epistasis in human metabolism. *Chaos* 20, 026104 (2010).
7. Imielinski, M. et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* 41, 1335–1340 (2009).

CHENG-ZHONG ZHANG

PERSONAL PROFILE

· chengz@broadinstitute.org
· 7 Cambridge Center Cambridge MA 02142
· +1-617-714-8681

EDUCATION

Caltech 2001-2007
Ph.D., Chemical Engineering, minor in Physics

Tsinghua University (Beijing) 1997-2001
B. Eng., Chemical Engineering

RESEARCH INTERESTS

- Single-Cell sequencing technologies and analysis
- Characterization of cancer genomes and tumor heterogeneity
- Chromosomal rearrangements and aneuploidy in cancer; biophysics of genome integrity
- Bioinformatic analysis of whole-genome sequencing data

CURRENT AND PAST RESEARCH

05/2011-present *Computational Biologist*, Broad institute of Harvard and MIT
Supervisor: Matthew L. Meyerson, M.D., Ph.D.

Single-cell genomic analysis

- Characterization of whole-genome amplification artifacts and optimization of single-cell genomic analysis
- Population-based analysis of single tumor cell genomes and reconstruction of tumor evolution history

Chromosomal translocations and aneuploidy

- Integrative analysis of DNA copy-number alterations and chromosomal translocations from whole-genome sequencing
- Correlation analysis of genomic rearrangements and statistical inference of tumor evolution history
- Characterization of DNA damages due to abnormal mitosis

12/2007-05/2011 *Postdoctoral Fellow*, Harvard Medical School, Supervisor: Timothy A. Springer, Ph.D.

Single-molecule biophysics

- Single-molecule studies of receptor-ligand interactions;
- Reconstruction of the free-energy landscape from single-molecule force spectroscopy;

09/2001-09/2007 *Graduate Student*, California Institute of Technology

LIST OF PUBLICATIONS

**denotes equal contributions*

1. "*Single-nucleus sequencing resolves heterogeneity in EGFR aberrations in glioblastoma.*" Josh Francis*, **Cheng-Zhong Zhang***, Cecile Maire*, ... *Under review* at Cancer Discovery.
2. "*Chromothripsis and beyond: rapid genome evolution from complex genomic rearrangement.*" **Cheng-Zhong Zhang***, Mitchell Leibowitz*, David Pellman. Genes and Development **27**, 2513-2530, 2013.
3. "*Pan-cancer patterns of somatic copy number alteration.*" Travis I. Zack, Steven E. Schumacher, ..., **Cheng-Zhong Zhang**, ... Nature Genetics **45**, 1134-1140, 2013.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Motifs and models of large-scale rearrangements in cancer	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Matthew Meyerson, DFCI / Broad Institute Gad Getz, MGH / Broad Institute	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Marcin Imielinski, Broad Institute / MGH / DFCI, Cheng-Zhong Zhang, Broad Institute / DFCI	

Background and preliminary data

In many tumor types, rearrangements and copy number changes cooperate to create a complex landscape of structural variation. Large-scale rearrangement events, such as chromothripsis, chromoanasynthesis, chromoplexy, and breakage-fusion-bridge cycles, comprise multiple DNA breakages that occur simultaneously or in close temporal succession in cancer genome evolution. Such events have been identified across multiple cancers, including medulloblastoma and prostate adenocarcinoma. The analysis of complex rearrangements is most difficult in highly altered genomes that have been subjected to several rounds of rearrangements and copy change at a single reference locus. Previous approaches to identifying large-scale structural rearrangements have relied on isolated analysis of copy number or rearrangement data.

Our preliminary WGS analyses show that signatures of large-scale rearrangement are most readily extracted through joint inference of rearrangement and copy state. We perform this analysis on tumor and matched normal WGS data by annotating a graph of intervals with copy states on both vertices and edges. Complex rearrangement events manifest as distinct graph motifs, which comprise a combination of connectivity and copy state features. Through analysis of a large pan-cancer cohort of WGS data, we aim to identify motifs of these and possibly novel complex events and characterize their genomic and clinical context. This analysis can be used to model the evolution of such complex loci and examine their possible “driver” consequences. Analytic predictions can be tested using multiplex FISH on tumor tissue sections and through genome engineering experiments.

Timelines & resources dedicated to project

Timeline: Initial interval-graph annotation of 2000 genomes by Sept 2014, characterization of motifs and context analysis by December 2014, FISH validation by Feb 2015, Manuscript preparation / submission, May 2015

Resources: tumor and normal whole genome read data, copy number segmentation, rearrangement adjacency calls for tumor and normal, tumor tissue sections for FISH

Research proposal

We will apply integrated structural variant analysis to annotate integer copy state, purity, and ploidy for 2000+ ICGC/TCGA whole genome sequenced tumor-normal pairs. This analysis will use our mixed-integer programming inference framework (karyoMIP) to integrate data from aberrant adjacencies with segmented DNA abundance data to infer copy state on genomic intervals and adjacencies. We then mine the resulting annotated interval graphs to group edges into complex rearrangement events and characterize their landscape across different tumor types.

Our analysis of TCGA WGS data has yielded preliminary graph-based signatures for chromoplexy, chromothripsis, and breakage-fusion-bridge (BFB) cycles. We plan to extend this analysis to the ICGC/TCGA pan-cancer dataset for *de novo* identification of motifs, which we will map to known and novel models of large-scale rearrangements. Analysis of these motifs will be used to build a taxonomy of complex rearrangement events and characterize their distribution across the ICGC/TCGA WGS data set. Analysis of the genomic and clinical context of event instances across the dataset will shed light onto the etiogenesis and biological mechanisms driving large-scale structural genomic changes. We will employ these rearrangement predictions in evolutionary simulations where we model temporal sequences of simple structural events (deletions, tandem duplications, balanced translocation), chromosomal copy changes, and large-scale rearrangements. These simulations will provide insight into the relative timing of large-scale events, and help untangle the history of loci shaped by series of multiple events (e.g. BFBs following a simple reciprocal rearrangement). We will validate histories and resulting reconstructions for select loci using multiplex FISH on tumor tissue sections. We will attempt to engineer select events / loci in immortalized primary cells using CRISPR/Cas9 systems to examine driver phenomena and validate evolutionary models.

This analysis will shed light on the landscape of large-scale rearrangement events across cancer through an integrated analysis of copy number and rearrangement. The size of this cohort will provide the power to characterize patterns of large-scale structural variation, both with respect to the background rearrangement process and positively-selected “driver” alterations. This integrated analysis will produce a taxonomy of complex events and probe both causes and consequences of large-scale structural variation.

Legacy plans

We will provide karyoMIP software, annotated interval graphs, motif definitions, and per-sample complex rearrangement annotations resulting from our analysis to the community.

Curriculum vitae for Matthew Meyerson, M.D., Ph.D.

Education and Training

1985	A.B., Chemistry and Physics, Harvard College
1993	M.D., Harvard Medical School
1994	Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996	Resident, Clinical Pathology, Massachusetts General Hospital
1995-1998	Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

1998-2005	Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006	Associate Member, Broad Institute of Harvard and MIT
2005-	Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
2005-2009	Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006-	Senior Associate Member, Broad Institute of Harvard and MIT
2009-	Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

Awards and Honors

1999	Pew Scholar in the Biomedical Sciences
2004	Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005	Clinical Investigator Award, American Lung Association
2009	Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010	Team Science Award, American Association for Cancer Research
2011	Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012	Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …, Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-13795.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

A. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science

2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Imielinski, Marcin	Research Fellow in Pathology
EMAIL ADDRESS marcin@broadinstitute.org	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Rutgers College	B.S., B.A.	5/00	Computer Science, Biological Sciences
University of Pennsylvania School of Medicine (UPenn)	M.D., Ph.D.	5/08	Medicine, Genomics and Computational Biology
Massachusetts General Hospital (MGH), Harvard Medical School (HMS)	Resident	6/11	Clinical Pathology
Brigham and Women's Hospital (BWH), MGH, HMS	Fellow	6/12	Molecular Genetic Pathology

A. Personal Statement

I am an M.D. with clinical training in molecular genetic pathology and a Ph.D. computational biologist with broad experience in genomics and systems biology. I'm fascinated by the potential of integrated 'omics and bi data analytics to transform cancer medicine and reveal fundamental features of tumor biology.

B. Positions and Honors

2000-2008 M.D. / Ph.D. Student, Genomics and Computational Biology Program, UPenn
2007-2010 Research Associate, Center for Applied Genomics, Children's Hospital of Philadelphia
2008-2011 Resident in Pathology, MGH / HMS
2011-2012 Clinical Fellow in Molecular Genetic Pathology, BWH / MGH / HMS
2010-Present Postdoctoral fellow in Dr. Matthew Meyerson lab, DFCI / Broad Institute
2012-Present Research Fellow, Department of Pathology, MGH / HMS

Honors: National Merit Scholar (1995), Presidential Scholar, Rutgers College (1995), Henry Rutgers Scholar, Rutgers College (1999), Best Student Poster Award, 5th International Conference for Systems Biology, Heidelberg, Germany (2004), BioAdvance Fellowship in Bioinformatics (2004), Best Paper Award in Session, 26th American Control Conference, New York, NY (2007), Trainee Research Award, American Society for Human Genetics (2009), Best Abstract, Pathology, MGH Clinical Research Day (2010), Most downloaded article in July 2010 for journal "Chaos" (2010), AACR Scholar-in-Training Award (2012), Best Poster in Anatomic Pathology, HMS Pathology Retreat (2013), Top 5 Abstract, Dana-Farber / Harvard Cancer Center Lung Cancer Research Symposium (2013)

C. Selected Peer-reviewed Publications (Selected from 36 peer-reviewed publications)

1. Imielinski, M. et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. *J Clin Invest* (2013), in press.
2. Berger, A.H. et al. Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* (2013), in press.
3. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cel* 150, 1107–1120 (2012).
4. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* 150, 251–263 (2012).
5. Imielinski, M. et al. Integrated proteomic, transcriptomic, and biological network analysis of breast carcinom reveals molecular features of tumorigenesis and clinical relapse. *Mol Cell Proteomics* 11, M111.014910 (2012
6. Imielinski, M. & Belta, C. Deep epistasis in human metabolism. *Chaos* 20, 026104 (2010).
7. Imielinski, M. et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* 41, 1335–1340 (2009).

CHENG-ZHONG ZHANG

PERSONAL PROFILE

· chengz@broadinstitute.org
· 7 Cambridge Center Cambridge MA 02142
· +1-617-714-8681

EDUCATION

Caltech 2001-2007
Ph.D., Chemical Engineering, minor in Physics

Tsinghua University (Beijing) 1997-2001
B. Eng., Chemical Engineering

RESEARCH INTERESTS

- Single-Cell sequencing technologies and analysis
- Characterization of cancer genomes and tumor heterogeneity
- Chromosomal rearrangements and aneuploidy in cancer; biophysics of genome integrity
- Bioinformatic analysis of whole-genome sequencing data

CURRENT AND PAST RESEARCH

05/2011-present *Computational Biologist*, Broad institute of Harvard and MIT
Supervisor: Matthew L. Meyerson, M.D., Ph.D.

Single-cell genomic analysis

- Characterization of whole-genome amplification artifacts and optimization of single-cell genomic analysis
- Population-based analysis of single tumor cell genomes and reconstruction of tumor evolution history

Chromosomal translocations and aneuploidy

- Integrative analysis of DNA copy-number alterations and chromosomal translocations from whole-genome sequencing
- Correlation analysis of genomic rearrangements and statistical inference of tumor evolution history
- Characterization of DNA damages due to abnormal mitosis

12/2007-05/2011 *Postdoctoral Fellow*, Harvard Medical School, Supervisor: Timothy A. Springer, Ph.D.

Single-molecule biophysics

- Single-molecule studies of receptor-ligand interactions;
- Reconstruction of the free-energy landscape from single-molecule force spectroscopy;

09/2001-09/2007 *Graduate Student*, California Institute of Technology

LIST OF PUBLICATIONS

**denotes equal contributions*

1. "*Single-nucleus sequencing resolves heterogeneity in EGFR aberrations in glioblastoma.*" Josh Francis*, **Cheng-Zhong Zhang***, Cecile Maire*, ... *Under review* at Cancer Discovery.
2. "*Chromothripsis and beyond: rapid genome evolution from complex genomic rearrangement.*" **Cheng-Zhong Zhang***, Mitchell Leibowitz*, David Pellman. Genes and Development **27**, 2513-2530, 2013.
3. "*Pan-cancer patterns of somatic copy number alteration.*" Travis I. Zack, Steven E. Schumacher, ..., **Cheng-Zhong Zhang**, ... Nature Genetics **45**, 1134-1140, 2013.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.	
Title of abstract	
Hunting for de novo centromere/telomere insertions in cancer genomes	
Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators (Name no more than 2; append 1 page CV for each)	
Matthew Meyerson, Gad Getz	
Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Cheng-Zhong Zhang, Marcin Imielinski	
Background and preliminary data	
Telomeres preserve genome integrity and centromeres ensure proper chromosome segregation during normal mitosis. Any genetically stable derivative chromosome must contain one centromere and either two telomeres (linear) or none (circular). However, individual cancer cells frequently harbor acentric or dicentric chromosomes as a result of genetic instability. Such unstable configurations can also arise due to chromosomal break and repair, whose resolution may involve novel centromere or telomere formation. In this project we will look for signatures in cancer genomes that are suggestive of novel centromere or telomere formation. The main aim of this project is to generate preliminary knowledge of such events and how they are involved in DNA repair.	
Timelines & resources dedicated to project	
The input for this analysis is generated from the integrative copy-number/chromosomal rearrangement analysis, which is expected to be complete by September 2014. Downstream analysis should be complete before December of 2014. Complementary validation by cytogenetic methods (FISH or SKY) will be performed in the first months of 2015.	

Research proposal
Both telomeres and centromeres consist of satellite repeats flanked by highly heterogeneous yet homologous pericentromeric and subtelomeric sequences; these features make it difficult to locate such sequences directly from shot-gun sequencing. Here we will combine genomic sequencing with cytogenetic methods to hunt for novel centromere or telomere formation. More specifically, we will rely on genomic sequencing to nominate candidate regions and apply long-fragment sequencing technology, FISH, and SKY for validation. **Nomination comes from two sources of information. First, we will identify chromosomal breaks that are not resolved by chromosomal translocations to loci in the reference assembly. As naked DNA ends cannot sustain through the cell cycle, such "unpaired" chromosomal breaks must have been repaired to genomic sequences not present in the reference assembly, such as subtelomeric, pericentromeric, or ribosomal sequences. We will perform local assembly near such breaks and compare the non-reference clipped sequence with known motifs of subtelomeric/pericentromeric/ribosomal sequences to infer the translocation partner loci of such breaks.** **Second, we will develop algorithms to walk the derivative chromosome combining copy-number and rearrangement events. In-silico chromosome walking can identify derivative chromosomes that lack centromeres or telomeres. This analysis will identify target chromosomes for which we can design probes to perform chromosomal painting.** **After identifying specific derivative chromosomes that showed an “unstable” conformation, we will perform complementary study and validation by multiple technologies, including nested PCR (for target region capture), long-fragment sequencing (3kb fragment jump library), and cytogenetic characterization (FISH or SKY).**
Legacy plans

Curriculum vitae for Matthew Meyerson, M.D., Ph.D.

Education and Training

1985	A.B., Chemistry and Physics, Harvard College
1993	M.D., Harvard Medical School
1994	Ph.D., Biophysics, Harvard University (thesis advisor: Ed Harlow)
1994-1996	Resident, Clinical Pathology, Massachusetts General Hospital
1995-1998	Post-doctoral fellow, Whitehead Institute (mentor: Robert Weinberg)

Research and Professional Experience

1998-2005	Assistant Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2004-2006	Associate Member, Broad Institute of Harvard and MIT
2005-	Director, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute
2005-2009	Associate Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School
2006-	Senior Associate Member, Broad Institute of Harvard and MIT
2009-	Professor of Pathology, Dana-Farber Cancer Institute, Harvard Medical School

Awards and Honors

1999	Pew Scholar in the Biomedical Sciences
2004	Tisch Family Outstanding Achievement Award for Translational Cancer Research
2005	Clinical Investigator Award, American Lung Association
2009	Paul Marks Prize in Cancer Research, Memorial Sloan Kettering Cancer Center
2010	Team Science Award, American Association for Cancer Research
2011	Caine Holter Hope Now Award, Uniting against Lung Cancer Foundation
2012	Ilchun Award in Molecular Medicine, Korean Society of Biochemistry & Molecular Biology

Publications (10 selected of 189 peer-reviewed original research publications)

1. Bhatt AS, …, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013 Aug 8;369(6):517-28.

2. Imielinski M, …, Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing Cell. 2012 Sep 14;150(6):1107-20.

3. The Cancer Genome Atlas Research Network. (M. Meyerson, corresponding author). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012 Sep 27;489(7417):519-25.

4. Kostic AD, …, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma Genome Res. 2012 Feb;22(2):292-8.

5. Beroukhim R, …, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature 2010 Feb 18;463(7283):899-905.

6. Bass AJ, …, Meyerson M. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cel carcinomas. Nat Genet. 2009 Nov;41(11):1238-1242.

7. The Cancer Genome Atlas Research Network. (L. Chin and M. Meyerson, corresponding authors). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oc 23;455(7216):1061-1068.

8. Weir BA, …, Meyerson M. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007;450(7171):893-898.

9. Paez JG, …, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy Science. 2004;304(5676):1497-1500.

10. Bhattacharjee A, …, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-13795.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Gad Getz	Director of Bioinformatics, Massachusetts General Hospital Cancer Center and Dept. of Pathology Director of Cancer Genome Computational Analysis, Broad Institute Associate Professor of Pathology, Harvard Medical School
eRA COMMONS USER NAME (credential, e.g., agency login) GADGETZ	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Hebrew University, Israel	B.Sc.	1992	Physics and Mathematics
Tel-Aviv University	M.Sc.	1998	Physics
Weizmann Institute of Science, Israel	Ph.D.	2003	Physics

A. Personal Statement

My research is focused on cancer genome analysis which includes identifying somatic events that cause cancer or germline events that increase risk for getting cancer, as well as identifying subtypes of the disease and their relationship to clinical parameters and/or treatment outcome. My background and expertise are in computational biology bringing rigorous statistical methods to the analysis of genomic data. In particular, I am interested in developing statistical tools to distinguish 'driver' from 'passenger' alterations in the cancer genome and by that identifying novel candidate genes, pathways and non-coding regions that promote tumorigenesis. In addition, I am working on questions regarding experimental design of cancer genome projects and estimating the power to detect cancer-related events. My group is also focused in developing tools to detect somatic events from massively parallel sequencing data including point mutations, insertions and deletions, copy-number changes and rearrangements. We are building these tools in a robust analytical pipeline to analyze data coming from various cancer genome projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). I am a co-PI on a major TCGA genome data analysis center (GDAC) that automatically analyzes genomic data from the entire TCGA and regularly provides data snapshots and results to the research community.

B. Positions and Honors

Positions:

1992-1997 Military Service - Captain
1997-1998 Tel Aviv. Univ. MSc student
1998-2000 Maximal Innovative Intelligence (part time)
1998-2003 Weizmann Institute of Science. PhD student
2004-2007 Broad Institute of MIT and Harvard. Postdoc
2007-2012 Broad Institute of MIT and Harvard. Head of Cancer Genome Analysis
2013- Director of Bioinformatics, MGH Cancer Center and Dept. of Pathology

Honors:

1991 Dean's excellence list. B.Sc. Hebrew University
1995 Prize for Creative Thinking. Israel Defense Forces
1997 Excellence award. M.Sc. Tel-Aviv University
2001 Sir Charles Clore Doctoral Scholarship, Weizmann Institute of Science

2002 Ph.D. Scholarship from the Planning and Budgeting Committee of the Israeli Council for High Education
2002 Student delegate to the International Achievement Summit (Barak Scholarship)
2004 Feinberg Graduate School prize of excellence

C. Selected Peer-reviewed Publications (15 publications)

1. **Getz G***, Hofling H*, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science. 2007 Sep 14;317(5844):1500.PMID: 17872428
2. Beroukhim R*, **Getz G***, …, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. PNAS. 2007 Dec 11; 104(50): 20007-20012. PMID: 18077431, PMCID: PMC2148413
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. Lead author of copy number and sequencing parts. PMID: 18772890, PMCID: PMC2671642
4. Ding L*, **Getz G***, Wheeler DA*, …, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23; 455(7216):1069-75. PMID: 18948947, PMCID: PMC2694412
5. Beroukhim R, Mermel CH, …, Lander ES*, **Getz G***, Sellers WR*, Meyerson M*. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905. PMID: 20164920, PMCID: PMC2826709
6. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, **Getz G**, Meyerson M, Beroukhim R. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Sep 26;45(10):1134-1140. PMID: 24071852, NIHMS ID: 517488, PMCID - In Process
7. Chin L, Hahn WC, **Getz G**, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55. PMID: 21406553, PMCID: PMC3059829
8. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, …, Hahn WC, Garraway LA, Meyerson M, Lander ES, **Getz G***, Golub TR*. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72. PMID: 21430775, PMCID: PMC3560292
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R*, **Getz G***. GISTIC2.0 facilities sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28; 12(4):R41. PMID: 21527027, PMCID: PMC3218867
10. Wang L, Lawrence MS, Wan Y, Stojanov P, …, Neuberg D, Brown JR, **Getz G***, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. NEJM. 2011 Dec; 365:2497-2506. PMID: 22150006, PMCID: PMC3685413
11. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, **Getz G**. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.2012 Dec; PMID: 23124520, PMCID: PMC3561864
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, **Getz G**. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Feb 10. PMID: 23396013, PMCID: PMC3833702
13. Landau DA, Carter SL, Stojanov P, …, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, **Getz G***, Wu CJ*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. PMID: 23415222, PMCID: PMC3575604
14. Dulak AM, Stojanov P, Peng S, Lawrence MS, …, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, **Getz G***, Bass AJ*. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature Genetics. 2013 March 24; 45(5):478-486 PMID: 23525077, PMCID: PMC3678719
15. Lawrence MS, Stojanov P, Polak P, …, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES*, **Getz G***. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 June 11; 499:214-218. PMID: 23770567, NIHMS ID:471461, PMCID - In Process

CHENG-ZHONG ZHANG

PERSONAL PROFILE

· chengz@broadinstitute.org
· 7 Cambridge Center Cambridge MA 02142
· +1-617-714-8681

EDUCATION

Caltech 2001-2007
Ph.D., Chemical Engineering, minor in Physics

Tsinghua University (Beijing) 1997-2001
B. Eng., Chemical Engineering

RESEARCH INTERESTS

- Single-Cell sequencing technologies and analysis
- Characterization of cancer genomes and tumor heterogeneity
- Chromosomal rearrangements and aneuploidy in cancer; biophysics of genome integrity
- Bioinformatic analysis of whole-genome sequencing data

CURRENT AND PAST RESEARCH

05/2011-present *Computational Biologist*, Broad institute of Harvard and MIT
Supervisor: Matthew L. Meyerson, M.D., Ph.D.

Single-cell genomic analysis

- Characterization of whole-genome amplification artifacts and optimization of single-cell genomic analysis
- Population-based analysis of single tumor cell genomes and reconstruction of tumor evolution history

Chromosomal translocations and aneuploidy

- Integrative analysis of DNA copy-number alterations and chromosomal translocations from whole-genome sequencing
- Correlation analysis of genomic rearrangements and statistical inference of tumor evolution history
- Characterization of DNA damages due to abnormal mitosis

12/2007-05/2011 *Postdoctoral Fellow*, Harvard Medical School, Supervisor: Timothy A. Springer, Ph.D.

Single-molecule biophysics

- Single-molecule studies of receptor-ligand interactions;
- Reconstruction of the free-energy landscape from single-molecule force spectroscopy;

09/2001-09/2007 *Graduate Student*, California Institute of Technology

LIST OF PUBLICATIONS

**denotes equal contributions*

1. "*Single-nucleus sequencing resolves heterogeneity in EGFR aberrations in glioblastoma.*" Josh Francis*, **Cheng-Zhong Zhang***, Cecile Maire*, ... *Under review* at Cancer Discovery.
2. "*Chromothripsis and beyond: rapid genome evolution from complex genomic rearrangement.*" **Cheng-Zhong Zhang***, Mitchell Leibowitz*, David Pellman. Genes and Development **27**, 2513-2530, 2013.
3. "*Pan-cancer patterns of somatic copy number alteration.*" Travis I. Zack, Steven E. Schumacher, ..., **Cheng-Zhong Zhang**, ... Nature Genetics **45**, 1134-1140, 2013.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Imielinski, Marcin	Research Fellow in Pathology
EMAIL ADDRESS marcin@broadinstitute.org	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Rutgers College	B.S., B.A.	5/00	Computer Science, Biological Sciences
University of Pennsylvania School of Medicine (UPenn)	M.D., Ph.D.	5/08	Medicine, Genomics and Computational Biology
Massachusetts General Hospital (MGH), Harvard Medical School (HMS)	Resident	6/11	Clinical Pathology
Brigham and Women's Hospital (BWH), MGH, HMS	Fellow	6/12	Molecular Genetic Pathology

A. Personal Statement

I am an M.D. with clinical training in molecular genetic pathology and a Ph.D. computational biologist with broad experience in genomics and systems biology. I'm fascinated by the potential of integrated 'omics and big data analytics to transform cancer medicine and reveal fundamental features of tumor biology.

B. Positions and Honors

2000-2008 M.D. / Ph.D. Student, Genomics and Computational Biology Program, UPenn
2007-2010 Research Associate, Center for Applied Genomics, Children's Hospital of Philadelphia
2008-2011 Resident in Pathology, MGH / HMS
2011-2012 Clinical Fellow in Molecular Genetic Pathology, BWH / MGH / HMS
2010-Present Postdoctoral fellow in Dr. Matthew Meyerson lab, DFCI / Broad Institute
2012-Present Research Fellow, Department of Pathology, MGH / HMS

Honors: National Merit Scholar (1995), Presidential Scholar, Rutgers College (1995), Henry Rutgers Scholar, Rutgers College (1999), Best Student Poster Award, 5th International Conference for Systems Biology, Heidelberg, Germany (2004), BioAdvance Fellowship in Bioinformatics (2004), Best Paper Award in Session, 26th American Control Conference, New York, NY (2007), Trainee Research Award, American Society for Human Genetics (2009), Best Abstract, Pathology, MGH Clinical Research Day (2010), Most downloaded article in July 2010 for journal "Chaos" (2010), AACR Scholar-in-Training Award (2012), Best Poster in Anatomic Pathology, HMS Pathology Retreat (2013), Top 5 Abstract, Dana-Farber / Harvard Cancer Center Lung Cancer Research Symposium (2013)

C. Selected Peer-reviewed Publications (Selected from 36 peer-reviewed publications)

1. Imielinski, M. et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. *J Clin Invest* (2013), in press.
2. Berger, A.H. et al. Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* (2013), in press.
3. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150, 1107–1120 (2012).
4. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* 150, 251–263 (2012).
5. Imielinski, M. et al. Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol Cell Proteomics* 11, M111.014910 (2012)
6. Imielinski, M. & Belta, C. Deep epistasis in human metabolism. *Chaos* 20, 026104 (2010).
7. Imielinski, M. et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* 41, 1335–1340 (2009).

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Landscape of germline cancer predisposing genes across human cancers

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Bo Peng, UT MD Anderson Cancer Center, TCGA GDAC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

Goal: Identify likely germline variants (genes) that predispose to different types of cancers.

Motivation: Whole-genome and whole-exome sequencing have been enthusiastically adopted by genetic epidemiologists to identify germline mutations that predispose to cancers. However, many such studies could not replicate findings from prior linkage and/or genome-wide association studies, and have failed to identify and validate novel genetic variants that predispose to these diseases. Besides statistical challenges in analyzing a large number of rare genetic variants among relatively small number of samples, quality of identified variants, small sample size, heterogeneity of cancers, and inability to compare lists of predisposing variants across cancer types usually leave researchers a large number of rare genetic variants that cannot be either excluded or selected for further validation.

Preliminary Studies: The PI has analyzed whole genome sequencing data for family and population based studies of multiple cancers including lung, triple negative breast cancer, Lynch syndrome, and early onset colon cancers. Only few variants have been successfully identified as potential cancer predisposing variants. Meta analyses and cross-validation were performed but failed to yield useful results.

ICGC/TCGA WGS datasets: Whole-genome sequencing data from ICGC/TCGA provide paired samples and family-data (subset of samples) that can be used to improve the quality of germline variant calling. The larger sample sizes significantly improve the statistical power of association tests, and allow us to compare mutation profiles across cancer types and to study subsets of samples across cancer types.

Timelines & resources dedicated to project

Two lines of research will be conducted roughly in parallel. Firstly, we will develop and refine a germline variant-calling algorithm that makes use of general control (1000 genomes), paired normal and cancer samples and family information to improve the accuracy of existing variant callers. The algorithm will be applied to the ICGC/TCGA WGS data. The development, validation and application of the algorithm will be performed throughout the project period.

In the meantime, lists of identify variants that predispose to different types of cancers will be analyzed within and across cancer types. The analyses will start with variants produced by existing variant calling algorithms, as soon as they are made available to the PI, and be revised and repeated with variants called from the improved algorithm from aim 1. The PI can dedicate up to 25% of his effort on this project.

Research proposal

Aim 1: Develop and apply a germline variant-calling algorithm to identify germline variants for each of the samples. The algorithm will be based on a probabilistic model that makes use of information from general control (1000 genomes data), pedigree (available for some of the samples), and variants (including copy number changes) called from normal and cancer samples. Briefly speaking, variants called from each normal sample will be compared to variants called from related individuals (relatives, if available), unrelated individuals (control, known germline mutations) and matched cancer sample. In the last case, the algorithm will check if a mutation exists also in cancer sample, altered by somatic mutation or copy number alternation. The algorithm will make use of sequencing information in BAM files of normal and cancer samples. It will be relatively fast because it is based on variants called from existing variant calling pipelines and does not need to scan through all aligned reads across the whole genome. It is worth noting that variant calling from cancer samples is generally less certain than calling from germline samples, especially for cancers that are mostly driven by copy numbers, the contribution of cancer samples to germline variant calling will therefore vary greatly.

Aim 2: Identify variants that predispose to different types of cancers. The analyses will generally be based on the assumptions that 1) similar to somatic mutations, there are germline mutations that are specific to particular types of cancers or subtypes of cancers that share similar genetic vulnerabilities, and 2) germline mutations can be shared by distinct cancer types as common trigger mutations that lead to different somatic mutations. In particular, we will filter our lists of variants by removing variants that are shared by the control population and variants that are not predicted to have any functional effect, and removing variants that are unrelated to cancer, for example variants that are common in certain subpopulations (e.g. race-specific). We will then cluster samples by list of germline mutations. Based on observations in similar studies of somatic mutations, It is expected that samples will be grouped by site or subtypes of cancers from different tissues of origin. Association analysis will be performed within such groups of samples, with increased statistical power due to larger sample sizes and reduced impact of heterogeneity. Finally, list of potential cancer predisposing variants will be validated using gene and pathway-based information, validated with existing findings and other studies, and compared across cancer types.

Legacy plans

The variant calling algorithm will be distributed either as a standalone software or a module of *variant tools*, which is a comprehensive toolset for the manipulation, annotation and analysis of genetic variants from whole genome or whole exome sequencing studies. The analyses of variants will be performed by *variant tools* and possibly R, and will be made publicly available on the *variant tools* website.

CURRICULUM VITAE
Bo Peng, Ph.D.

Assistant Professor, Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX

EDUCATION

Degree-Granting Education

Shanghai Jiao Tong University, Shanghai, China, BS, 1996, Applied Mathematics

University of Houston, Houston, TX, MA, 2001, Applied Mathematics

Rice University, Houston, TX, PHD, 2006, Biostatistics

Postgraduate Training

Research Fellowship, Genetic Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, Christopher I. Amos, 5/2006-5/2008

Positions and Honors

2008-2012	Instructor, Department of Epidemiology, Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX
2012-present	Assistant Professor, Department of Genetics, Division of Basic Science Research, The University of Texas MD Anderson Cancer Center, Houston, TX

Honors

2003	R.L. Anderson Student Paper Award, Research Conference on Statistics, SRCOS/ASA
2004-2006	Predoctoral W.M. Keck Fellowship, W.M. Keck Foundation
2006-2009	R25 Postdoctoral Fellowship in Cancer Prevention, National Institutes of Health
2011	Richard C. Devereaux Outstanding Young Investigator Award in Lung Cancer Prevention, Prevent Cancer Fundation

Selected Peer-reviewed Publications (Selected from 21 peer-reviewed publications)

1. **Peng B**, Kimmel M. simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21(18):3686-7, 9/2005. PMID: 16020469.
2. **Peng B**, Amos CI, Kimmel M. Forward-time simulations of human populations with complex diseases. PLoS Genet 3(3):e47, 3/2007. e-Pub 2/2007. PMCID: PMC1829403.
3. **Peng B**, Amos CI. Forward-time simulations of non-random mating populations using simuPOP. Bioinformatics 24(11):1408-9, 6/2008. e-Pub 4/2008. PMCID: PMC2691961.
4. **Peng B**, Li B, Han Y, Amos CI. Power analysis for case-control association studies of samples with known family histories. Hum Genet 127(6):699-704, 6/2010. e-Pub 4/2010. PMID: 20383776.
5. **Peng B**, Amos CI. Forward-time simulation of realistic samples for genome-wide association studies. BMC Bioinformatics 11:442, 2010. e-Pub 9/2010. PMCID: PMC2939614.
6. **Peng B**, Liu X. Simulating Sequences of the Human Genome with Rare Variants. Hum Hered 70(4). e-Pub 1/2011. PMID: 21212684.
7. San Lucas FA, Wang G, Scheet P, **Peng B**. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics 28(3):421-422, 12/2011. PMCID: PMC3268240.
8. **Peng B**, Kimmel M and Amos CI. Forward-time population genetics simulations, methods, implementation and applications. Wiley & Sons, 2011. PMID: 9780470503485.
9. Gorlova OY, Ying J, Amos CI, Spitz MR, **Peng B**, Gorlov IP. Derived SNP alleles are used more frequently than ancestral alleles as risk-associated variants in common human diseases. J Bioinform Comput Biol 10(2):1241008, 4/2012. PMCID: PMC3655427
10. **Peng B**, Chen HS, Mechanic LE, Racine B, Clarke J, Clarke L, Gillanders E, Feuer EJ. Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. Bioinformatics 29(8):1101-2, 4/2013. e-Pub 2/2013. PMCID: PMC3624809.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Pan-Cancer Analysis of Intra-Tumor Heterogeneity and Complex Rearrangements

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Benjamin J. Raphael
Department of Computer Science and Center for Computational Molecular Biology, Brown University, USA
TCGA and ICGC Bioinformatics Analysis Groups

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Iman Hajirasouliha, Brown University Layla Oesper, Brown University	Suzanne Sindi, University of California, Merced

Background and preliminary data

Cancer genomes are highly rearranged and often contain extensive duplications and deletions. In addition, most tumors exhibit extensive intra-tumor heterogeneity, with individual cells in a tumor typically having different complements of somatic mutations. Since most cancer sequencing projects (including ICGC) sequence a tumor sample containing various subpopulations of tumor cells as well as admixture by normal (non-cancerous) cells.

Our group has developed several algorithms to detect and analyze cancer genome rearrangements from whole-genome sequencing data, and to use these rearrangements to analyze intra-tumor heterogeneity. These include:
(1) **GASV** (Sindi, *et al*. 2009) and **GASVPro** (Sindi, *et al.* 2012) detect all types of structural variants from paired-end sequencing data, combining signals from both discordant read pairs and read depth.
(2) **PREGO** (Oesper, *et al.* 2012) reconstructs the block organization of a cancer genome using both discordant read pairs and read depth. PREGO reconstructs the most likely sequence of genomic blocks or segments from the reference genome that "spell out" the cancer genome.
(3) **THetA** (Oesper *et al*., 2013) uses copy number aberrations to infer tumor purity as well as the number and composition of tumor subpopulations from a single tumor sample.
We have used each of these algorithms on whole-genome sequencing data from TCGA and other studies.

In addition to the above, we have also developed approaches to analyze genome rearrangement mechanisms from paired-end sequencing data. Several recent studies report that some rearrangements exhibit a complicated structure with multiple, closely located breakpoints. An extreme example of this phenomenon is *chromothripsis*, proposed by Stephens *et al.* (2011), which posits that a small portion of the genome in a cancer cell undergoes a cataclysmic event resulting in a shattering of genetic material that is subsequently pieced back together in apparently random order. **We have developed a preliminary graph theoretic model of chromothripsis and have applied it to 24 genomes from 5 different studies** (Stephens *et al*., 2011, Rausch *et al*., 2012, Malhotra *et al*., 2013, Nik-Zainal *et al*., 2012, Zakov *et al.,* 2013) where each genome has been previously classified as containing either a chromothripsis event, a breakage/fusion/bridge cycle or is the result of a step-wise sequence of events. We find that we are able to identify chromothripsis events with a true positive rate of 52% and a false positive rate of 0%. A potentially related phenomenon is *chromoplexy*, which involves fewer rearrangements than chromothripsis affecting a larger number of chromosomes (Baca *et al*., 2013). While chromothripsis is proposed to be a one-time event, chromoplexy may occur several times during tumor evolution.

Timelines & resources dedicated to project

Timeline: Initial run of GASV/GASV-Pro and PREGO complete – April 2014. Initial complex rearrangement and THetA analysis complete – July 2014. Model/Algorithm Revisions Complete – October 2014. Chromothripsis Analysis Complete – November 2014. Additional THetA Analysis (Including primary tumor/relapse analysis) Complete – January 2015. Manuscript preparation – January to February 2015. Manuscript submission – 20th March 2015.

Resources: Our pan-cancer analysis of complex rearrangements and intra-tumor heterogeneity is staffed by 1 postdoc, 1 graduate student and 1 staff programmer. We have already installed and tested our software described above on TCGA data using the Bionimbus Protected Data Cloud, and thus are well prepared to compute on the ICGC data. In additional to cloud computing resources supplied by the ICGC pan-cancer consortium, we can utilize two compute clusters. The first cluster at Brown's Center for Computing and Visualization contains >3000 Intel Xeon cores, with each node containing >= 24Gb of memory. The second cluster at Brown's Department of Computer Science contains 1820 cores and includes machines with up to 256 Gb RAM. We also have two 16-core workstations, each with 256Gb RAM for development and testing, and used exclusively by our research group.

Research proposal

Using our existing and new-developed algorithms will perform three complementary analyses.

First, **we will apply our THetA algorithm to all ~2000 WGS tumor/normal pairs.** THetA will determine large-scale patterns of intra-tumor heterogeneity across and within tumor types, including the number of tumor subpopulations or whether all samples exhibit a dominant clonal population. We will investigate how intra-tumor heterogeneity varies across tumor types. Tumor purity estimates output by THetA may also provide useful input for other programs that infer single nucleotide variants (SNVs). In addition, we will also use THetA to analyze the ~200 datasets where multiple samples, such as primary tumor and relapse are available. **This analysis will allow us to determine the relationship between tumor subpopulations in the primary tumor and relapse,** which may yield important information about the progression of different tumors.

Concurrently, analysis of large pan-cancer datasets will motivate algorithmic improvements to THetA. One such improvement is the type of model selection criterion used to determine the number of tumor subpopulations. Currently the Bayesian Information Criterion (BIC) is used, but we will investigate the use of non-parametric methods such as a Dirichlet Process.

Second, **we will identify genomes with complex rearrangements and investigate the prevalence of different rearrangement mechanisms (chromothripsis, chromoplexy, breakage/fusion/bridge cycles, sequential rearrangements) both within and across different cancer types.** We will use our GASV/GASVPro algorithms to identify rearrangements in whole-genome sequences, and use PREGO to combine the individual rearrangement breakpoints into a cancer genome with rearrangements and duplications. We will use different signatures identified by PREGO to identify the underlying rearrangement mechanisms.

Finally, we will use **our graph theoretic model of a chromothripsis to determine the rate of chromothripsis within each cancer type, as well as to investigate the differences between chromothripsis and chromoplexy**. In particular, the inclusion of bone cancer samples will allow us to further investigate the observation by Stephens *et al*. (2011) that chromothripsis is more prevalent in this type of cancer.

Legacy plans

All of the software developed in our research group, including the GASV, GASVPro, PREGO, and THetA programs described above, are freely available for researchers at: http://compbio.cs.brown.edu/software. We are also starting to release software on source-code repositories including Google Code and GitHub. We are in the process of packaging the group's software into virtual machines: one for Amazon EC2 and one for local use.

As part of this project, we anticipate that we will improve the efficiency and ease of usability of our software, and will release updated versions of the software. Additionally, we plan to write a companion methodology paper to further disseminate these developments and their application to large-scale cancer studies.

BENJAMIN J. RAPHAEL
Associate Professor
Department of Computer Science & Center for Computational Molecular Biology
Brown University
Providence, RI 02912
Phone: (401) 863-7643 Email: braphael@brown.edu Web: http://compbio.cs.brown.edu

EDUCATION

1996-2002 **Ph.D.** in Mathematics, University of California, San Diego.

1992-1996 **S.B.** in Mathematics, **S.B. Minor** in Biology, Massachusetts Institute of Technology.

EXPERIENCE

2013-present **Director**, Center for Computational Molecular Biology, Brown University
2011-present **Associate Professor**, Department of Computer Science & Center for Computational Molecular Biology, Brown University
2006-2011 **Assistant Professor**, Department of Computer Science & Center for Computational Molecular Biology, Brown University

SELECTED RECENT PUBLICATIONS

C.Kandoth, [16 additional authors], **B.J. Raphael**, L. Ding. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* 502(7471):333-9

L. Oesper, A. Mahmoody, **B.J. Raphael**. (2013) THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology* 14:R80.

The Cancer Genome Atlas Research Network. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43-9

M.D.M. Leiserson, D. Bloch, R. Sharan*, **B.J. Raphael***. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLOS Computational Biology.* May;9(5):e1003054. *equal contribution

The Cancer Genome Atlas Research Network. (2013) Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*. 368(22):2059-74.

Sindi, S. Onal, L. Peng, H. Wu, **B.J. Raphael**. (2012) An Integrative Probabilistic Model for Identification of Structural Variation in Sequencing Data. *Genome Biology* 13(3):R22.

F. Vandin, E. Upfal, **B.J. Raphael**. (2012) *De novo* Discovery of Mutated Driver Pathways in Cancer. *Genome Research.* 22(2):375-85. [Journal version of paper accepted at *RECOMB 2011*].

Vandin F, Upfal E, **B.J. Raphael** . (2011) Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology.* 18(3):507-22. [Journal version of paper accepted at *RECOMB 2010*].

International Cancer Genome Consortium. (2010) International network of cancer genome projects. *Nature*. 464(7291):993-8.

S. Sindi, E. Helman, A. Bashir, **B.J. Raphael**. (2009) A Geometric Approach for Classification and Comparison of Structural Variants. *Bioinformatics* 25: i222-i230. *[Proceedings of ISMB/ECCB 09*]

Iman Hajirasouliha, Ph.D.

Contact Information

Box 1910, Computer Science Department
Brown University
Providence, Rhode Island, 02906

Office: +1 401 863 6044
Web: http://www.imanh.org
E-mail: imanh@cs.brown.edu

Education

Simon Fraser University, Burnaby, BC, Canada

Ph.D in Computing Science — **Jan 2008 – August 2012**

M.Sc. in Computing Science — **Sep 2005 – Dec 2007**

Sharif University of Technology, Tehran, Iran

B.Sc. in Computer Engineering — **Sep 2001 – Jul 2005**

Honours and Awards

◇ NSERC-CGS Michael Smith Foreign Study Supplements ($6,000), Summer 2012
◇ NSERC Alexander Graham Bell Canada Graduate Scholarship ($70,000), 2010–2012
◇ Simon Fraser University Graduate Fellowship ($6,250), Summer 2012
◇ **Best Paper Award**, HiTSeq 2011: Conference on High Throughput Sequencing Analysis and Algorithms (Special Interest Group of ISMB 2011: July 15-16, 2011 Vienna, Austria)

Current Affiliation

Department of Computer Science, Brown University, Providence, RI, USA

Post Doctoral Research Associate — **Jan 2013 – present**

Member of the Raphael Lab for Computational Biology. Research on Bioinformatics Algorithms, Combinatorial Optimization, Cancer Genomics.

Selected Publications

1000 Genomes Project Consortium.
An integrated map of genetic variation from 1,092 human genomes.
Nature 2012 Nov 1;491(7422):56-65.

Iman Hajirasouliha, Alexander Schönhuth, David Juan, Alfonso Valencia, S. Cenk Sahinalp
Mirroring co-evolving trees in the light of their topologies
Bioinformatics (2012) May 1;28(9):1202-8.

Fereydoun Hormozdiari, Iman Hajirasouliha, Andrew McPherson, Evan E. Eichler, S. Cenk Sahinalp.
Simultaneous structural variation discovery in multiple paired-end sequenced genomes.
Genome Research, 2011 Dec;21(12):2203-12
◇ Featured on the cover of the journal!
◇ Highlighted in: **Nature Biotechnology**, 29, 1101 (2011).

The 1000 Genomes Project Consortium
A map of human genome variation from population-scale sequencing.
Nature 2010, 467: 1061-1073.
◇ Featured on the cover of the journal!
◇ See the 1000 Genomes Project page for media coverage

Iman Hajirasouliha, Fereydoun Hormozdiari, Can Alkan, Jeffrey M. Kidd, Inanc Birol, Evan E. Eichler, S. Cenk Sahinalp
Detection of locus and content of novel sequence insertions using paired-end next-generation sequencing **Bioinformatics** 2010 May 15;26(10):1277-83.

Iman Hajirasouliha, Fereydoun Hormozdiari, S. Cenk Sahinalp, and Inanc Birol
Optimal pooling for genome re-sequencing with ultra-high-throughput short-read technologies
Bioinformatics 2008 Jul 1;24 (13):i32-40.

Layla Oesper

Brown University
Department of Computer Science
Box 1910
Providence, RI 02912

Phone: (719) 235-7738
Email: layla@cs.brown.edu [contact]
Homepage: http://www.cs.brown.edu/people/layla [link]

Education

Brown University, Providence, RI, USA.
Ph.D. Candidate, Computer Science, In Progress since Fall 2010.
Sc.M., Computer Science, 2012. GPA: 4.0
National Science Foundation Graduate Research Fellowhip (2011-2014).

University of Wisconsin at Madison, Madison, WI, USA.
Certificate, Computer Science, 2010. GPA: 4.0

Pomona College, Claremont, CA, USA.
B.A. Mathematics, *Magna Cum Laude,* 2005. GPA: 3.8

Related Publications

Layla Oesper, Ahmad Mahmoody and Benjamin J. Raphael. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology* 2013. 14:R80. [publisher link]. [THetA Software].

Layla Oesper, Ahmad Mahmoody and Benjamin J. Raphael. Inferring Intra-tumor Heterogeneity from High-Throughput DNA Sequencing Data. *17th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2013).* LNCS 7821, 171-172. [publisher link].

Layla Oesper, Anna Ritz, Sarah J. Aerni, Ryan Drebin and Benjamin J. Raphael. Reconstructing Cancer Genomes from Paired-end Sequencing Data. *BMC Bioinformatics* 2012 (Proceedings of 2nd Annual RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-seq)). 13(Suppl 6):S10. [publisher link]. [PREGO Software]

Related Presentations

Layla Oesper, Gryte Satas, Max Song, Simone Dantas and Benjamin J. Raphael. Analysis of Complex Genomic Rearrangements using High-Throughput DNA Sequencing Data. (Platform Presentation) *Wellcome Trust Scientific Conferences/Cold Spring Harbor Laboratory conference on Genome Informatics* November 2013.

Layla Oesper and Benjamin J. Raphael. Reconstructing the Organization of Cancer Genomes. (Invited Talk) 3^{rd} *Workshop on Computational Advances for Next Generation Sequencing (CANGS) in conjunction with* 3^{rd} *IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)* 2013.

Layla Oesper, Ahmad Mahmoody and Benjamin J. Raphael. Inferring Intra-tumor Heterogeneity from High-Throughput DNA Sequencing Data. (Poster Presentation) *CSHL Meeting on the Biology of Genomes*, 2013.

Suzanne Sindi
University of California, Merced
Assistant Professor of Applied Math
Office Phone: (209) 228-4224
Email: ssindi@ucmerced.edu

Research Interests

Mathematical Biology, Computational Biology, Stochastic Processes, Kinetics and Transmission of Prion Diseases, Structural Variation

Education

PhD, University of Maryland, College Park, 2006.
Major: Applied Mathematics and Scientific Computation
Advisor: Yorke, J. A., Hunt, B. R.

BA, *Summa cum laude*, California State University, Fullerton, 2001. (Major: Mathematics, Minor Computer Science)

Professional Positions

Assistant Professor, Applied Math, University of California, Merced (2012-2013).

NSRA Postdoctoral Fellow, Brown University. (2009 - 2012).

Prager Assistant Professor, Brown University. (2006 - 2009).
Division of Applied Mathematics & Center for Computational Molecular Biology

Publications (Last 3 Years)

Sindi, S., Raphael, B. "Identification of Structural Variation" (Book Chapter - Genome Analysis: Current Procedures and Applications.) *In Press*

Olofsson, P., **Sindi, S.** A Crump-Mode-Jagers Branching Process Model of Prion Loss in Yeast. *Journal of Applied/Advances in Probability. In Press*

Weinreich, D., **Sindi, S.**, Watson, R. (2013). Finding the boundary between evolutionary basins of attraction, and implications for Wright's fitness landscape analogy. *Journal of Statistical Mechanics: Theory and Experiment.* (Date Published - 2013).

Sindi, S., Olofsson, P. (2013). A Discrete-Time Branching Process Model of Yeast Prion Curing Curves. *Mathematical Population Studies, 20*, 1-13. (Date Published - 2013).

Sindi, S., Onal, S., Peng, L., Wu, H.T, Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology, 13(3)*. (Date Published - March 2012).

Cáceres, A., **Sindi, S.**, Raphael, B. J., Cáceres, M., González, J. (2012). Identification of polymorphic inversions from genotypes. *BMC Bioinformatics.*

Sindi, S., Raphael, B. (2010). Identification and Frequency Estimation of Inversion Polymorphisms from Haplotype Data. *Journal of Computational Biology, 17(3)*, 517-531.

Derdowski, A., **Sindi, S.**, Klaips, C., DiSalvo, S., Serio, T. (2010). A Size Threshold Limits Prion Transmission and Establishes Phenotypic Diversity. *Science, 330(6004)*, 680-683.

Current Research Support

Sindi, Suzanne (Co-PI), Gary Lupyan (Principal Investigator). "Selection as an Organizing Principal: From Molecules to Minds" NSF-INSPIRE (November 2013 – November 2016).

Sindi, Suzanne (Key Personnel), Serio, Tricia (Principal Investigator), "The Role of Competitive Forces in Prion Propagation and Appearance," NIH - National Institutes of Health. (September 2012 - August 2016).

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Network Analysis of Somatic Mutations in ICGC Whole-Genome Sequences

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Benjamin J. Raphael
Department of Computer Science and Center for Computational Molecular Biology, Brown University, USA
TCGA and ICGC Bioinformatics Analysis Groups

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Max Leiserson, Brown University Hsin-Ta Wu, Brown University	Fabio Vandin, University of Southern Denmark and Brown University

Background and preliminary data

Cancer sequencing projects have demonstrated that most cancer types exhibit extensive mutational heterogeneity with driver mutations found in a large number of genes in different individuals with the same cancer. A major reason for this heterogeneity is that driver mutations target genes in signaling and regulatory pathways. These pathways may be perturbed in numerous ways by multiple types of alterations (e.g. nonsynonymous coding mutations, copy number changes, rearrangements, epigenetic alterations, non-coding mutations, etc.) in any of the genes in the pathway. Thus, obtaining a comprehensive understanding of the somatic alterations in cancer requires the analysis of *combinations* of alterations in *multiple* genes.

We have developed two complementary algorithms, HotNet (Vandin *et al.* 2011) and Dendrix (Vandin, *et al.* 2012), to identify significant *combinations* of driver mutations in cohorts of tumors. HotNet identifies subnetworks of a protein-protein (or protein-DNA) interaction network that are mutated in a statistically significant number of samples using a heat diffusion model to *simultaneously* assess both the significance of mutations in individual proteins and the local topology of a protein's interactions. Dendrix identifies collections of candidate driver pathways *de novo*, without prior knowledge of pathways or protein interactions. Dendrix searches for gene sets that present a pattern of mutually exclusive mutations, a combinatorial constraint derived from knowledge of the somatic mutational process. HotNet has been used in multiple TCGA analysis projects including OV, KIRC, AML, THCA, STAD, and with results published in the corresponding TCGA papers in *Nature* and *NEJM*. Dendrix was used in TCGA Pan-Cancer (Kandoth, *et al*., *Nature*) and Dendrix++, our recent extension using a statistical score with higher sensitivity for rare mutations, was used in TCGA AML project.

During TCGA Pan-Cancer Project, we developed HotNet2, which includes algorithmic improvements necessary to analyze large numbers of samples (>3000 whole exomes in TCGA Pan-Cancer) with a large variation in mutation frequencies. HotNet2 was used in two TCGA Pan-Cancer publications (*Nature Genetics* and *Cell*, in review).

Timelines & resources dedicated to project

Timeline: Initial run of HotNet2 and Dendrix++ complete – April 2014. Development of HotNet2 for non-coding mutations complete – October 2014. Additional runs of HotNet2 and Dendrix with non-coding mutations and rearrangements Complete – January 2015. Manuscript preparation – January to February 2015. Manuscript submission – 20th March 2015.

Resources: Our pan-cancer network analysis is staffed by 2 Ph.D. students and 1 Sc.M. student. In additional to cloud computing resources supplied by the ICGC pan-cancer consortium, we can utilize two compute clusters. The first cluster at Brown's Center for Computing and Visualization contains >3000 Intel Xeon cores, with each node containing >= 24Gb of memory. The second cluster at Brown's Department of Computer Science contains 1820 cores and includes machines with up to 256 Gb RAM. We also have two 16-core workstations, each with 256Gb RAM for development and testing, and used exclusively by our research group.

Research proposal

We will use HotNet2 and Dendrix++ to identify significant combinations of coding and non-coding mutations in the ICGC whole-genome data. Previous analyses with these algorithms have focused on nonsynonymous coding mutations and copy number aberrations. We hypothesize that the whole-genome sequencing data from ICGC will improve these analyses by allowing us to incorporate a wider spectrum of mutations including non-coding mutations and genome rearrangements. Further, the additional cancer types in ICGC (compared to TCGA) will expand our view of how pathways are perturbed in similar and different ways across tumor types. Together, these datasets will enable HotNet/Dendrix to provide a cleaner and more fine-grained view of the mutated driver pathways across cancer types compared to our previous analyses.

Below we provide additional details for the proposed analysis with each algorithm.

HotNet2
We are ready to run HotNet2 immediately on coding variants and copy number aberrations identified by other ICGC groups. In parallel, we will work to incorporate non-coding variants -- and the regulatory interactions they may influence – into the interaction network and gene scores used by HotNet2 to identify significantly mutated subnetworks. We propose to develop such a method, using regulatory interactions identified by the ENCODE project. We also intend to incorporate altered splicing into the analysis in collaboration with Matt Meyerson and Angela Brooks at DFCI (see additional abstract).

Dendrix++
We are ready to run Dendrix++ immediately on coding variants and copy number aberrations identified by other ICGC groups. In contrast to HotNet2 that analyzes combinations of mutations at the gene-level, Dendrix++ analyzes combinations of mutational events, and thus can separately analyze mutations of different types. Thus, Dendrix++ is well suited to the analysis of rearrangements and other non-coding mutations that may not be localized to a specific gene. Furthermore, because patterns of mutual exclusivity can occur due to different genes being mutated in different cancers, we propose to develop a method to exclude cancer-type exclusivity from the Dendrix++ analysis.

Legacy plans

All of the software developed in our research group, including the HotNet and Dendrix programs described above, are freely available for researchers at: http://compbio.cs.brown.edu/software.

We are also starting to release software on source-code repositories including Google Code and GitHub. We are in the process of packaging the group's software into virtual machines: one for Amazon EC2 and one for local use.

As part of this project, we anticipate that we will improve the efficiency and ease of usability of our software, and will release updated versions of the software. Additionally, we plan to write a companion methodology paper to further disseminate these developments and their application to large-scale cancer studies.

BENJAMIN J. RAPHAEL
Associate Professor
Department of Computer Science & Center for Computational Molecular Biology
Brown University
Providence, RI 02912
Phone: (401) 863-7643 Email: braphael@brown.edu Web: http://compbio.cs.brown.edu

EDUCATION

1996-2002 **Ph.D.** in Mathematics, University of California, San Diego.

1992-1996 **S.B.** in Mathematics, **S.B. Minor** in Biology, Massachusetts Institute of Technology.

EXPERIENCE

2013-present **Director**, Center for Computational Molecular Biology, Brown University
2011-present **Associate Professor**, Department of Computer Science & Center for Computational Molecular Biology, Brown University
2006-2011 **Assistant Professor**, Department of Computer Science & Center for Computational Molecular Biology, Brown University

SELECTED RECENT PUBLICATIONS

C.Kandoth, [16 additional authors], **B.J. Raphael**, L. Ding. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* 502(7471):333-9

L. Oesper, A. Mahmoody, **B.J. Raphael**. (2013) THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology* 14:R80.

The Cancer Genome Atlas Research Network. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43-9

M.D.M. Leiserson, D. Bloch, R. Sharan*, **B.J. Raphael***. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLOS Computational Biology.* May;9(5):e1003054. *equal contribution

The Cancer Genome Atlas Research Network. (2013) Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*. 368(22):2059-74.

Sindi, S. Onal, L. Peng, H. Wu, **B.J. Raphael**. (2012) An Integrative Probabilistic Model for Identification of Structural Variation in Sequencing Data. *Genome Biology* 13(3):R22.

F. Vandin, E. Upfal, **B.J. Raphael**. (2012) *De novo* Discovery of Mutated Driver Pathways in Cancer. *Genome Research.* 22(2):375-85. [Journal version of paper accepted at *RECOMB 2011*].

Vandin F, Upfal E, **B.J. Raphael** . (2011) Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology.* 18(3):507-22. [Journal version of paper accepted at *RECOMB 2010*].

International Cancer Genome Consortium. (2010) International network of cancer genome projects. *Nature*. 464(7291):993-8.

S. Sindi, E. Helman, A. Bashir, **B.J. Raphael**. (2009) A Geometric Approach for Classification and Comparison of Structural Variants. *Bioinformatics* 25: i222-i230. *[Proceedings of ISMB/ECCB 09*]

Mark D.M. Leiserson

Ph.D. candidate in Computer Science and Computational Biology, Brown University

Brown University
Box 1910
Providence, RI 02912

Phone: (203) 927-8678
Email: mdml@cs.brown.edu
Homepage: http://maxleiserson.com/

Education

Ph.D., Computer Science and Computational Biology, Brown University *Expected May 2016*

Advisor: Professor of Computer Science Benjamin J. Raphael.

M.Sc., Computer Science, Brown University *May 2013*

Advisor: Professor of Computer Science Benjamin J. Raphael.

Thesis: *Methods for Identifying Driver Pathways in Cancer*.

B.Sc. (*cum laude*), Computer Science, Tufts University *May 2011*

Advisor: Assistant Professor of Computer Science Benjamin J. Hescott.

Thesis (*highest honors*): *Inferring Mechanisms of Compensation from E-MAP and SGA Data Using Local Search Algorithms for Max Cut*.

Honors / Awards

National Science Foundation Graduate Research Fellow, 2012-present.

Tufts University Senior Thesis in Computer Science completed with Highest Honors, 2011.

Runner-up, Computing Research Association Outstanding Undergraduate Award, 2011.

Tufts University Senior Award, Benjamin G Brown Scholarship for Promise in Scientific Research, 2011.

Research Positions

Brown University, Department of Computer Science and Center for Computational Molecular Biology *2012-present*
National Science Foundation Graduate Research Fellow

Brown University, Department of Computer Science and Center for Computational Molecular Biology *2011-2012*
Graduate research assistant

Tufts University, Department of Computer Science *2008-2011*
Undergraduate research assistant

Selected Publications

M.D.M. Leiserson, D. Blokh, R. Sharan, B. Raphael. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Biol*, 9(**5**):e1003054.

M.D.M. Leiserson, D. Tatar, L. Cowen, B. Hescott. (2011) Inferring Mechanisms of Compensation from E-MAP and SGA Data Using Local Search Algorithms for Max Cut. *Journal of Computational Biology*, **18**(11):1399-1409.

Last updated: November 26, 2013
http://maxleiserson.com/docs/cv.pdf

HSIN-TA WU

115 Waterman St., 4th Flr., Providence, RI, 02912, USA
Tel: +1 401.450.7621 Email: bournewu@cs.brown.edu

EDUCATION

Sep. 2010-present	**Ph. D.** in Computational Biology and Computer Science, **Brown University** Advisor: Benjamin J. Raphael
Sep. 2008-Jul. 2010	**Sc. M.** in Computer Science, **Brown University**
Sep. 2004-Jul. 2006	**M. S.** in Bioinformatics, **National Yang Ming University**, Taipei, Taiwan
Sep. 2000-Jun. 2004	**B. S.** in Computer Science, **National Cheng Chi University**, Taipei, Taiwan

PUBLICATIONS

JOURNALS

- Hsin-Ta Wu, Iman Hajirasouliha, Benjamin J. Raphael. (2013) A combinatorial algorithm to identify recurrent copy number aberrations. In review.
- The Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45(10):1113-20.
- The Cancer Genome Atlas Research Network. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43-9.
- The Cancer Genome Atlas Research Network. (2013) Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *The New England Journal of Medicine* 369(1):98.
- Souriya Vang, Hsin-Ta Wu, Andrew Fischer, Daniel H. Miller, Shannon MacLaughlan, Elijah Douglass, Margaret Steinhoff, Colin Collins, Peter J. S. Smith, Laurent Brard, Alexander S. Brodsky. (2013) Identification of Ovarian Cancer Metastatic miRNAs. *PLOS One* 8(3):e58226.
- Suzanne S Sindi, Selim Önal, Luke C Peng, Hsin-Ta Wu and Benjamin J. Raphael. (2012) An Integrative Probabilistic Model for Identification of Structural Variation in Sequencing Data, *Genome Biology* 13(3):R22.
- X. Yuan, Z. Z. Hu, H. T. Wu, M. Torii, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, and C. H. Wu (2006) An online literature mining tool for protein phosphorylation. *Bioinformatics* 22(13):1668-1669.

CONFERENCE PAPER

- Dai, H.J., P.T. Lai, C.H. Huang, Y.C. Chang, Y.Y. Bow, H.T. Wu, Richard T.H. Tsai and W.L. Hsu (2009). IASL-IISR Interactor Normalization System: Using a Multi-stage Gene Normalization Algorithm and SVM-based ranking. *Proceedings of BioCreAtIvE II.5 Challenge Evaluation Workshop*, Madrid, Spain.

THESES

- H. T. Wu, J. H. Chiang, U. C. Yang (2006), Identify disease associated gene by literature mining, *Master Thesis*.

COMPUTER SKILLS AND PROFICIENCY

Scope	Specific tools/ languages	Level of expertise
Object-oriented languages	Java	Proficient
	C++	Expert
Scripting languages	Python, Perl, Javascript	Proficient
Database design	MySQL, PostgreSQL, ORACLE	Proficient
Shell scripting	BASH	Proficient
Data exchange	XML, XSL	Knowledgeable
OS administration	Linux, Unix, FreeBSD	Proficient

Fabio Vandin

Department of Computer Science, and Center for Computational Molecular Biology
Brown University
115 Waterman St., 4th Floor, Providence, RI 02912 USA
Mobile: +1-401-286-1402, *Fax:* +1-401-863-7657
E-mail: fabio_vandin@brown.edu
WWW: www.cs.brown.edu/~vandinfa/home.html

Current Position

Adjunct Assistant Professor of Computer Science (Research)
Department of Computer Science, and Center for Computational Molecular Biology,
Brown University.

Education

Ph.D. in Information Engineering, University of Padova, 04/2010.
Thesis Title: *Mining of Significant Patterns: Theory and Practice.*
Advisor: Prof. A. Pietracaprina.

Sample publications

F. Vandin, E. Upfal, and B. J. Raphael. Algorithms for Detecting Significantly Mutated Pathways in Cancer. **Journal of Computational Biology**, 18(3):507-22, 2011.
F. Vandin, E. Upfal, and B. J. Raphael. De novo Discovery of Mutated Driver Pathways in Cancer. **Genome Research**, 22(2):375-85, 2012. Epub 2011 Jun. 7.
The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. **New England Journal of Medicine**, May 1^{st} 2013.
The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. **Nature Genetics**, 45(10):1113-20 (26 September 2013).
C.Kandoth*, M.D. McLellan*, F. Vandin, K.Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J.F. McMichael, M.A. Wyczalkowski, M.D.M. Leiserson, C.A. Miller, J.S. Welch, M.J. Walter, M.C. Wendl, T.J. Ley, R.K. Wilson, B.J. Raphael, L. Ding. Mutational landscape and significance across 12 major cancer types. **Nature**, 502, 333-339 (17 October 2013).

Research Grants

National Science Foundation, **BIGDATA: Mid-Scale: Analytical Approaches to Massive Data Computation with Applications to Genomics**. Role: co-PI (PI: Eli Upfal). Total amount: $1,566,685. *10/2012-09/2016.*

National Science Foundation, **AF: Small: Algorithmic Problems in Protein Structure Studies**. Role: PI for the last two years of the award, total amount: $225,001. Collaborative with Gopal Pandurangan, and Chris Bailey-Kellogg. *06/2011-08/2013.*

Program Committees

SPIRE, International Symposium on String Processing and Information Retrieval 2011 (PC Member).
RECOMB-seq, Second Annual RECOMB Satellite Workshop on Massively Parallel Sequencing 2012 (PC Member).
RECOMB, International Conference on Research in Computational Molecular Biology 2014 (PC Member).

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Joint analysis of cancer-specific expression and RNA processing patterns across cancer types

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Gunnar Rätsch, Ph.D., MSKCC, Nikolaus Schultz, Ph.D., MSKCC

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Kjong Lehmann, MSKCC	

Background and preliminary data

While there have been many case-control association studies of various cancer types across multiple tumors in order to determine disease risk from germline mutations and in order to determine somatic driver mutations, little effort has been spent on analyzing expression and splicing patterns across various tumors to understand underlying molecular differences and their genomic causes. We are currently in the final stages of this quantitative trait analysis of > 3,000 TCGA tumor samples across 12 cancer types. While this is the largest dataset of matched genetic and RNA processing data, many cancer-specific challenges need to be addressed in order to perform a quantitative trait association analysis across the very heterogeneous tumor types. Our analysis has been focused on understanding splice pattern difference across cancer and within cancer types and the discovery of the genetic variants responsible for those changes. During this pioneer effort we have developed a pipeline to jointly re-analyze the TCGA whole exome sequencing data as well as the RNA-seq data including re-mapping, SNP-calling, expression and splice quantification as well as appropriate quality control filtering. While this joint analysis has not been done previously, it is nevertheless crucial in order to perform an appropriate joint QTL analysis across all TCGA tumor types. We have also developed a novel strategy to address the phenotypic and genetic heterogeneity observed across tumors in order to allow us to control the Type-1 error. Within our common variant and rare variant association analysis we have discovered genetic variants in trans and *cis* factors associated with changes in splice patterns of various cancer census genes. In order to gain power from the different cancer types we have also designed a meta-analysis pipeline allowing us to detect variants commonly associated with RNA-seq processing alterations across cancer types. We anticipate that the incorporation of whole genome sequencing data compared to the whole exome sequencing data used so far, will yield significant better resolution in promoter regions to understand regulatory expression changes. Further, the anticipated additional matched RNA-seq and WGS data will help improve power of our rare variant association strategy to overcome challenges in the discovery of rare somatic variations to gain new insights into the genetic mechanisms of splicing and expression changes in cancer.

Timelines & resources dedicated to project

Joint preprocessing and reanalysis of TCGA/ICGC RNA-seq/WGS/WXS can be achieved by May/June 2014. This involves the re-mapping (~2 Month) /variant calling (~ 1 Month) / expression quantification and splicing quantification (~ 1 Month)/ quality control (~ 1 Month). By July/August 2014 first common variant association analysis should become available with rare variant association analysis and meta-analysis across cancer types completed by Dec 2014. Manuscript preparation and submission by early 2015

Research proposal

The extended dataset will include additional cancer types with whole genome sequencing data as well as RNA-seq data which can be leveraged to not only gain additional power to detect the underlying genetic causes of RNA processing differences across cancer types, but also to understand cancer type specific expression changes and developments. The anticipated combined TCGA/ICGC dataset with ~1,500 samples of RNA-sequencing and cDNA microarray data with matching whole genome sequencing datasets will particularly help us explore the effect of rare somatic variants. We are specifically interested in understanding the genetic bases of expression changes during cancer progression, which has been challenging previously, due to low resolution of Affymetrix SNP chips as well as low coverage of promoter regions for exome sequencing data. In addition, the whole genome data will help us to refine our exome-based analyses for RNA-processing phenotypes (as the causal variant can be several hundred base-pairs into the intron). The matched methylation data will contribute towards this analysis and allow us to explore the interaction between genetic variants and methylation changes and their effect onto expression changes during cancer progression. Major efforts will include the clean joint analysis of these heterogeneous datasets in order to not only improve the quality of SNP calls but also to remove analysis specific artifacts from the data. While this pipeline has already been set up, it will require significant computation time as well as solutions to address technical and quality differences between ICGC and TCGA datasets. Common variant association study utilizing linear mixed models, will be performed to undertake a QTL analysis of expression and splicing patterns while accounting for population structure and tumor specific heterogeneity. We will also estimate possible hidden confounding effects to improve power. Rare variant association studies will be developed tailored towards finding rare regulatory variants equivalent to a rare variant association study developed for the detection of rare splicing variants in the TCGA dataset previously. Meta-analysis will allow us to leverage information across all tumor types to find genetic mechanism undetectable by the analysis of a specific cancer type itself. Since the amount of RNA processing data is limited in comparison to the whole genome and whole exome sequencing data, we would develop methods for phenotype imputation to utilize the remaining SNP data available.

Legacy plans

The Rätsch group has a long-standing history of providing galaxy services for all the tools developed within this group as well as useful independent counter parts. In order to continue this dedication to make our results and intermediate products available, we will make all association summary statistics publicly available to allow for use in other studies. In addition, we plan to extend RNA-seq related resources that we already developed for 12 Pan-Cancer tumor types: namely, a unified alignment of all RNA-seq libraries, the joint detection of alternative splicing events and the variant-aware quantification of gene expression. We believe that these resources will also be of interest for others in the project.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 14th November, 2013 (midnight your local time). Explanatory notes follow the form.

Title of abstract

Timing mutational processes in cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Paul Spellman, OHSU

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Molecular profiling of tumor at different points of its progression will provide insights into early biology of tumors and facilitate early diagnosis of cancer. It is difficult to obtain tumor samples at different points. So, we need a model that can temporally order chromosomal abnormalities from a single time point sample. By obtaining maximum likelihood estimates (MLE) from mutation allele frequencies in copy-neutral LOH (CN-LOH) and copy gain segments of a chromosome, chromosomal abnormalities can be temporally ordered allowing a reconstruction of the events that occurred in each tumor. We have evaluated this method and introduced a full MLE model that is stable and accounts for sequencing errors.

This method has been implemented in timing mutational signatures in breast cancer using breast cancer whole genomes from TCGA. Distinct mutation signatures separated most of Her2, Basal, and Luminal samples. This suggests that there are different biological mechanisms involved in DNA damage and repair mechanisms underlying each breast cancer subtype. Her2 breast cancer samples were enriched for C>T and C>G mutations at TCX tri-nucleotides. Recent studies have shown that APOBEC proteins are implicated in the underlying mutational process for this signature. Basal breast cancer samples were enriched for C>G mutations at GCX and CCX tri-nucleotides. Luminal samples were enriched for ACG>ATG, CCG>CTG and GCG>GTG mutations. Mutation signals for Basal and Her2 were stronger than for Luminal, since there were only 6 Luminal samples. We also broke down mutational signatures into early and late events and observed that mutations from APOBEC mutagenesis occur late.

We used limited breast cancer samples in this analysis; whole genomes from ICGC Pancan will allow us to examine timing of mutation signatures and to study mutational patterns across different cancers.

Timelines & resources dedicated to project

Two studies in TCGA Pancan have shown that mutagenesis from APOBEC is carcinogenic and is seen in several cancer types. We have observed in breast cancer that APOBEC mutagenesis occurs late during tumor progression. Timing events using whole genomes from ICGC Pancan will allow timing APOBEC mutagenesis in all cancer types.

To do timing analysis on a sample, we would require mutation calls, allele frequencies, copy number estimates, and purity of the sample.

Research proposal

Timing analysis

Mutations in single gain, double gains, and CN-LOH regions will be considered for timing, as these regions provide direct measure of relative age of the event. Regions with less than 20 mutations will be eliminated because higher number of mutations is required for stable estimates of time of event occurrence. Copy number estimates will be used to identify CN-LOH, single and double gain segments of a chromosome. Allele frequencies of the mutations will be adjusted for normal contamination.

If π is the probability of a mutation occurring at a certain stage (A), $C\pi$ is normalization constant and q is the probability of an allele frequency. Then based on a full MLE model q can be estimated as $q=(1/C\pi)*A\pi$ (Purdom *et al*.,2013). A homozygous mutation with a high q value indicates early mutation, implying that mutation has occurred before a chromosomal event, resulting in a double copy number. A heterozygous mutation with a high q value indicates late mutation, implying that a mutation occurred after a chromosomal event, so they appear haploid in copy number. Confidence intervals for timing of events will be estimated from several bootstrap runs.

Mutation signatures

Mutations will be classified into 96 tri-nucleotide mutation types as described by Nik-Zainal *et al*. Counts for each mutation type will be scaled, by taking arcsine of the square root of count proportions. Hierarchical clustering will be performed on scaled counts to study mutational signature patterns across cancer types. In addition timing estimates from timing analysis will allow us to examine timing of mutational signatures.

References

Nik-Zainal, S., *et al*. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell,* 149, 979-993.

Purdom, E., *et al.* (2013). Timing Chromosomal Abnormalities within Cancer Samples. *Bioinformatics.* (In review)

Legacy plans

R package cancerTiming is available to the research community to do timing analysis.

International Cancer Genome Consortium

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Pathogen detection in 2000 WGS data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

PI: Xiaoping Su, Ph.D.
Associate Professor, TCGA affiliate, Dept. of Bioinformatics and Computational Biology
UT MD Anderson Cancer Center, Houston, TX 77030, U.S.A.
Email: xsu1@mdanderson.org

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Yunxin Chen, M.S. Dept. of Bioinformatics and Computational Biology UT MD Anderson Cancer Center Houston, TX 77030, U.S.A. Email: ychen20@mdanderson.org	Gabriel G. Malouf, M.D Assistant Professor Groupe Hospitalier Pitié-Salpêtrière Department of Medical Oncology University Pierre and Marie Curie (Paris VI), AP-HP, Paris, France. Email: gabriel.malouf@psl.aphp.fr

Background and preliminary data

Background: The association between viral infection and neoplasia is well established in a wide variety of cancer types. One of the best understood causal relationships is between human papillomaviruses (HPV) and squamous neoplasia of the head-and-neck regions. However, determining the potential role of viruses in tumorigenesis has proven to be extremely elusive. Even with the development of newer technologies, establishing the presence of the virus, the temporality of the infection, and the influence of viral factors in carcinogenesis remains challenging. This work will help understand the association between viral infection and other common types of malignancies. More specifically, our analysis will provide the landscape of virus integration sites, as well their genomic and epigenomic features.

Preliminary data: We developed a software, VirusSeq, for cancer-associated virus and its integrated site discovery by paired-end whole genome sequencing (NGS) data or RNA-Seq. VirusSeq will search a comprehensive virus database (Genome Information Broker for Viruses- GIB-V) for any known virus in patient samples, and quantify virus representation by a measure of the virus genome coverage (or overall count of mapped reads) for all patient samples to determine candidate tumor-associated viruses. VirusSeq will also identify the integrated genes by DNA virus if it indeed integrates into human cancer genome. VirusSeq has been published in Bioinformatics (Chen et al., 2013), and was used to analyze more than 3500 TCGA RNA-Seq tumor samples, which has been published in Journal of Virology (Khoury et al., 2013).

References:

Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29(2):266-7, 1/2013. e-Pub 11/2012. PMCID: PMC3546792.

Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, Thompson EJ, Meric-Bernstam F, Medeiros LJ, Weinstein JN, Su X. The Landscape of DNA Virus Associations Across Human Malignant Cancers

Using RNA-Seq: An Analysis of 3775 Cases. J Virol. 87(16):8916-26, 8/2013. e-Pub 6/2013. PMID: 23740984.

Timelines & resources dedicated to project

Timelines:

1. Viruses and its integration sites detection in both WGS data and RNA-Seq
   – November 2013 to December 2014
2. Viral copy number analysis from WGS data and its correlation with the fraction of APOBEC-mediated mutations
   – June 2014 to September 2014
3. Correlation of virus integration sites with histone chromatin marks in a tissue-specific type
   – September 2014 to December 2015
4. Manuscript preparation – January 2015 to February 2015
5. Manuscript submission – 20th March 2015.

Resources dedicated to project

1. Xiaoping Su: 20% effort
2. Gabriel G Malouf: 10% effort
3. Yunxin Chen: 90% effort

Research proposal

We plan to fulfill the 4 following aims:
Aim 1: We will apply our own software VirusSeq to analyze 2000 WGS samples for detection of cancer-associated viruses and their integration sites.
Aim 2: We will apply our own software VirusSeq to analyze 1500 RNA-Seq samples for detection of cancer-associated viruses and their integration sites. It is possible that some viruses may be able to contribute to the carcinogenic process without being expressed. Thus, these viruses might be detected by WGS data, but won't be detected by RNA-Seq. We will make appropriate distinctions regarding the difference of findings from DNA (WGS) versus RNA sequencing studies and how each contributes to our understanding of associations between viruses and tumors.
Aim 3: we will investigate viral copy number at the DNA level and test whether this is related to the fraction of APOBEC-mediated mutations across different tumor types. Also we will look at whether these APOBEC-mediated mutations are enriched at sites close to where viral integration has occurred since this has been linked to localized genomic instability.
Aim 4: we will investigate the correlation of virus integration sites with histone chromatin marks in a tissue-specific type. This will be done using data extracted from the roadmap epigenomics project.

Legacy plans

Our legacy plans are:

1. Describe the landscape of viral integration across tumor sub-types, in the aim to discover whether there are regions in the genome that are more prone than others for virus integration.
2. Perform genome-wide integrative analysis for viral integration and histone chromatin marks, to unravel a putative histone code for virus integration across tumor-subtypes.

CURRICULUM VITAE

Xiaoping Su, Ph.D.

TCGA GDAC Affiliation:

- I am the official member of TCGA MDACC GDAC, and I have also led the virus detection in TCGA Bladder (BLCA) AWG.

Primary Academic Appointment:

- Xiaoping Su, Ph.D.
  Associate Professor
  Department of Bioinformatics and Computational Biology
  Division of Quantitative Sciences
  The University of Texas MD Anderson Cancer Center, Houston, TX 77030, U.S.A.
  Phone: 713-792-5508
  Email: xsu1@mdanderson.org

EDUCATION:

- Chongqing University, Chongqing, China, BS, 1989, Computer Sciences
- Mount Sinai School of Medicine of New York University, New York, NY, PHD, 2000, Biostatistics

Selected Peer-reviewed Publications:

1. **Su X**, Zhang L, Zhang J, Meric-Bernstam F, Weinstein JN. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics 28(17):2265-6, 9/2012. e-Pub 6/2012. PMCID: PMC3426843.
2. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, **Su X**. VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29(2):266-7, 1/2013. e-Pub 11/2012. PMCID: PMC3546792.
3. Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, Thompson EJ, Meric-Bernstam F, Medeiros LJ, Weinstein JN, **Su X**. The Landscape of DNA Virus Associations Across Human Malignant Cancers Using RNA-Seq: An Analysis of 3775 Cases. J Virol. 87(16):8916-26, 8/2013. e-Pub 6/2013. PMID: 23740984.
4. Mullighan CG, Phillips LA, **Su X**, Ma J, Miller CB, Shurtleff SA, Downing JR. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. Science 322(5906):1377-80, 11/2008. PMCID: PMC2746051.
5. Mullighan CG, **Su X**, Zhang J, Radtke I, Phillips LA, Miller CB, Ma J, Liu W, Cheng C, Schulman BA, Harvey RC, Chen IM, Clifford RJ, Carroll WL, Reaman G, Bowman WP, Devidas M, Gerhard DS, Yang W, Relling MV, Shurtleff SA, Campana D, Borowitz MJ, Pui CH, Smith M, Hunger SP, Willman CL, Downing JR. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med 360(5):470-80, 1/2009. e-Pub 1/2009. PMCID: PMC2674612.
6. Obenauer JC, Denson J, Mehta PK, **Su X**, Mukatira S, Finkelstein DB, Xu X, Wang J, Ma J, Fan Y, Rakestraw KM, Webster RG, Hoffmann E, Krauss S, Zheng J, Zhang Z, Naeve CW. Large-scale sequence analysis of avian influenza isolates. Science 311(5767):1576-80, 3/2006. e-Pub 1/2006. PMID: 16439620.
7. **Su X**, Wallenstein S. New approximation for the distribution of r-scan statistics. Statistics and Probability Letters 46(4):411-419, 2/2000.
8. **Su X**, Wallenstein S, Bishop D. Nonoverlapping clusters: approximate distribution and application to molecular biology. Biometrics 57(2):420-6, 6/2001. PMID: 11414565.
9. Benson G, **Su X**. On the distribution of k-tuple matches for sequence homology: a constant time exact calculation of the variance. J Comput Biol 5(1):87-100, 1998. PMID: 9541873.

CURRICULUM VITAE

Yunxin Chen, M.S.
Bioinformatics Analyst
Department of Bioinformatics and Computational Biology
Division of Quantitative Sciences
The University of Texas MD Anderson Cancer Center,
Houston, TX 77030, U.S.A.
Phone: (713)-563-7204 (Office)
Email: YChen20@mdanderson.org

Education:

- **Master of Science in Computer Sciences** (12/2000)
  City University of New York, NY
- **Bachelor in Chemical Engineering** (1992)
  Beijing University of Technology, Beijing, China

Publications:

1. **Chen Y**, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29(2):266-7, 1/2013. e-Pub 11/2012. PMCID: PMC3546792.
2. Khoury JD, Tannir NM, Williams MD, **Chen Y**, Yao H, Zhang J, Thompson EJ, Meric-Bernstam F, Medeiros LJ, Weinstein JN, Su X. The Landscape of DNA Virus Associations Across Human Malignant Cancers Using RNA-Seq: An Analysis of 3775 Cases. J Virol. 87(16):8916-26, 8/2013. e-Pub 6/2013. PMID: 23740984.

Bioinformatics Experience:

1. Implementation of bioinformatics algorithm for detection of virus and its integration sites using whole genome sequencing (NGS) or RNA-Seq of human cancer samples.
2. Development of whole exome next-generation sequencing (Illumina) analysis pipeline for detection of SNVs and somatic mutations, and for full annotation of somatic mutations. This pipeline has been applied to Xp11 renal cell carcinoma whole exome sequencing project, and has successfully detected novel non-synonymous somatic mutations.

Bioinformatics Skills:

- **Next-gen Sequence Analysis Package**: BWA, Bowtie/TopHat, Mosaik, BLAST/BLAT, VarScan2, ANNOVAR/SIFT, Samtools/Picard.
- **Programming languages**: JAVA, Perl, C/C++, SQL.
- **Operating Systems**: Linux, Unix, Windows.

CURRICULUM VITAE
Gabriel G Malouf, M.D., M.Sc

Primary Academic Appointment:
Assistant Professor
Department of Medical Oncology, Groupe Hospitalier Pitié-Salpêtrière,
Assistance Publique Hôpitaux de Paris, Faculty of Medicine Pierre et Marie Curie,
Institut Universitaire de Cancérologie GRC5, University Paris 6,
43 boulevard de l'Hôpital, Paris, France
Phone : + 33 1 42 16 05 18
E-mail: gabriel.malouf@psl.aphp.fr

Education:

- Master of science in biochemistry, University of Paris VII, France (2009)
- French National Board in medical oncology, University of Paris V, France 2012
- French state MD, University of Paris V, France 2005
- Post-doctoral research training (November 2009- October 2011): Leukemia department and Cancer Epigenetic Center, mentor (Jean-Pierre Issa), University of Texas, MD Anderson Cancer Center, Texas, USA

Selected Peer-reviewed Publications :

1. Architecture of Epigenetic Reprogramming Following Twist1 Mediated Epithelial-Mesenchymal Transition. **Malouf GG**, Taube JH, Lu Y, Roysarkar T, Panjarian S, Estecio MR, Jelinek J, Yamazaki J, Raynal NJ, Long H, Tahara T, Tinnirello A, Ramachandran P, Zhang X, Liang S, Mani SA and Issa JP. Genome Biology, *in press*
2. Transcriptional Profiling of Pure Fibrolamellar Hepatocellular Carcinoma Reveals an Endocrine Signature. **Malouf GG,** Job S, Paradis V, Fabre M, Faivre S, de Reyniès A, Raymond E. Hepatology, *in press*
3. NRAS Mutation Is the Sole Recurrent Somatic Mutation in Large Congenital Melanocytic Nevi. Charbel C, Fontaine RH, **Malouf GG**, Picard A, Kadlub N, El-Murr N, How-Kit A, Su X, Coulomb-L'hermine A, Tost J, Mourah S, Aractingi S, Guégan S. J Invest Dermatol. 2013 Oct 15.
4. Epigenetic silencing of microRNA-203 is required for EMT and cancer stem cell properties. Taube JH, **Malouf GG**, Lu E, Sphyris N, Vijay V, Rosen JM, Issa JP, Calin GA, Chang JT, Mani SA. Scientific Reports, 2013;3:2687.
5. Genomic Heterogeneity of Translocation Renal Cell Carcinoma. **Malouf GG**, Monzon FA, Couturier J, Molinie V, Escudier B, Camparo P, Su X, Yao H, Tamboli P, Lopez-Terrada D, Picken M, Varella-Garcia M, Multani A, Pathak S, Wood CG, Tannir NM. Clin Cancer Res. 2013 Jul 1.
6. The epigenome of AML stem and progenitor cells. Yamazaki J, Estecio MR, Lu Y, Long H, **Malouf GG**, Graber D, Huo Y, Ramagli L, Liang S, Kornblau SM, Jelinek J, Issa JP. Epigenetics. 2013 Jan;8(1):92-104.
7. Histone deacetylase inhibitors as anti-neoplastic agents. Batty N, **Malouf GG**, Issa JP. **Cancer Lett.** 2009 Aug 8;280(2):192-200. Epub 2009 Apr 3.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Charting the structural genome and transcriptome variant landscape in cancer

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Roel G.W. Verhaak, PhD
Ken Chen, PhD

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Kosuke Yoshihara, MD, PhD **Wanding Zhou, PhD**	**Luay Nakhleh, PhD, Rice University**

Background and preliminary data

High throughput sequencing facilitates the measurement of quantity, structure and composition of DNA and RNA molecules in a biological sample. ***Somatic variation in the genome and the transcriptome plays a critical role in cancer and accurate characterization may thus provide important insights into tumorigenesis.*** This proposal builds upon the following preliminary studies from our team: 1) Our software has accurately and reproducibly processed thousands of DNA and RNA sequences, for detection of somatic structural genomic and transcript variants across many cancer types; 2) We have integrated the results from our DNA and RNA pipelines to detect different categories of somatic rearrangements, such as balanced translocations, inversion-tandem duplications, transposable element insertion, virus integration, and deletion-associated events; 3) we have publicly released our methods and they have been widely adopted by TCGA, the 1000 Genomes project and many research groups. Included are 1) BreakDancer (Chen et al., Nature Methods, 2009), 2) CREST (Wang et al., Nature Methods, 2011), both of which discovers genomic rearrangement breakpoints, 3) TIGRA (Chen et al., Genome Research, 2013), which performs targeted assembly of discovered breakpoints and outputs breakpoint sequences and precise breakpoint structure, 4) BreakDown (Xian et al, in preparation), which computes genotype likelihood and estimate individual breakpoint allelic fractions in clonally heterogeneous data, 5) PRADA (Bioinformatics minor revision) and 6) BreakFusion (Chen et al., Bioinformatics, 2012), which discover gene fusions from RNA-seq data, and 7) BreakTrans (Chen et al., Genome Biology, 2013), which associates genomic and transcriptomic breakpoints.

In our recently published studies of the somatic genomic landscape of glioblastoma (GBM, Zheng et al, Genes Dev 2013; Brennan et al, Cell 2013) we confirmed 41 of 49 fusion transcripts detected in the RNA sequencing data of 27 GBM in the matching whole genome sequencing data, showing our capacity to integrate RNA-seq and DNA-seq output. Similarly, We have applied these tools to identify over 20 fusion transcripts in 20 TCGA AML samples (N Engl J Med., 2013) and over 100 fusion transcripts in 45 TCGA breast cancer samples that have matched RNA-seq and DNA-seq data, including fusions that involve multiple genomic breakpoints (Chen et al Genome Biology, 2013). In a pilot effort to compare fusion transcripts across many types of cancers, and to illustrate the ability of our to process large numbers of samples, we have analyzed the paired-end RNA sequencing data of 3,787 TCGA tumors and 303 TCGA tumor-associated normal samples, spanning twelve tumor types and arrived at a comprehensive list of 7,636 somatic cancer-associated fusions. We classified the final fusion list by tumor type and fusion type, thereby defining the spectrum of somatic fusion genes across cancers. This effort illustrated, for example, that transcript fusions in a tumor type that is characterized by a high frequency of somatic copy number variation, such as ovarian carcinoma, are dominated by genes within 1Mb that are fused as a result of inversion-duplication. Our findings show our ability to process massive amounts of DNA-seq and RNA-seq data.

Timelines & resources dedicated to project

Depending on the resources available, we estimate that it will take approximately six to nine months to successfully analyze all data sets using our pipelines. We will contribute the intellectual input and time of both principal investigators, as well as the efforts of at least two computationally trained and experienced postdoctoral researchers. We are dedicated to share our results and tools with the TCGA, ICGC and general research communities and to attend scientific meetings that may come forth from these efforts. We are supported by TCGA, the 1000 Genomes project and at least 2 NIH grants: R01-CA172652 (Delineating Heterogeneous Structural Complexity in Cancer Genomes, PI: Ken Chen) and U41-HG007497 (An Integrative Analysis of Structural Variation for the 1000 Genomes Project, PI: Charles Lee, MDA-PI: Ken Chen). In addition to the computational cloud infrastructure provided by ICGC, our institutional high performance cluster, which consists of over 8000 CPU and 5 PB storage, is available for we smaller-scale integrative analysis.

Research proposal

Structural variation of the genome and transcriptome plays a major role in cancer (Mittelman et al, Nat Rev Cancer 2007) but effective detection of rearrangement lesions has been hampered by our lacking ability to recognize true events and eliminate false positives. We propose to use our extensive prior experience in analyzing large amounts of sequencing data using our established pipelines to **1.** Effectively and accurately learn and establish the structural DNA and RNA alteration landscape through analysis of 2,000 whole genome sequences and 1,500 RNA sequences and **2.** Provide methodological insights into how to detect cancer-driving structural DNA and RNA variant lesions with high sensitivity. **3.** Increase our understanding on how to detect and validate fusion transcripts and genomic variants when RNA sequencing data or array-based copy number profiles but now whole genome sequencing are available **4.** Characterize the nucleotide architecture of rearrangements to gain insight in evolutionary mechanisms and variant functions. Importantly, in addition to transcribed and non-transcribed gene-gene alterations, we will include intragenic variants (such as the EGFR vIII) and lesions targeting non-coding RNAs. The latter will be facilitated but the unique alignment strategy implemented in PRADA, in which read pairs are mapped to a combined transcriptome-genome reference, and in post-processing all transcriptome alignments are collapsed into single genomic coordinates. As a result, both read pairs mapping to exonic regions as well as to unannotated transcripts are captured. BreakDancer/CREST/TIGRA identifies somatic rearrangements by comparing tumor-normal pairs. As matched normal RNAseq is typically lacking, we have previously analyzed 303 tumor-adjacent TCGA normal RNAseq samples to establish a list of fusion predictions consisting of data artefacts and non-somatic fusion transcripts (such as the *TFG-GPR128* transcripts resulting from a germline copy number variant on chromosome 13), which will be a crucial tool to adequately reduce the number of false predictions in the analysis of tumor samples.

We have previously observed that approximately 55% of fusion transcripts detected in the RNA sequencing data of 164 GBM samples are associated with DNA breakpoints that can be identified in DNA copy number data generated using the Affymetrix SNP6.0 platform (Zheng et al, Genes Dev, 2013). We will use the integration of RNA sequencing and whole genome sequencing structural variant predictions to further understand which RNA events can be validated using DNA copy number data, thereby providing further opportunity of analysis of the large number of samples in ICGC and TCGA in which RNA sequencing or DNA copy number profile, but not whole genome sequencing, are available.

Our first milestone will be reached when the 2,000 whole genome sequences have been analyzed using BreakDancer, CREST and TIGRA. The second milestone is to detect fusion transcripts in the approximately 1,500 matching cDNA samples using PRADA and BreakFusion. Upon completion, the two output lists will be integrated by BreakTrans and other tools, which will serve multiple purposes. First, genomic lesions that can be matched to fusion transcripts will validate both DNA and RNA rearrangement predictions. Second, the properties of validated DNA variants will be used to learn which invalidated alterations are likely to be accurate as well. The large amount of computational resources necessary to complete the analysis will be provided through this grant.

Legacy plans

We anticipate that we will significantly improve our existing analytical pipelines through the integration of DNA and RNA rearrangement predictions. In particular, we expect to be able to make recommendations on how to sensitively detect true events, which may benefit tumor characterization projects in which only RNA sequencing or DNA sequencing data are available. Importantly, we will include such validation strategies in our publicly available pipelines and thus provide mechanisms for the research community to benefit.

CURRICULUM VITAE
Roeland GW Verhaak, PhD

PRESENT TITLE AND AFFILIATION

Primary Appointment: Assistant Professor, Department of Bioinformatics and Computational Biology
Adjunct Appointment: Assistant Professor, Department of Genomic Medicine

EDUCATION

MSc: Radboud University, Nijmegen, Netherlands, 2000, Biomedical Science and Computer Science
PhD: Erasmus University Medical Center, Rotterdam, Netherlands, PHD, 2006, Medicine
Research Fellowship: Broad Institute of Harvard and MIT, Cambridge, MA, Meyerson Lab, 2007-2010

SELECTED PUBLICATIONS (from 44 publications total)

1. Valk PJ, **Verhaak RG**, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 350(16):1617-28, 4/2004.
2. **Verhaak RG**, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood 106(12):3747-54, 12/2005.
3. **Verhaak RG**, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17(1):98, 1/2010.
4. **Verhaak RG**, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J Clin Invest 123(1):517-25, 1/2013.
5. Zheng S, …, **Verhaak RG**. A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. Genes Dev 27(13):1462-72, 7/2013. e-Pub 6/2013.
6. Yang JY, Yoshihara K, …, **Verhaak RG**. Predicting time to ovarian carcinoma recurrence using protein markers. J Clin Invest 123(9):3740-50, 9/2013.
7. Brennan CW, **Verhaak RG**, et al. The somatic genomic landscape of glioblastoma. Cell 155(2):462-77, 10/2013.
8. Yoshihara K, …, **Verhaak RG**. Inferring tumour-purity and stromal and immune cell admixture from expression data. Nat Commun 4:2612, 2013.

SELECTED GRANTS (from 8 funded grants total)

1. Investigator, 25%, An Integrative Pipeline for Analysis & Translational Application of TCGA Data (GDAC), 5 U24 CA143883 05, NIH/NCI, PI - John N. Weinstein, 9/29/2009-7/31/2014,
2. Principal Investigator, Elucidating the mechanisms that shape the genome of post-treatment GBM, N/A, The University of Texas M D Anderson Cancer Center, 6/15/2013-6/4/2014,
3. Core Leader, 20%, Brain Tumor Therapeutic Efficacy by Quantitative Magnetic Resonance, 2P01 CA085878-01A1, NIH/NCI Sub Contract from the University of Michigan, PI - Brian Ross, 7/1/2013-6/30/2018,
4. Co-Leader, Project 2 (Targeting the PI3K Pathway in Malignant Glioma), 5%, SPORE in Brain Cancer, 2 P50 CA127001-06, NIH/NCI, PI - Lang, Frederick, 9/17/2013-8/31/2018
5. Co-director, Biostatistics and Bioinformatics Core, 5%, SPORE in Brain Cancer, 2 P50 CA127001- 06, NIH/NCI, PI - Lang, Frederick, 9/17/2013-8/31/2018

HONORS AND AWARDS

- Netherlands Genomics Initiative Fellowship, 2006
- Fundamental and pre-clinical fellowship, Dutch Cancer Society KWF, 2008-2010
- Wilson S. Stone Memorial Award, UT MD Anderson Cancer Center, 2011
- Peter Steck Memorial Award, Pediatric Brain Tumor Foundation, 2013

NAME	POSITION TITLE
Chen, Ken	Assistant Professor
	MD Anderson Cancer Center

A. Education

Tsinghua University, Beijing, China	BE	1996	Precision Instruments
University of Illinois, Urbana-Champaign, IL	PHD	2004	ECE
University of California, San Diego, CA	Research Fellowship	2004-2005	Biochemistry

B. Positions and Honors

2009-2011 Research Instructor, Department of Genetics, Washington University, St. Louis, MO

2011-present Assistant Professor, Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX

C. Selected Peer-reviewed Publications

1. **Chen K**, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Res 17(5):659-66, 5/2007. e-Pub 4/2007. PMCID: PMC1855178.
2. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216):1061-8, 10/2008. e-Pub 9/2008. PMCID: PMC2671642.
3. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, **Chen K**, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456(7218):66-72, 11/2008. PMCID: PMC2603574.
4. **Chen K**, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6(9):677-81, 9/2009. e-Pub 8/2009. PMCID: PMC3661775.
5. Koboldt DC, **Chen K**, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25(17):2283-5, 9/2009. e-Pub 6/2009. PMCID: PMC2734323.
6. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-73, 10/2010. PMCID: PMC3042601.
7. Mills RE, Walter K, Stewart C, Handsaker RE, **Chen K**, et al. Korbel JO, 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59-65, 2/2011. PMCID: PMC3077050.
8. Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S, Wallis J, **Chen K**, Payton JE, Fulton RS, Veizer J, Schmidt H, Vickery TL, Heath S, Watson MA, Tomasson MH, Link DC, Graubert TA, DiPersio JF, Mardis ER, Ley TJ, Wilson RK. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 305(15):1577-84, 4/2011. PMCID: PMC3156695.
9. Larson DE, Harris CC, **Chen K**, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics. e-Pub 12/2011. PMCID: PMC3268238.
10. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, **Chen K**, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 8(8):652-4, 2011. e-Pub 6/2011. PMCID: PMC3527068.
11. Ding L, et al. Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing. Nature 481, 1/2012. e-Pub 1/2012. PMCID: PMC3267864.
12. **Chen K**, Wallis JW, Kandoth C, Kalicki-Veizer JM, Mungall KL, Mungall AJ, Jones SJ, Marra MA, Ley TJ, Mardis ER, Wilson RK, Weinstein JN, Ding L. BreakFusion: Targeted Assembly-based Identification of Gene Fusions in Whole Transcriptome Paired-end Sequencing Data. Bioinformatics. e-Pub 5/2012. PMCID: PMC3389765.
13. Welch JS, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell 150(2):264-78, 7/2012. PMCID: PMC3407563.
14. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56-65, 11/2012. PMCID: PMC3498066.
15. The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med. e-Pub 5/2013. PMID: 23634996.
16. **Chen K**, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: A Targeted Iterative Graph Routing Assembler for breakpoint assembly. Genome Res. e-Pub 12/2013. PMID: 24307552.
17. **Chen K**, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, McLellan MD, Hoadley KA, Mardis ER, Ley TJ, Perou CM, Wilson RK, Ding L. BreakTrans: uncovering the genomic architecture of gene fusions. Genome Biol 14(8). e-Pub 8/2013.
18. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, **Chen K.** CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. PLoS One 8(10):e77945, 2013. e-Pub 10/2013. PMID: 24205039.

Curriculum vitae

Name: Kosuke Yoshihara
Address: Department of Bioinformatics and Computational Biology,
University of Texas MD Anderson Cancer Center
1400 Pressler Street, Houston, Texas, U.S.A. Zip code: 77030
E-mail: kyoshihara@mdanderson.org
Citizenship: Japan
Gender: Male
Date of Birth: October 27, 1978

Academic Degree:

April 1, 1997-March 31, 2003 M.D. Niigata University School of Medicine
April 1, 2005-March 23, 2009 Ph.D. Niigata University Graduate School of Medical and Dental Sciences (Molecular Biology)

Employment of experience:

April 1, 2003-March 31, 2004 Niigata University Medical and Dental Hospital
April 1, 2004-March 31, 2005 Niigata City General Hospital
April 1, 2009-December 31, 2011 Niigata University Graduate School of Medical and Dental Sciences (Department of Obstetrics and Gynecology)

Licenses and Certifications:

May 2003 Medical License: Japan, #430857
October 2008 Obstetrics and gynecology specialist approved by Japan Society of Obstetrics and Gynecology

Memberships:

American Association for Cancer Research (Associate member)

Selected Publications:

1. Yoshihara K, …, Verhaak RG. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
2. Yang JY, Yoshihara K, …, Verhaak RG. Predicting time to ovarian carcinoma recurrence using protein markers. J Clin Invest. 2013; 123: 3740-50.
3. Yoshihara K, et al. High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by down-regulation of antigen presentation pathway. Clin Cancer Res. 2012;18:1374-85

Wanding Zhou

(832)·900·9468 ◇ wzhou1@mdanderson.org ◇ https://wanding.bitbucket.org
Department of Bioinformatics and Computational Biology ◇ UT MD Anderson Cancer Center

ACADEMIC TRAINING

MD Anderson Cancer Center - Dept of Bioinformatics and Computational Biology 2013 -
Odyssey Fellow *Houston, TX*

· Advisor - Ken Chen

EDUCATION & TRAINING

Rice University - Department of Bioengineering 2007 - 2013
Doctor of Philosophy, GPA 4.07 *Houston, TX*

· Dissertation - Modeling and Evolutionary Analysis of Microbial Metabolism
· Advisor - Luay K. Nakhleh

Fudan University - School of Life Science 2003 - 2007
Bachelor of Science, GPA 3.72 *Shanghai, China*

· Thesis - Molecular Dynamic Simulation and Surface Analysis of the Aggregation of β-amyloid Fibril
· Advisor - Guanghong Wei

University of Hong Kong - Faculty of Science 2005
Exchange Student, GPA 3.95 *Hong Kong, China*

PUBLICATIONS

Refereed Journal Publications

· N. Berestovsky, W. Zhou, D. Nagrath, and L. Nakhleh, *Modeling Integrated Cellular Machinery Using Hybrid Petri-Boolean Networks*, PloS Computational Biology, 9(11):e1003306, 2013
· W. Zhou and L. Nakhleh, *Quantifying and assessing the effect of chemical symmetry in metabolic pathways*, Journal of Chemical Information and Modeling, 52(10):2684-2696, 2012
· W. Zhou and L. Nakhleh, *Convergent Evolution of Modularity in Metabolic Networks Through Different Community Structures*, BMC Evolutionary Biology, 12:181, 2012, *(accompanying software)*
· W. Zhou and L. Nakhleh, *The Strength of Chemical Linkage as A Criterion for Pruning Metabolic Graphs*, Bioinformatics, 27(14):1957-1963, 2011
· W. Zhou and L. Nakhleh, *Properties of Metabolic Graphs: Biological Organization or Representation Artifacts?*, BMC Bioinformatics, 12:132, 2011

Working Manuscripts

· W. Zhou, Tenghui Chen, Hao Zhao, Agda Karina Eterovic, Funda Meric-Bernstam, Gordon B. Mills and Ken Chen, *Bias from removing read duplication in ultra-deep sequencing experiments*, in preparation
· W. Zhou and L. Nakhleh, *An Evolutionary Analysis of the Gain and Loss of Metabolic Capabilities in Proteobacteria*, in preparation
· W. Zhou and L. Nakhleh, *ReAA: An Open-source Tool for the Analysis of Metabolic Reaction Atom Mappings*, in preparation

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Luay Nakhleh	Associate Professor

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Technion, Israel	B.Sc.	1992-1996	Computer Science
Texas A&M University, College Station, TX	M.CS.	1997-1998	Computer Science
University of Texas at Austin	Ph.D.	1999-2004	Computer Science

A. Positions and Honors

2004-2010	Assistant Professor, Department of Computer Science, Rice University, Houston, TX
2010-present	Associate Professor, Department of Computer Science, Rice University
2006	The Early Career Award, The Department of Energy, U.S.A.
2009	The Early Career Award, National Science Foundation, U.S.A.
2010	The Alfred P. Sloan Research Fellowship.
2012	Guggenheim Fellowship

B. Selected Peer-reviewed publications (Selected from over 79 peer-reviewed publications**)**

1. Ruths, D., **L. Nakhleh**, M.S. Iyengar, S.A.G. Reddy, and P.T. Ram, 2006. Graph-theoretic hypothesis generation in biological signaling networks. Journal of Computational Biology, 13(9): 1546-1557.
2. Ruths, D., J.T. Tseng, **L. Nakhleh**, and P.T. Ram, 2006. De novo signaling pathway predictions based on protein-protein interaction, targeted therapy, and protein microarray analysis. Proceedings of the
3. RECOMB Satellite Workshop on Systems Biology and Proteomics. Lecture Notes in Bioinformatics, LNBI #4532: 109-119.
4. Ruths, D., M. Muller, J.T. Tseng, **L. Nakhleh**, and P.T. Ram, 2008. The signaling Petri net-based simulator: A non-parametric strategy for characterizing the dynamics of cell-specific signaling networks." PLoS Computational Biology, 4(2): e1000005.
5. Ruths, D., **L. Nakhleh**, and P.T. Ram, 2008. Rapidly exploring structural and dynamic properties of signaling networks using PathwayOracle. BMC Systems Biology, 2:76.
6. Ruths, T., D. Ruths, and **L. Nakhleh**, 2009. GS2: An efficiently computable measure of GO-based similarity of gene sets. Bioinformatics, 25(9): 1178-1184.
7. Ruths, D. and **L. Nakhleh**, 2010. Deriving predictive models of signaling network dynamics from qualitative experimental data. Proceedings of the 9th Annual International Conference on Computaitonal Systems Biology, 136-145.
8. Zhou, W. and **L. Nakhleh**, 2011. Properties of metabolic graphs: Biological organization or representation artifacts? BMC Bioinformatics, 12: 132.
9. Zhou, W. and **L. Nakhleh**, 2011. The strength of chemical linkage as a criterion for pruning metabolic graphs. Bioinformatics, 27(14): 1957-1963.
10. Y. Lu et al., 2011. Kinome siRNA-phosphoproteomic screen identifies networks regulating AKT signaling. Oncogene, 30(45): 4567-4577.
11. N. Berestovsky, R. Fukui, and **L. Nakhleh**, 2012. On the performance of particle swarm optimization for parameterizing kinetic models of cellular networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2012.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of cross-cancer signatures of somatic mutation and tumor heterogeneity from WGS data.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

- Wenyi Wang, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, TCGA Prostate AWG
- David Wheeler, Human Genome Sequencing Center, Baylor College of Medicine.

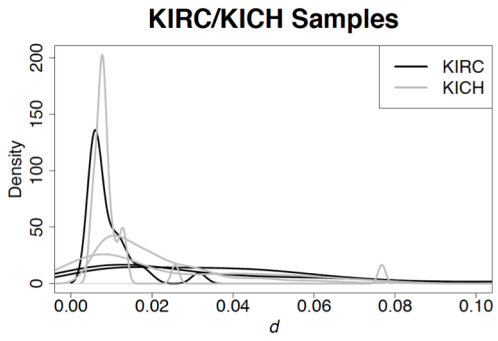
Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

- Yu Fan, Postdoctoral fellow, The University of Texas MD Anderson Cancer Center

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

Cancer is an evolutionary process involving accumulation and selection of somatic mutational events. The use of next-generation sequencing (NGS) on matched tumor–normal sample pairs is critical for discovery of somatic variation. However, accurate detection of somatic point mutations remains a challenge, due to the genetic heterogeneity in the tumor sample, from admixture of normal cells and the presence of multiple subclones of tumor cells. We present a Bayesian phylogenetic method, Mutation Somatic Evolution estimation (MuSE), for describing the evolution from the reference allele to the tumor and the normal allelic composition at a single nucleotide position. Our proposed method incorporates the probability of sequencing errors and computes the unknown allele frequencies, multiple alternative alleles and the rates of nucleotide transition/transversion. All model parameters are estimated using the maximum likelihood or the Markov chain Monte Carlo (MCMC) method. The figure on the left shows how our evolution distance measure *d* can be used to depict position specific mutational events (group of positions clustering at distinct values of d) across samples (shown in lines) and across cancer types (KIRC and KICH).



Timelines & resources dedicated to project

We will re-analyze the WGS BAM files with MuSE, a newly-developed approach to somatic mutation calling (Yu, Wheeler and Wang, *in preparation*). MuSE is packaged into an executable package we plan to run on one or more of the PanCancer data installations. MuSE will operate on aligned tumor and normal paired BAM files as they become available in the cloud between November 2013 and June 2014. Mutation calling should be completed no later than September 2014. Analysis of tumor heterogeneity will commence by September 2014 and be completed by the end of the year. The acquisition of tumor/metastasis and primary tumor/recurrent pairings will give us an opportunity to study the evolution of a tumor as it becomes increasingly aggressive and more difficult to treat.

Research proposal

Based on the virtual-tumor benchmarking approach, MuSE outperforms MuTect in the sensitivity and specificity comparison when the coverage is above 50X and allele fraction is greater 0.2, therefore we plan to use MuSE to call all WES data to complement the variant calling of the 2000 WGS samples. MuSE provides evolutionary distance *d* per genomic position, which reflects the expected allelic changes within the tumor cell population at that position. We will use *d* as input to identify subclasses of mutation positions, estimate the heterogeneity and clonal architecture for every sample and every cancer type. Common patterns in these mutation classes will likely emerge across cancer types.

Legacy plans

The computational steps used to produce publication-ready results on the pan-cancer data set will be embodied in executable code that we will install in one or more of the data servers with the help of IT support personnel at the 5 centers. Our code will be sufficiently well documented to enable replication by third parties. The Human Genome Sequencing Center at Baylor College of Medicine has a long history of data sharing and open source code sharing.

CURRICULUM VITAE of Wenyi Wang

EDUCATION

2003-2007 JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH Baltimore, MD
PhD, Biostatistics

2001-2003 COLUMBIA UNIVERSITY COLLEGE OF PHYSICIANS AND SURGEONS
MA, Human nutrition New York City, NY

1997-2001 FUDAN UNIVERSITY Shanghai, China
BS, Honor Science Program, Biology

PUBLICATIONS (Selected)

Articles

1. Ahn J, Liu S, **Wang W***, Yuan Y*. Bayesian latent-class mixed-effect hybrid models for dyadic longitudinal data with non-ignorable dropouts. *Biometrics* in press. *corresponding authors
2. The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 2013, 45: 1113-1120.
3. Shen P*, **Wang W***, Chi AK, Fan Y, Davis RW, Scharfe C. Target capture using double-stranded DNA probes. *Genome Medicine* 2013, 5:50 doi:10.1186/gm454. *authors contributed equallly
4. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, Wistuba II, and **Wang W**. DeMix: deconvolution for mixed cancer transcriptomes. *Bioinformatics* 2013 doi: 10.1093/bioinformatics/btt301.
5. Srivastava S, **Wang W**, Zinny PO, Colen RR, Baladandayuthapani V. Integrating multi-platform genomic data using hierarchical bayesian relevance vector machines. *EURASIP Journal on Bioinformatics and Systems Biology* 2013:9 doi:10.1186/1687-4153-2013-9.
6. Peng G, Fan Y, Palculict TB, Shen P, Ruteshouser EC, Chi A, Davis RW, Huff V, Scharfe C, **Wang W**. Rare variant detection using family-based sequencing analysis. *Proceedings of the National Academy of Sciences.* ePub, February 20, 2013, doi: 10.1073/pnas.1222158110.
7. Zhang N, Xu Y, O'Hely M, Speed TP, Scharfe C, **Wang W**. SRMA: an R package for sequence based calling in candidate genes with custom resequencing microarrays. *Bioinformatics.* e-Pub 05/2012.
8. Shen P*, **Wang W***, Krishnakumar S, Palm C, Chi AK, Enns GM, Davis RW, Speed TP, Mindrinos MN, Scharfe C. High-quality DNA sequence capture of 524 disease candidate genes. *Proceedings of the National Academy of Sciences.* 2011, Apr 19;108(16):6549-54. Epub 2011 Apr 5.

*authors contributed equally

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2. Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME David A. Wheeler, Ph.D.	POSITION TITLE Associate Professor of Molecular and Human Genetics
eRA COMMONS USER NAME:	

EDUCATION/TRAINING (*Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Maryland, College Park, MD	B.S.	1972	Biochemistry/Zoology
The George Washington University, Washington, DC	M.S.	1976	Biochemistry
The George Washington University, Washington, DC	Ph.D.	1983	Genetics

A. Personal Statement

Dr. Wheeler develops methods for discovery of genome variation in human and animal populations using DNA sequencing technologies with the goal of relating polymorphism to human disease, especially cancer. His work in this area involves large-scale multi-center national and international projects such as TCGA and ICGC and other cancer sequencing projects that aim to comprehensively catalogue all mutations leading to cancer. Dr. Wheeler is a recognized expert in mutation discovery and analysis in the cancer genome.

B. Positions and Employment

2004-2006 Co-Director for Bioinformatics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

2004-2013 Associate Professor, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

2006-present Director, Cancer Genomics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

2010-present Assistant Director, Human Genome Sequencing Center, Baylor College of Medicine, Houston TX.

2013-present Professor, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

C. Selected Peer-reviewed Publications

1. **Wheeler DA**, Srinivasan M, Egholm M, Shen Y, Chen L et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-6. PMID: 18421352
2. Ding L, Getz G, **Wheeler DA**, Mardis ER, McLellan MD et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455:1069-1075. PMCID: PMC2694412
3. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, **Wheeler DA**, Gibbs RA, Yu F (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20:273-80. Epub 2009 Dec 17. PMCID: PMC2813483
4. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, ..., **Wheeler DA**, Pearson JV, McPherson JD, Gibbs RA, Grimmond SM. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491: 399-405.
5. Powell BC, Jiang L, Muzny DM, Treviño LR, Dreyer ZE, Strong LC, **Wheeler DA**, Gibbs RA, Plon SE. (2012). Identification of TP53 as an Acute Lymphocytic Leukemia Susceptibility Gene Through Exome Sequencing. Pediatric Blood and Cancer 60: E1-3.
6. The Cancer Genome Atlas Research Network. (2013). Integrative analysis of genomic and molecular alterations in clear cell renal cell carcinoma. Nature 499: 43-49.

D. Research Support

1U24CA143843-04 ($10M, PI: Wheeler) 09/29/09 – 07/31/14 NCI:The BCM Tumor Genome Characterization Center . The major goals of this project are to analyze sets of tumors plus, when appropriate, matched normal tissue to characterize and enumerate the somatic changes occurring in 500 patients for each of 20-25 tumors types over the next 5 years.

Yu Fan

Department of Bioinformatics & Computational Biology, Unit 1410 • The University of Texas MD Anderson Cancer Center • P. O. Box 301402 • Houston TX 77230 • E-Mail: yfan1@mdanderson.org

Education

2011: Ph.D., Ecology and Evolutionary Biology (Phylogenetics), University of Connecticut
2004: Molecular Evolution and Systematics, Institute of Oceanology Chinese Academy of Sciences
2001: B.S., Marine Biology, Ocean University of China

Experience

2012-Current	Postdoctoral Fellow, MD Anderson Cancer Center
2004-2011	Research/Teaching Assistant, University of Connecticut
2001-2004	Research Assistant, Institute of Oceanology Chinese Academy of Sciences
2000-2001	Research Assistant, Ocean University of China

Publications

Lewis, P. O., W. Xie, M. H. Chen, **Y. Fan**, L. Kuo. Posterior predictive Bayesian phylogenetic model selection. *Systematic Biology*. doi:10.1093/sysbio/syt068.
Fan Y., R. Wu, M. H. Chen, L. Kuo and P. O. Lewis. 2013. A conditional autoregressive model for detecting natural selection in protein-coding DNA sequences. Topics in Applied Statistics Springer Proceedings in Mathematics & Statistics 55: 203-212.
Shen P., W. Wang, A. K. Chi, **Y. Fan**, R. W. Davis and C. Scharfe. 2013. Multiplex target capture with double-stranded DNA probes. *Genome Medicine* 5: 50.
Peng G., **Y. Fan**, T. B. Palculict, P. Shen, E. C. Ruteshouser, A. K. Chi, R. W. Davis, V. Huff, C. Scharfe, W. Wang. 2013. Rare variant detection using family-based sequencing analysis. *Proceedings of the National Academy of Sciences* 110(10): 3985-90.
Wang W., **Y. Fan** and T. P. Speed. 2013. DNA variant calling in targeted sequencing data. Advances in Statistical Bioinformatics. Cambridge University Press.
Fan Y., R. Wu, M. H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution* 28(1): 523-532.
Xie, W., P. O. Lewis, **Y. Fan**, L. Kuo, and M. H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology* 60(2): 150-160.
Wickett, N. J., **Y. Fan**, P. O. Lewis, and B. Goffinet. 2008. Distribution and evolution of pseudogenes, gene losses, and a gene rearrangement in the plastid genome of the nonphotosynthetic liverwort, *Aneura mirabilis* (Metzgeriales, Jungermanniopsida). *Journal of Molecular Evolution* 67(1): 111-122.
Fan Y., X. Z. Li, L. S. Song, and Z. H. Cai. 2004. Phylogenetic Relationships of Five Species of Dorippinae (Crustacea, Decapoda) Revealed by 16S rDNA Sequence Analysis. *Acta Oceanologica Sinica* 23(3): 513-519.

Awards and Honors

- Selected Poster Awardee in the Symposia on Cancer Research 2013 - "Genomic Medicine"
- Invited Speaker in the International Chinese Statistical Association (ICSA) 2012 Applied Statistics Symposium
- Doctoral Dissertation Fellowship of the University of Connecticut 2011
- Ecology & Evolutionary Biology (EEB) Department Summer Award 2011
- Nominated for the EEB Department Excellence in Student Teaching Award 2011
- The paper “Choosing among partition models in Bayesian phylogenetics” was awarded in the conference “Frontiers of Statistical Decision Making and Bayesian Analysis - in Honor of James O. Berger” 2010
- College of Liberal Arts and Sciences Fellowship to the Top Graduate School Applicants of the University of Connecticut 2004
- Institute of Oceanology Chinese Academy of Sciences Fellowship 2001-2004
- Ocean University of China Fellowship 1997-1999

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Analysis of cross-cancer miRNA mutation patterns from WGS data.

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

- David Wheeler, Human Genome Sequencing Center, Baylor College of Medicine.
- Sean McGuire, Department of Molecular and Cellular Biology, Baylor College of Medicine

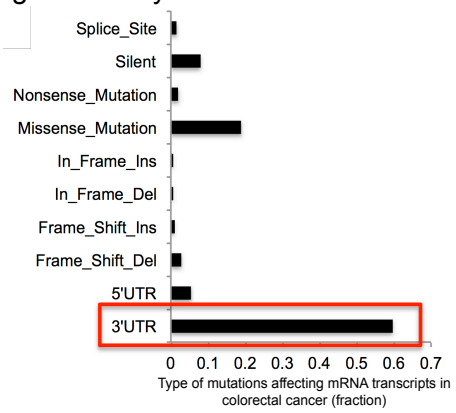
Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

- Mark Hamilton. Graduate Student, Department of Molecular & Cell Biology, Baylor College of Medicine, Houston, TX.
- Christian Coarfa, Department of Molecular and Cell Biology, Baylor College of Medicine, Houston, TX

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

It is well known that miRNAs modulate tumorigenesis through suppression of specific genes. As many tumor types rely on overlapping oncogenic pathways, a core set of microRNAs may exist, which consistently drives or suppresses tumorigenesis in many cancer types. As such, mutation of microRNA target sites on mRNA 3' untranslated regions (UTRs) may represent an important, but currently understudied, mechanism of tumorigenesis. This hypothesis is supported by the fact that ~50-60% of tumor mutations affecting mRNAs fall within the 3'UTR. Improved definition of cis-regulatory microRNA target sites and integration of mRNA expression data to better assess functional impacts of these mutations is necessary to separate any putative driver mutation event in this region. We recently integrated The Cancer Genome Atlas pan-cancer data set with a microRNA target atlas composed of publicly available Argonaute Crosslinking Immunoprecipitation (AGO-CLIP) data, a technology that genomically defines microRNA binding sites. Through this analysis we defined mutations in microRNA target sites using the AGO-CLIP microRNA target atlas and TCGA exome-sequencing data using a novel algorithm, miSNP, which is designed to identify mutations in microRNA binding sites and then calculate if the target site mutations are associated with changes in mRNA expression (1). While this analysis successfully identified functional mutations in microRNA binding sites, the analysis was limited because our results were based on fewer than 1000 3' UTRs present in whole exome sequencing data. We now propose to replicate and extend these findings in the WGS pan-cancer data of the ICGC where well-annotated 3'UTR mutation profiles will provide us with a powerful ability to analyze mutational impact on microRNA binding.



(1) Mark P. Hamilton, Kimal Rajapakshe, Sean M. Hartig, Boris Reva, Michael D. McLellan, Cyriac Kandoth, Li Ding, Travis I. Zack, Preethi H. Gunaratne, David A. Wheeler, Cristian Coarfa & Sean E. McGuire. (2013). Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. Nature Comm. *in press*.

Timelines & resources dedicated to project

This project will be executed starting with the availability of the standard somatic point mutation data sets

in September 2014 and be completed by January 2015.

Research proposal

Part 2. The procedure developed for the analysis of the mutational impact on microRNA seed targets, called miSNP, will be run on all WGS and WES data sets. The AGO-CLIP atlas will be expanded and integrated with traditional 3'UTR motif analysis to provide a more robust set of microRNA target sites for the analysis. Information on mutation of predicted miRNA seeds will be compared to miRNA expression profiles, which will be available for at least half the patients in the study. This analysis will examine mutations in the mRNA 3'UTR region which contains ~50-60% of all mutations affecting mRNAs, providing a powerful new method to understand this possible affect of mutation on cis-regulatory sites. The analysis will be as reported in ref 1.

Legacy plans

The computational steps used to produce publication-ready results on the pan-cancer data set will be embodied in executable code that we will install in one or more of the data servers with the help of IT support personnel at the 5 centers. Our code will be sufficiently well documented to enable replication by third parties. The Human Genome Sequencing Center at Baylor College of Medicine has a long history of data sharing and open source code sharing.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2. Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME David A. Wheeler, Ph.D.	POSITION TITLE Associate Professor of Molecular and Human Genetics
eRA COMMONS USER NAME:	

EDUCATION/TRAINING (*Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Maryland, College Park, MD	B.S.	1972	Biochemistry/Zoology
The George Washington University, Washington, DC	M.S.	1976	Biochemistry
The George Washington University, Washington, DC	Ph.D.	1983	Genetics

A. Personal Statement

Dr. Wheeler develops methods for discovery of genome variation in human and animal populations using DNA sequencing technologies with the goal of relating polymorphism to human disease, especially cancer. His work in this area involves large-scale multi-center national and international projects such as TCGA and ICGC and other cancer sequencing projects that aim to comprehensively catalogue all mutations leading to cancer. Dr. Wheeler is a recognized expert in mutation discovery and analysis in the cancer genome.

B. Positions and Employment

2004-2006 Co-Director for Bioinformatics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

2004-2013 Associate Professor, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

2006-present Director, Cancer Genomics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

2010-present Assistant Director, Human Genome Sequencing Center, Baylor College of Medicine, Houston TX.

2013-present Professor, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

C. Selected Peer-reviewed Publications

1. **Wheeler DA**, Srinivasan M, Egholm M, Shen Y, Chen L et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-6. PMID: 18421352
2. Ding L, Getz G, **Wheeler DA**, Mardis ER, McLellan MD et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455:1069-1075. PMCID: PMC2694412
3. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, **Wheeler DA**, Gibbs RA, Yu F (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20:273-80. Epub 2009 Dec 17. PMCID: PMC2813483
4. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, ..., **Wheeler DA**, Pearson JV, McPherson JD, Gibbs RA, Grimmond SM. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491: 399-405.
5. Powell BC, Jiang L, Muzny DM, Treviño LR, Dreyer ZE, Strong LC, **Wheeler DA**, Gibbs RA, Plon SE. (2012). Identification of TP53 as an Acute Lymphocytic Leukemia Susceptibility Gene Through Exome Sequencing. Pediatric Blood and Cancer 60: E1-3.
6. The Cancer Genome Atlas Research Network. (2013). Integrative analysis of genomic and molecular alterations in clear cell renal cell carcinoma. Nature 499: 43-49.

D. Research Support

1U24CA143843-04 ($10M, PI: Wheeler) 09/29/09 – 07/31/14 NCI:The BCM Tumor Genome Characterization Center . The major goals of this project are to analyze sets of tumors plus, when appropriate, matched normal tissue to characterize and enumerate the somatic changes occurring in 500 patients for each of 20-25 tumors types over the next 5 years.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
McGuire, Sean E	Assistant Professor
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YYYY	FIELD OF STUDY
Harvard University, Cambridge, MA	BA	6/1994	Biochemistry
Baylor College of Medicine, Houston, TX	PHD	6/2003	Medicine
Baylor College of Medicine, Houston, TX	MD	6/2004	Medicine
The University of Texas Health Science at Houston, Houston, TX	Clinical Internship	7/2004-6/2005	Internal Medicine
The University of Texas MD Anderson Cancer Center, Houston, TX	Clinical Residency	7/2005-6/2009	Division of Radiation Oncology

A. Personal Statement: My laboratory has generated an atlas of all publically available Argonaute Cross-link Immunoprecipitation (AGO-CLIP) datasets to facilitate accurate microRNA-target prediction. Currently we are generating the first AGO-CLIP datasets in prostate cancer, and will begin applying this approach to xenografts to more closely model in vivo microRNA dynamics including the generation of circular microRNA "sponges" based on a core motif common to the pan-cancer oncogenic microRNAs identified above and are testing them for therapeutic activity *in vivo*.

B. Positions and Honors

Positions and Employment

2009-present Assistant Professor, Department of Radiation Oncology, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX
2009-present Assistant Professor, Department of Molecular & Cellular Biology, Baylor College of Medicine, Houston, TX

Honors

2003 Harold Weintraub Award, Fred Hutchinson Cancer Center
2004 Donald B. Lindsey Prize, Society for Neuroscience
2004 Larry Sandler Award, Genetics Society of America for Outstanding Ph.D.
2004 Presentation of the Larry Sandler Memorial Lecture, 43rd Annual Drosophila Research Conference
2010 Distinguished Young Alumnus Award, Baylor College of Medicine

C. Selected Peer-reviewed Publications

9. McGuire AL, Cho MK, **McGuire SE**, Caulfield T. Medicine. The future of personal genomics. *Science* 317(5845):1687, 9/2007. PMCID: PMC2220016.
10. **McGuire SE,** McGuire AL. Don't throw the baby out with the bathwater: enabling a bottom-up approach in genome-wide association studies. *Genome Res* 18(11):1683-5, 11/2008. PMID: 18974262.
13. **McGuire SE,** Lee AK, Cerne JZ, Munsell MF, Levy LB, Kudchadker RJ, Choi SL, Nguyen QN, Hoffman KE, Pugh TJ, Frank SJ, Corn PG, Logothetis CJ, Kuban DA. PSA Response to Neoadjuvant Androgen Deprivation Therapy Is a Strong Independent Predictor of Survival in High-Risk Prostate Cancer in the Dose-Escalated Radiation Therapy Era. *Int J Radiat Oncol Biol Phys* 85(1):e39-46, 1/2013. e-Pub 10/2012. PMID: 23102837.
15. Weinstein JN, Collison EA, Mills GB,… **McGuire SE**,… Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45(10):1113-20, 10/2013. PMID 24071849.
16. Hamilton MP, Rajapakshe K, Hartig SM, Reva B, McLellan MD, Kandoth C, Ding L, Zack TI, Gunaratne PH, Wheeler DA, Coarfa C, **McGuire SE**. Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. *Nature Communications* 4:2730, 11/2013. PMID 24220575.

MARK P. HAMILTON - BIOGRAPHICAL SKETCH

NAME Hamilton, Mark Patrick	POSITION TITLE Student

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Baylor College of Medicine, Houston Texas	MD	2017 (anticipated)	Medicine
Baylor College of Medicine, Houston Texas	PhD	2015 (anticipated)	Molecular and Cellular Biology
Austin College, Sherman Texas	B.A.	2008	Biology and English

Personal Statement

I am currently a GS3 MSTP student at Baylor College of Medicine. My long-term goal is to become an outstanding physician scientist with a clinical and laboratory focus in oncology and computational analysis of tumors. My current goal is to achieve my PhD in Molecular and Cellular Biology by completing projects that will train me in the modern high-throughput informatic techniques that will define the next generation of cancer research. My research focuses on integrative analysis of microRNA function in human tumors using genomic data sets. Specifically I employ Argonaut crosslinking immunoprecipitation (AGO-CLIP) technology to determine global microRNA binding sites in tumor cells. I have spent the last two and a half years of my PhD work analyzing AGO-CLIP data, in addition to traditional microRNA-sequencing, RNA-sequencing, and DNA-sequencing data sets. I have recently developed a new method to probe mutations in microRNA binding sites and successfully employed this approach in the TCGA Pan-Cancer dataset to identify new methods of microRNA regulation of tumor cells. I have successfully authored two first-author peer-reviewed publications and co-authored three additional publications. These publications include a first-author publication in *Nature Communications* detailing my work on the Pan-Cancer project and a consortia authorship in *Nature Genetics*. As such I have extensive experience working with sequencing data across multiple tumors and have developed a novel pipeline specifically designed to incorporate RNA-sequencing, genome sequencing, and AGO-CLIP data to determine mutations in microRNA binding sites. This analysis is important because it may confer relevance to mutations in the 3'UTR region of genes, which makes up ~60% of all mutations affecting mRNAs. Despite the relative abundance of 3'UTR mutations, they are often ignored when analyzing genome-sequencing data. I believe I am uniquely suited to provide a strong contribution to the proposed application through my experience in analyzing 3'UTR mutations in pan-tumor datasets.

Research Positions

Graduate Research, August 2011 – present – I currently work in the lab of Dr. Sean McGuire, MD/PhD, at Baylor College of medicine. My research focuses on Pan-Tumor mechanisms by which microRNAs drive cancer.

Surgery and Medicine Clinical Rotations, January – June 2011 – Completed my Surgery and Medicine clinical rotations at Baylor College of Medicine with Honors in both.

University of Texas Southwestern Research Technician, August 2008-August 2009 – Worked in the laboratory of Dr. Philipp Scherer at UT Southwestern Health Science Center using HPLC coupled to protein electrophoresis for protein analysis of adiponectin isoforms under various parameters.

Austin College Honors Research, Fall 2007-Spring 2008 – Worked in the lab of Dr. Lance Barton using microarray technology to elucidate the effects of the PA28gamma protein on yeast systems.

MD Anderson Cancer Center in Houston, Summer 2007 – Worked in the lab of Dr. Hector Martinez-Valdez MD/PhD to create bacterial clones expressing fragments of the AKNA protein for use in the production of monoclonal antibodies.

Austin College Welch Research, Summer 2005 – Worked in the lab of Dr. Brad Smucker to synthesize nano-cages for use as solar cell dyes

Honors

The 2013 BCM Anthony R. Means Award for Academic Excellence; the 2013 BCM Trenton Award for Academic Excellence; the TCGA Pan-Cancer project miRNA/non-coding RNA project lead; named a 2013 BCM BRASS Scholar; named the 2009 BCM McNair Scholar; honors in both Surgery and Medicine clinical rotations; induction into the Alpha Chi, Beta Beta Beta, Phi Beta Kappa and Sigma Xi honors societies; the 2007 AC S. D. Heard Fellowship in English (best student in English); the 2007 AC Dr. and Mrs. J. C. Erwin Fellowship in Pre-Medical Studies (best student in pre-medical studies); Boy Scouts of America, Eagle Scout Rank.

CRISTIAN COARFA - BIOGRAPHICAL SKETCH

NAME Coarfa, Cristian	POSITION TITLE Assistant Professor

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Rice University	Ph.D	2007	Computer Science
Rice University	M.S.	2004	Computer Science
POLITEHNICA University, Bucharest, Romania	B.S.	1998	Computer Science

A. Personal Statement

My research focuses on achieving biological insight via integrative analysis, interpretation, and visualization of genomic, transcriptomic, epigenomic, and proteomic assays, and enabling of scientific advancements via collaborative tools and environments. I have extensive experience with ChIP-Seq, RNA-Seq, Small RNA-Seq, and Bisulfite-Seq as part of the NIH Epigenomics Roadmap Data Analysis and Coordination Center at Baylor College of Medicine, where I successfully lead the processing of more than 3100 experiments and over 15TBp of sequencing data. I developed methods for integration of genetic and epigenetic variation, reference pipelines for epigenetic assays, and bioinformatic tools for high-throughput reads mapping and structural variants detection. In particular, I am involved in integrative analysis of structural variation and epigenomic variation, which yielded new insights into epigenomic regulation across prostate and breast cancer. I developed tools and engaged in fruitful collaboration in the area of aberrant RNA-level and DNA-level gene fusions. I co-authored over 40 peer-reviewed articles in the fields of high-performance computing and computational biology, and I demonstrated a record of successful integrative 'omics research, with applications to cancer studies, and in the deployment of cloud-powered analysis toolsets and pipelines. I am enthusiastic about participating in the proposed application, and firmly believe that it will lead to significant and impactful scientific advances.

B. Positions and Honors

Positions and Employment

2006-2007 Research Associate, Baylor College of Medicine, Houston, TX
2007-2008 Postdoctoral Research Associate, Baylor College of Medicine, Houston, TX
2008-2012 Assistant Research Professor, Baylor College of Medicine, Houston, TX
2012-present Assistant Professor, Baylor College of Medicine, Houston, TX

Honors

2005 Best Paper Award at the 2004 International Workshop on High Performance Computing in Medicine and Biology (ICPADS-HiPCoMB05), Fukuoka, Japan
1999 Rice University Fellowship
1998 First Graduate of 1998 class of 140 graduates in Computer Science

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Integrated DNA and RNA somatic mutation discovery and characterization

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Matthew D. Wilkerson, PhD
University of North Carolina at Chapel Hill, Department of Genetics, Lineberger Comprehensive Cancer Center
TCGA Analysis Working Groups: Lung squamous, lung adenocarcinoma, head and neck squamous, prostate.

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Patrick K. Kimes	NA

Background and preliminary data

Identifying somatic mutations with high sensitivity and high specificity is critical for cancer genome characterization and for prioritizing patient treatment. DNA sequencing (whole exome: DNA-WES or whole genome: DNA-WGS) is the standard in cancer mutation profiling. However, DNA sequencing can produce uneven coverage and low sensitivity in low purity or high-clonality tumors. RNA sequencing (RNA-seq) covers the expressed exome, with depth proportional to expression, offering a potential boost in statistical power. **Our core hypothesis is that adding RNA-seq to DNA-WGS will boost mutation detection sensitivity and deepen cancer genome characterization.**

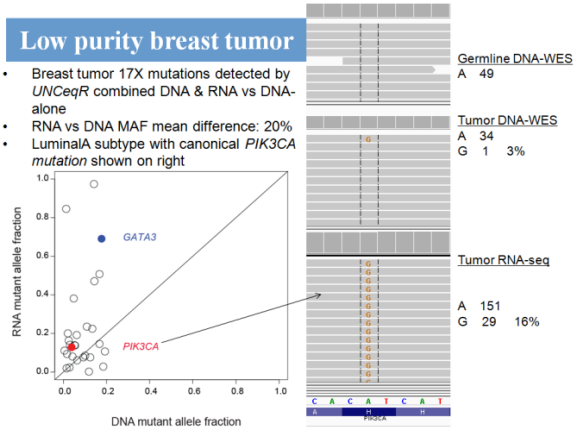
Specifically, we aim to:

1. To determine if combining RNA-seq and DNA-WGS increases mutation detection sensitivity and specificity versus DNA-WGS alone.

2. To determine if low purity tumors have greater mutation signal in RNA-seq versus DNA-WGS

3. To predict and characterize somatic driver genes by total expression, mutation-specific expression, and isoform usage.

In earlier work, we developed a novel methodology and software, called *UNCeqR* (manuscript under review), that detects somatic mutations by combining RNA-seq with DNA-WES using meta-analysis statistics. In application to over 1,000 TCGA breast and lung tumors and simulated datasets, we demonstrated and validated that the addition of RNA-seq boosts mutation detection sensitivity and maintains specificity as DNA-only methods. Mutations detected by our RNA and DNA combined method often had low DNA mutant allele fraction but with a corroborating, higher mutant allele fraction in RNA. Strikingly, low purity tumors (measured by DNA microarrays) displayed the largest gains in new mutations after adding RNA-seq and also had the largest excess of mutation signal in RNA versus DNA (See figure for example of one low purity breast tumor). Our data suggested that rare cancer cells in a low purity tumor can express very high levels of mutant transcript, increasing the RNA mutation signal relative the DNA mutation signal which is diluted by non-cancerous cells. Compared to prior TCGA mutations based on DNA alone, our DNA and RNA integrated method yielded large increases of coding mutations in driver genes, which will lead to deeper and more accurate integrated analyses because the mutation positive and mutation negative groups are better defined (i.e. *UNCeqR* yielded a 36% increase in coding



GATA3 mutations in breast cancer and a 26% increase in coding *CDKNA* lung cancer mutations relative to prior TCGA DNA-only predictions). The ICGC WGS pan-cancer analysis study is an excellent opportunity for our research proposal because it provides a large cohort of patient-matched DNA-WGS and RNA-seq and a broad sampling of different tumor types. In contrast to earlier work using DNA-WES, the availability of DNA-WGS will enable new mutation discoveries and potentially recurrent mutations in non-protein-coding genomic regions.

Timelines & resources dedicated to project

Timeline milestones:

1. Somatic mutation detections of TCGA-specific-sub-cohort
    a. Input: tumor and germline DNA-WGS BAMs and RNA-seq BAM
    b. Output: DNA and RNA integrated VCF and MAF
2. Tumor and mutation analysis of TCGA-specific-sub-cohort
3. Somatic mutation detections of ICGC-specific-sub-cohort
    a. Input and output same as #1.
4. Tumor and mutation analysis of ICGC-specific-sub-cohort mutations
5. Final Report

Dependencies. We have access to the TCGA-sub-cohort BAMs as the RNA sequencing was completed at UNC and we can obtain DNA-WGS via our TCGA credentials. For milestones #3 and #4, we are dependent on ICGC BAM file access.

Resources: Dr. Wilkerson has greater than 2 petabytes of high performance disk storage that is accessible from 3 high performance compute clusters (500-1000 cores each), which has been used for this type of computation previously and is sufficient for the proposed study. Access to ICGC computing infrastructure with local BAM storage, if available, would eliminate file transfers and accelerate completion of the proposed research.

Research proposal

Aim 1. To determine if combining RNA-seq with DNA-WGS increases mutation detection sensitivity and specificity versus DNA-WGS alone. Our previously-validated algorithm, *UNCeqR*, will detect and rank somatic mutations in patient-matched DNA-WGS and RNA-seq. The combined results will be compared to DNA-WGS only based results (from *UNCeqR* and other methods) using gold standards of orthogonal sequencing, replicate sequencing, and sequencing with simulated mutations. These results will form the basis of Aim 2 and 3 and will be shared with all ICGC groups for integrated analysis.

Aim 2. 2. To determine if low purity tumors have greater mutation signal in RNA-seq versus DNA-WGS. We will test for associations between tumor purity with tumor mutation properties (total count with RNA-seq and DNA-WGS versus DNA-WGS only; and mutation signal RNA vs DNA). We will use multiple ratings of tumor purity: DNA-sequencing based, DNA-copy number based and light-microscopy based. We hypothesize that RNA mutation signal excess (greater signal in RNA versus DNA) will be a characteristic of select cancer types and low purity tumors, leading to deeper and more accurate characterization.

Aim 3. To predict and characterize somatic driver genes by total expression, mutation-specific expression, and isoform usage. Current driver gene appraisal is typically based on DNA mutation recurrence and gene properties. Here, we will evaluate new variables to predict driver genes: differential total expression in mutants versus non-mutants and mutation-specific expression, via regression modeling and training gene sets. Recurrent gene mutations will be evaluated for association with cumulative transcript isoform usage using our tool, *SigFuge* (under review). Our approaches provide RNA empirical evidence for evaluating non-coding mutations, as well as coding.

Legacy plans

Intermediate results will be deposited into a secure ICGC repository during the study. Final results, analytical scripts and documentation will be deposited. *UNCeqR* is free for non-profit use.

Matthew D. Wilkerson, Ph.D.

Research Assistant Professor of Genetics
Department of Genetics
Lineberger Comprehensive Cancer Center
University of North Carolina at Chapel Hill
450 West Drive, Campus Box #7295, Chapel Hill, NC 27559, 919-843-5763
mwilkers@med.unc.edu

Education and Training

Postdoctoral Fellow in Cancer Genomics		2008-2012
University of North Carolina at Chapel Hill	Chapel Hill, NC	
Ph.D. in Bioinformatics and Computational Biology		2007
Iowa State University	Ames, IA	
B.S. in Biological Sciences		2003
University of Notre Dame	Notre Dame, IN	

Honors and Awards

Ruth L. Kirschstein National Research Service Award Fellowship (F32 NRSA), National Cancer Institute, PI: Matthew D. Wilkerson, $153,810, 2009-2012

Future of Science Fund Scholarship, Keystone Symposia, Changing Landscape of Cancer Genome, 2011

Clinical/Translational Second Place Poster Clinical/Translational Second Place Poster, 2011

Premium for Academic Excellence, Iowa State University, 2003

Dean's List, University of Notre Dame, 2001, 2002, 2003

Reviewer for: *Journal of Clinical Investigation, Clinical Cancer Research,* Italian Association for Cancer Research, *Bioinformatics, Plos One, Nucleic Acids Research, BMC Genomics*

Selected Roles in The Cancer Genome Atlas (TCGA) analysis working groups

Lung squamous cell carcinoma: Data Coordinator, Writing Committee, RNA Analysis Leader
Lung adenocarcinoma: Writing Committee, RNA-seq Analysis Leader, Fusion transcript analysis
Head neck squamous cell carcinoma: Writing Committee, Virus analysis, RNA-seq analysis
Prostate cancer: RNA-seq analysis, Fusion transcript analysis subgroup leader, RNA mutation discovery

Selected Publications (8 of 30, full list at http://www.unc.edu/~mwilkers)

- The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM). (**Wilkerson MD** 1 of 140 authors) A Genomics-Based Classification of Human Lung Tumors. *Science Translational Medicine*. Vol. 5, Issue 209, p. 209ra153
- The Cancer Genome Atlas Research Network. (**Wilkerson MD** Data Coordinator, Writing Committee, RNA Analysis Leader) (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012 Sep 9.
- Cabanski CR, **Wilkerson MD**, et al. (2013) BlackOPs: increasing confidence in variant detection through mappability filtering. *Nucleic Acids Research*. Oct;41(19):e178.
- **Wilkerson MD**, et al. (2013) Prediction of Lung Cancer Histological Types by RT-qPCR Gene Expression in FFPE Specimens. *The Journal of Molecular Diagnostics*. Jul;15(4):485-97
- **Wilkerson MD**, et al. (2012) Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability and methylation. *PLoS One*. 7(5): e36530.
- **Wilkerson MD**, et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical Cancer Research*. Oct 1;16(19):4864-75.
- Veerhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, **Wilkerson MD**, et al. (2010) Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. Volume 17, Issue 1, 98-110.
- **Wilkerson MD**, Hayes DN. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, Jun 15;26(12):1572-3

Patrick Kosuke Kimes

Contact Information

Statistics and Operations Research
Hanes Hall B01, CB #3260
University of North Carolina at Chapel Hill
Chapel Hill, NC 27566

Education

The University of North Carolina (expected) **May 2015**
Chapel Hill, North Carolina
Ph.D. – Statistics (Candidate)
Certificate – Bioinformatics/Computational Biology
Advisors: Yufeng Liu, J. S. Marron, D. Neil Hayes

Pomona College **May 2009**
Claremont, California
B.A. – Mathematics, GPA: 3.77

Honors and Awards

Travel Grant, *UW Summer Institute in Statistical Genetics* **July 2012**
- stipend and fee waiver to attend short courses in Seattle, WA

Hoeffding Award, *UNC Statistics* **December 2011**
- highest score on statistics Ph.D. comprehensive written exams

Senior Service Award, *Pomona College* **May 2009**
- significant contributions to campus life through leadership

Pomona College Scholar, *Pomona College* **Spring 2006 to Fall 2007**

Summer Research Grant (SURP), *Pomona College* **Summer 2007**
- stipend to study rank-based nonparametric approaches to differential gene analysis

Papers

1 Kimes PK, Cabanski CR, Wilkerson MD, Johnson AR, Makowski L, Maher CA, Liu Y, Perou CM, Marron JS, and Hayes DN. *SigFuge: single gene unsupervised clustering of RNA-seq data reveals heterogeneity among cancer samples*, **submitted**.
- accompanying Bioconductor package: `SigFuge`

2 TCGA Research Network (one of over 300 contributors), *Comprehensive genomic characterization of head and neck squamous cell carcinomas*, **submitted**.

3 TCGA Research Network (one of over 300 contributors), *Diversity of lung adenocarcinoma revealed by integrative molecular profiling*, **submitted**.

Posters

1 *SigFuge: unsupervised discovery in RNA-seq data*, Lineberger Comprehensive Cancer Center Postdoc/Faculty Research Day (Chapel Hill, NC), September 2013.

2 *Adaptive Nonparametric Tests for the Two-Sample Location Model with Applications to Microarray Data*, Pomona College Summer Research Poster Conference (Claremont, CA), September 2007.

Technical Skills

Computing:
- *Proficient:* MATLAB, `R`/`Bioconductor`, LaTeX
- *Familiar:* C++ (STL, GSL), Emacs, PERL, Bash, Git

Languages:
- *Native:* English, Japanese

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Characterization of DNA copy number variation in HLA region in the human genome

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators

(Name no more than 2; append 1 page CV for each)

Li Zhang

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)

Background and preliminary data

The leukocyte antigen (HLA) region in the human genome encodes for major histocompatibility complex (MHC). HLA region is located in 6p21 containing several sub regions called HLA-A, HLA-B, HLA-C, HLA-DR, HLA-DP, and HLA-DQ. These regions are hyper variable in terms of copy number, and fast evolving, compared with the rest of the genome. MHC plays an important role in the immune system. Mutations and DNA copy number variations (CNV) in the regions have been linked to autoimmune diseases and cancer. Using data collected in the digital database of human genome variation (DGV; URL: http://dgv.tcag.ca), we performed cluster analysis of CNV segments in order to identify their common modes in the HLA region. DGV currently included 55 existing studies of germline variation of copy number variation (CNV). The original data were generated from microarrays and lately next generation sequencing technologies.

Our results showed that CNVs obtained from microarrays are mostly in agreement, while the results from the next generation sequencing, particularly those from 1000 genome project, were drastically fewer. For CNVs that span more than 1000 base pairs in the genome, we found that of the 6 sub regions of the HLA systems, only one of them appears to contain common CNVs collected from 1000 genome project, while all of the 6 regions were recognized to have common CNVs from microarray studies.

For CNVs that span fewer than 1000 base pairs, we found that the distribution of CNV segment length has a prominent peak around 340 base pairs. Because 340 bp is too short to be detected in the microarrays, there were such reported finding from microarray studies. However, several studies using sequencing technologies have the same peak at this segment length. The mechanism of such CNVs is not yet known, and we are not even sure that they are real, and not artefacts of sequencing.

Furthermore, we analyzed CNVs using data collected from TCGA project. The samples were collected from histologically normal cells adjacent to tumors. We compared the distributions of CNV in TCGA normal samples with that collected in DGV in the HLA region. We found that there are common alteration regions in TCGA data that do not correspond to common regions in DGV. Such regions may be related to cancer.

Note: The PI of this proposed project is partially funded by GDAC in TCGA.

Timelines & resources dedicated to project

1. Jan-2014 to May- 2014: develop algorithms for data processing to generate the CNVs from the whole genome sequencing data.
2. May-2014-July 2014: Aim 1 and Aim 2.
3. July-2014 – Sept 2041: Aim 3.

Research proposal

We propose to investigate the CNVs in the HLA region in the 2K genomes. We aim to characterize the common modes of CNVs in HLA at nucleotide resolution and compare with that identified from TCGA data. The results will help us to better define germline CNVs and how they may be related to cancer.

Specifically, we plan to:

(1) Determine the distribution of CNVs that span more than 1000 bp in the HLA regions.

(2) Determine the distribution of CNVs that span less than 1000 bp but greater than 50 bp in the HLA regions.

(3) Compare the distribution of the resulted >1000 bp CNVs with that obtained previous analysis of normal samples of microarray data in the TCGA project and determine the features that are cancer specific.

The HLA region is one of the most intensively studied in the human genome. The common variable regions are already known. However, the current knowledge is not precise at the nucleotide resolution. Our previous analysis of 1000 genome project data showed that the sequencing results are actually quite different from microarray studies. Thus, the discrepancy is major challenge before we can make further progress using sequencing for CNV analysis. We hope that the 2k genome project will provide better data to resolve the issue. We will test several existing methods for estimating copy number changes and find out which method will generate results that are consistent with previous research.

For short (<1000 bp) CNVs, we plan to find out if the CNVs with the characteristic length of 340 bp identified from 1000 Genome Projects can be reproduced in the 2k Genome Project. If they can be reproduced, we will conclude that are less likely to be caused by sequencing artefacts. We hypothesized that the breakpoints of such segments may be related to nucleosome binding. We plan to data from ENCODE project if the distribution of breakpoints of the 340bp CNVs are related to the regulatory sites identified in ENCODE project.

Finally, by comparing the 2k Genome Project results with that in TCGA normal sample data, we will be able to identify the features that are specific to the seemingly normal cells adjacent to the tumors. Such feature may be important in the carcinogenesis.

CURRICULUM VITAE

Li Zhang, Ph.D.

Title and Affiliation: Associate Professor, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX.

Eduction: PHD, University of North Carolina At Chapel Hill, Chapel Hill. 1995.

Research interests and Expertise:

1. Microarray data analysis
2. Biomarker research
3. Bioinformatic algorithm development

Selected Publications (from 64 peer review articles):

1. Zhang L, Miles MF, Aldape KD. A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol 21(7):818-21, 7/2003. e-Pub 6/2003.
2. Zhang L, Wu C, Carta R, Baggerly K, Coombes KR. Response to Preprocessing of oligonucleotide array data. Nat Biotechnol 22(6):658, 6/2004.
3. Wu C, Carta R, Zhang L. Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Res 33(9):e84, 2005. PMCID: PMC1140085.
4. Zhang L, Wei Q, Mao L, Liu W, Mills GB, Coombes K. Serial dilution curve: a new method for analysis of reverse phase protein array data. Bioinformatics 25(5):650-4, 3/2009. e-Pub 1/2009. PMCID: PMC2647837.
5. Saintigny P*, Zhang L*, Fan Y-H, El-Naggar AK, Papadimitrakopoulou V, Feng L, lee JJ, Kim ES, Hong W-K, Scott M, Lippman SM, Mao L. Gene expression profiling predicts the development of oral cancer. Cancer Prevention Research. 2011. 4(2):218-29. (* Co-first authors)
6. Qiu P, Zhang L. Identification of markers associated with global changes in DNA methylation regulation in cancers. BMC Bioinformatics. 2012, 13 (Suppl 13):S7.
7. Su X, Zhang L, Zhang J, Meric-Bernstam F, Weinstein JN. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics. 2012 Sep 1;28(17):2265-6. Epub 2012 Jun 28.
8. Zhang L, Mitani Y, Caulin C, Rao P-H, Kies MS, Saintigny P, Zhang N, Weber RS, Lippman SM, and El-Naggar AK. Detailed genome wide SNP analysis of major salivary carcinomas localizes subtype specific chromosomal sites and oncogenes of potential clinical significance. J. Am. Pathology. Am J Pathol. 2013 Jun;182(6):2048-57.
9. Zhang L, Zhang L. Use of autocorrelation scanning in DNA copy number analysis. Bioinformatics. 2013 Nov 1; 29(21):2678-82.

.

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings Jennifer.Jennings@oicr.on.ca by 27th November, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Non-PCR related reads duplication and adjustment in NGS data

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Jianhua Zhang and Lynda Chin, Department of Genomic Medicine, M. D. Anderson Cancer Center, The University of Texas, Houston, 77054 (Affiliated with TCGA)

Name(s) & institute(s) of junior investigators
(Name no more than 2; append 1 page CV for each)

Name(s) & institute(s) of non-ICGC collaborators
(Name no more than 2; append 1 page CV for each)

Background and preliminary data

High throughput sequencing has been the main platform employed by ICGC and TCGA to study cancer genomics. One issue we have observed while working on sequencing data is the existence of non-PCR related read duplications in BAM files of various sequencing depths and particularly in whole exome sequencing data. Reasons for read duplications may vary between experiments but the physical limitation of unique DNA fragments that can be sequenced is a definite factor (Figure 1a). In a simulation study, we observed that read duplications may occur at a mean sequencing depth as low as 1 X (Figure 1b) with an average occurrence at depths around 10 X (Figure 1c). In fact, the number of unique starting locations of sequenced reads is determined by the sequencing length (L) following the formula 2 * L + 1, where 2 accounts for both strands. For example, in a sequencing experiment of 75 bp the number of unique starting locations is 151, meaning when the sequencing depth is greater than 151, reads will duplicate regardless. Practically, reads begin to duplicate at much lower depths. In a whole exome sequencing study, we observed that it requires deeper sequencing to reach the point when all the positions within a 151 pb window are covered by at least two reads than the depth at which all the sequenced reads are duplicated (Figure 2a, b), indicating that reads accumulate at some locations more often than the others. The fact that read duplications do not occur randomly suggests that read duplications may affect the accuracy of variant calling, tumor purity estimations, and tumor clonal structure inferences. In an extreme example, we observed that a mutation call was based on reads accumulating at the same starting location. After removing the duplicating reads, there was not enough evidence to call the mutation. On the other hand, if read duplications occur more often in non-variant bearing reads, a variant may not be called or the allele frequency may be calculated wrong. In another study, we noticed that the total number of unique non-variant bearing reads saturates when sequencing depth increases while the number of unique variant bearing reads rarely reaches the point of saturation over a wide range of depths. Removing the duplicating reads may thus increase the sensitivity of detecting low frequency variants.

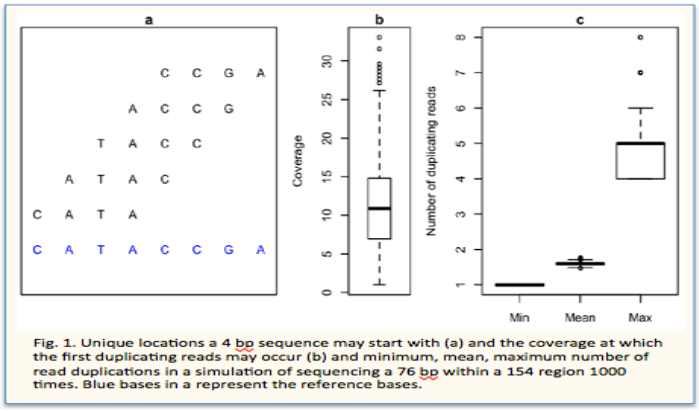


Fig. 1. Unique locations a 4 bp sequence may start with (a) and the coverage at which the first duplicating reads may occur (b) and minimum, mean, maximum number of read duplications in a simulation of sequencing a 76 bp within a 154 region 1000 times. Blue bases in a represent the reference bases.

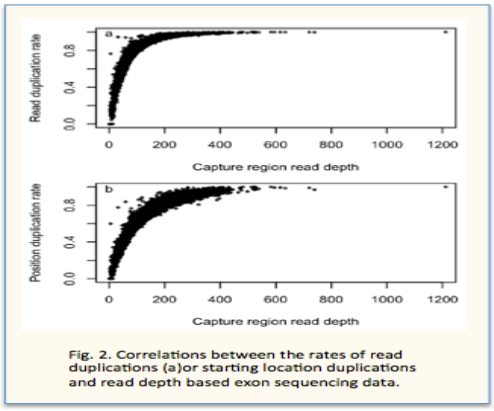


Fig. 2. Correlations between the rates of read duplications (a)or starting location duplications and read depth based exon sequencing data.

ICGC/TCGA has both whole genome and exome sequencing data of various sequencing depths, which provide an excellent platform for studying the issues related to non-PCR read duplications in sequencing data. Here we propose

a research project to investigate whether removing duplicating reads would improve the accuracy and sensitivity of variant detection. We plan to address the questions by a) removing duplicating reads from BAM files before making variant calls and b) calculating the adjusted base counts and allele frequency for variant calls derived from duplicates containing BAM files, and c) comparing results from a and b plus the variant calls using duplicates containing BAM files to examine the effects of read duplications on downstream analyses. The results also provide other ICGC/TCGA working groups with additional information about the genomes.

Timelines & resources dedicated to project

1. Pre-process BAM files to remove the duplicating reads and call variants (Dec. 2013 to Sept. 2014. 0.5 FT will be allocated to the task). An existing JAVA module developed in house will be used. It takes 4 – 5 hours to remove the duplicating reads from a BAM file of 20 GB on one computing core.

2. Calculate the adjusted base counts and allele frequency after removing the duplicating reads for all the variants called using duplicates containing BAM files (Dec. 2013 to Sept. 2014. 0.5 FT will be allocated to the task). An existing R package developed in house will be used. It takes 5 – 6 hours to calculate the adjusted base counts and allele frequency for about 1,000,000 variants on one computing core.

3. Examine 100 BAM files for the extend and distribution of read duplications (Sept. 2013 – May 2014. 0.5 FT will be allocated)

4. Compare the results from 1 and 2 plus variant calls otherwise made (Sept. 2014 – March. 2015. 1 FT will be allocated)

Research proposal

Pre-process BAM files and make SNV/indel calls

All the whole genome and whole exome sequencing BAM files released by ICGC/TCGA will be processed using a JAVA module we developed to remove the duplicating reads (reads with the same starting location, same sequence, and on the same strand). The pre-processed BAM files will be subjected to variant detection using the same algorithm and parameters used by ICGC/TCGA working group to detect variants without the duplicating reads removal. Variants called after the duplicating reads have been removed will be compared with those otherwise called using duplicates containing BAM files to determine whether removing the duplicating reads would improve variant calls. Variant calls will also be released to ICGC/TCGA working group to provide them with additional information about the variants of the genomes.

Post-call read duplication removal/adjustment

Variant calls released to the working groups by ICGC/TCGA using duplicates containing BMA files will be extracted and the adjusted base counts and allele frequency will be calculated after removing the duplicating reads from BAM files using an R package we developed. The results will be compared with the original calls to determine whether the adjusted values would add to the understanding of the called variants or can be used as filters. The adjusted values will also be released to ICGC/TCGA working groups to provide additional information about the Variants called.

Read duplication assessment

A subset of 50 bam files will be randomly selected from the whole genome and whole exome sequencing platforms, respectively, and duplicating reads and reads that do not have duplicates will be extracted from the BAM files. Their distribution along the genome and correlations with GC contents of the region, sequencing depth, copy number data, and repeats will be examined. A JAVA module has already been developed to extract reads in different groups.

In all the above comparisons, programmatic and manual inspections of a subset of variants using IGV will be conducted.

Legacy plans

An R package that generates adjusted base counts and allele frequency for each variant post variant calls and a JAVA module that produces bam files with duplicating reads removed will be made available to the public.

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Jianhua Zhang	Associate Director - Bioinformatics
eRA COMMONS USER NAME (credential, e.g., agency login)	Institute for Applied Cancer Science M. D. Anderson Cancer Center

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Yunnan University, Kunming, Yunnan, P. R. China	B. Sc.	1982	Botany
University of St. Thomas, St. Paul, Minnesota, USA	M. Sc.	2000	Software engineer
University of Western Ontario, London, Ont., Canada	Ph. D.	1991	Botany
Department of Biology, McGill University, Montreal, Canada	Post-Doctor	1993	Biology

Positions and Honors.

Nov. 2011 -	Associate Director – Bioinformatics, M. D. Anderson Cancer Center
Nov. 2001 – Oct. 2011	Research Scientist/Group Leader, Dana-Farber Cancer Institute
March 1998 – Oct. 2001	Senior Research Associate, University of Minnesota
May 1993 – Feb. 1998	Research Scientist, Agriculture and Agri-Food Canada
Aug. 1991 – April 1993	Post-doctoral Fellow, McGill University

Selected peer-reviewed publications (in chronological order).

The CANCER GENOME ATLAS NETWORK. 2013. Integrated Genomic Characterization of Endometrial Carcinoma. Nature 497, 67-73

HU J, HO AL, YUAN L, HU B, HUA S, HWANG SS, ZHANG J, HU T, ZHENG H, GAN B, WU G, WANG YA, CHIN L, DEPINHO RA. 2013. Neutralization of terminal differentiation in gliomagenesis. PNAS 110:14520-14527.

GENOVESE G, ERGUN A, SHUKLA SA, CAMPOS B, HANNA J, GHOSH P, QUAYLE SN, RAI K, COLLA S, YING H, WU CJ, SARKAR S, XIAO Y, ZHANG J, ZHANG H, KWONG L, DUNN K, WIEDEMEYER WR, BRENNAN C, ZHENG H, RIMM DL, COLLINS JJ, CHIN L. 2012. microRNA Regulatory Network Inference Identifies miR-34a as a Novel Regulator of TGF-β Signaling in Glioblastoma. Cancer Discovery. 2(8) 736 – 74 9

LARMAN TC, DEPALMA SR, HADJIPANAYIS AG, THE CANCER GENOME ATLAS RESEARCH NETWORK, PROTOPOPOV A, ZHANG J, GABERIEL SB, CHIN L, SEIDMAN C, KUCHERLAPATI R, SEIDMAN JG. 2012. Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci USA 109: 14087 - 14091.

DING Z, WU CJ, JASKELLOFF M, IVANOVA E, KOST-ALLMOVA M, PROTOPOPOV A, CHU GC, WANG G, LU X, WANG W, XIAO Y, ZHANG H, ZHANG J, ZHANG J, GAN B, PREEY SR, JIANG S, LI L, HORNER JW, WANG YA, CHIN L, DEPINHO RA. 2012. Telomerase Reactivation following Telomere Dysfunction Yields Murine Prostate Tumors with Bone Metastases. Cell, 148: 896 - 907.

XI R, HADJIPANAYIS AG, LUQUETTE LJ, KIM TM, LEE E, ZHANG J, JOHNSON MD, MUZNY DM, WHEELER DA, GIBBS RA, KUCHERLAPATI R, PARK PJ. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci USA, 108:E1128-36,

THE CANCER GENOME ATLAS NETWORK. 2011. Integrated genomic analyses of ovarian carcinoma. Nature 474: 609 – 615.

COLLABORATORS/KEY PERSONNEL - BIOGRAPHICAL SKETCH

Provide the following information for all Collaborators/Key Personnel.
(there is no page limit)

NAME	**POSITION TITLE**
Lynda Chin, MD	Professor and Chair, Dept of Genomic Medicine Scientific Director, Institute for Applied Cancer Science

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Brown University, Providence, RI	BS	09/84-06/88	Neuroscience
Albert Einstein College of Medicine, Bronx, NY	MD	09/89-06/93	Medicine
Columbia Presbyterian Medical Center, NY, NY	Internship	07/93-06/94	Internal Medicine
Albert Einstein College of Medicine, Bronx, NY	Residency	07/94-06/97	Dermatology
Albert Einstein College of Medicine, Bronx, NY	Postdoctoral	07/93-06/97	Molecular Genetics

A. Positions and Honors.

Positions

1996 – 1997	Chief Resident, Dermatology, Albert Einstein College of Medicine (AECOM), NY
1998 – 2004	Assistant Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute (DFCI), Boston, MA
1999 – 2004	Scientific Director, Arthur & Rochelle Belfer Cancer Genomics Center, DFCI, Boston, MA
2005 – 2009	Associate Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2008 –	Member, scientific steering committee, International Cancer Genome Consortium (ICGC).
2009 – 2011	Professor, Dept of Dermatology, Harvard Medical School and Dept of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
2009 –	Associate Member, the Broad Institute of MIT and Harvard, Boston, MA
2009 – 2011	Co-director, Melanoma Program, Dana-Farber/Harvard Cancer Center, Boston, MA
2009 – 2011	Scientific Director, the Belfer Institute for Applied Cancer Science, DFCI, Boston, MA
2009 –	Member, Executive Subcommittee, The Cancer Genome Atlas (TCGA), USA
2011 –	Professor and Chair, Department of Genomic Medicine, UTMDACC, Houston, TX
2012	Elected, Institute of Medicine of National Academies (IOM)

B. Selected peer-reviewed publications (in chronological order).

1. Ding Z, …, Chin L, Depinho RA. Telomerase Reactivation following Telomere Dysfunction Yields Murine Prostate Tumors with Bone Metastases. ***Cell***. 2012 Mar 2;148(5):896-907. Epub 2012 Feb 16. PMID: 22341455
2. The Cancer Genome Atlas Research Network, et al. Comprehensive genomic characterization of squamous cell lung cancers. ***Nature***. e-Pub 9/2012. PMC3466113
3. Kwong LN, …, Chin L. Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma. ***Nat Med*** 18(10):1503-10, 9/2012. e-Pub 9/2012. PMID:22983396
4. Hodis E, …, Chin L. A landscape of driver mutations in melanoma. ***Cell*** 150(2):251-63, 7/2012. PMID: 22817889
5. Berger MF, …, Chin L, Garraway LA. Melanoma genome sequencing reveals frequent PREX2 mutations. ***Nature*** 485(7399):502-506, 5/2012. PMC3367798

Abstract of proposed research for WGS pan-cancer analysis

Please submit to Jennifer Jennings **Jennifer.Jennings@oicr.on.ca** by ~~27th November~~ 31st December, 2013 (5pm your local time). Explanatory notes follow the form.

Title of abstract

Investigation of how germline variation informs somatic mutation profiles and its effect on both cancer risk and outcomes

Name(s), institute(s) & ICGC/TCGA affiliation(s) of principal investigators
(Name no more than 2; append 1 page CV for each)

Bin Zhu, Nilanjan Chatterjee, Stephen Chanock – National Cancer Institute

Name(s) & institute(s) of junior investigators (Name no more than 2; append 1 page CV for each)	**Name(s) & institute(s) of non-ICGC collaborators** (Name no more than 2; append 1 page CV for each)
Simina Boca – National Cancer Institute	

Background and preliminary data

Over last several years, analysis of somatic mutation data generated from TCGA and ICGC studies has provided a comprehensive catalog of mutation signatures and underlying driver genes for over two dozen cancers. In parallel, hundreds of germline variants which affect cancer risk have been discovered, consisting of highly penetrant genes identified mainly in family studies, and, more recently, common susceptibility loci with small effects identified from genome-wide association studies, only a small fraction of which are shared by multiple tumor types. The contribution of less common and rare germline variants, possibly with stronger effect sizes, remains to be discovered with next generation sequencing technologies. The spectrum of susceptibility germline variants discovered thus far points towards disease specific differences in the distribution of the number and effects of variants of different minor allele frequencies. As the field continues to discover more susceptibility alleles, it is critical to investigate how germline variation can inform our understanding of somatic alterations, both with respect to the complex process of cancer pathogenesis and to the effect of germline variation (not necessarily the drivers of carcinogenesis) on cancer outcomes.

Under the leadership of Dr. Stephen Chanock, the Division of Cancer Epidemiology and Genetics (DCEG) of NCI has generated data from genome-wide association studies for more than 70,000 cancer cases and 30,000 controls, covering over a dozen tumor types. These studies have contributed substantively to the discovery of a majority of known genome-wide association studies loci. Dr. Nilanjan Chatterjee, the Chief of the Biostatistics Branch (BB) of DCEG, has been the lead statistician in many of these studies and has extensive experience in modeling cancer risks associated with germline variations. Dr. Chatterjee, together with Dr. Bin Zhu, a Tenure-Track Investigator in BB, has recently initiated a methodological study for developing a rigorous statistical framework for testing driver gene hypothesis taking into account patient level characteristics such as germline mutation status, non-genetic risk-factors and histopathological characteristics of tumor. Drs. Chatterjee and Zhu are currently testing these methodologies in simulation studies and in preliminary analyses of existing TCGA datasets. The project is being assisted by Dr. Simina Boca, a post-doctoral Fellow in BB, who did her PhD dissertation at Johns Hopkins University under the mentorship of Dr. Giovanni Parmigiani, her prior work including the development of tools for somatic mutation data analysis. Preliminary results indicate that methods including use of optimal test-statistics and a proper resampling algorithm for generating the null distribution, can improve the power to detect driver genes in comparison to standard algorithms such as MutSig, even in the absence of any additional modeling of patient-level characteristics. Further development and testing of the methodologies are underway.

Timelines & resources dedicated to project

Timelines: The methodological development has been started and will be finalized and applied to the existing TCGA datasets by August 2014. From September 2014 to December 2014, comprehensive analysis will be

applied to core variant calls. Manuscripts will be submitted in March 2015, following board timelines.

Resources: This study will utilize the cloud-based computing centers which will be made available. We will require the core variant calls starting in September 2014. Additional computational and bioinformatics resources will be obtained through existing resources at the National Institutes of Health.

Research proposal

We aim to analyze the large pan-cancer datasets generated by ICGC/TCGA studies to explore links between germline risk variants and somatic mutations. Below we describe a proposal for a set of specific analyses, focused on discovery of common and rare germline susceptibility variants and how they inform our understanding of critical somatic alterations as well as cancer outcomes.

1. Are there undiscovered associations between less common and rare germline variants and cancer risk?
We will first consider the associations between less common and rare germline variants and cancer risk, using the ICGC/TCGA studies. This analysis will utilize statistical methods which test individual variants, as well as methods which test variants in a genomic region, such as CAST, C-alpha, or SKAT, to find high or moderate penetrance alleles. Additionally, a subset-based meta-analysis approach (ASSET), developed by Dr. Chatterjee's group, will be applied to search for variants which are possibly associated with more than one cancer type. It is important to use these additional approaches in a complementary manner to individual variant testing due to the fact that the ICGC/TCGA studies consist of sample size in the order of tens or hundreds, which will yield low power for finding associations with germline variants that have relatively small effect sizes. For rarer variants with strong effects, the statistical approach will be augmented by incorporation of somatic signatures of the gene/region of interest. The sets above are available for strategic sequencing or genotyping required to confirm new susceptibility alleles.

2. Do germline risk variants and somatic pathogenic mutations tend to aggregate at same genetic loci?
There are well known examples, such as the classic two-hit model for retinoblastoma, in which pathogenic somatic mutations and rare highly penetrant germline variants alter the same genetic loci with comparable functional consequences. The links between loci containing common or rare susceptibility variants and somatic pathogenic mutations have remained relatively unexplored. For each cancer in the pancancer dataset, we will survey the NHGRI database as well as a comprehensive survey of cancer predisposing genes to identify all susceptibility SNPs discovered to date and map underlying or nearby exonic regions. We will then test whether this small subset of exonic regions contains an excess of somatic driver mutations using methods similar to those currently being used for testing enrichment of somatic coding mutations within specific genes. We will perform such tests "globally," i.e. at the level of the whole gene-set, but also locally at the level of each gene within the subset, looking at both coding and non-coding regions. For the latter, we will use ENCODE and other putative functional predictors for prioritization in the analysis. As we will focus on a relatively small number of genomic regions, the power to identify driver gene within such subsets may be higher than finding driver genes from the much larger collection (~25,000) of all annotated genes in the genome. In an exactly symmetric approach, we will investigate whether there is an enrichment of cancer susceptibility alleles within cancer driver genes that are already identified or to be identified from the continuing ICGC/TCGA studies. We plan to perform this analysis using large genome-wide association studies datasets generated by DCEG together with the large consortia in which DCEG has participated and the remaining datasets available in dbGaP.

3. How do germline risk variants and other patient characteristics interact with somatic pathogenic mutations?
Independent of whether germline risk-variants and somatic pathogenic mutations tend to aggregate at the same genetic loci, fundamentally it is of interest to understand how these factors interact in the multi-step model of carcinogenesis. We will explore possible mechanisms of such interaction using a number of alternative analytic approaches using both germline and somatic mutation data that would be available as part of the ICGC/TCGA studies. First, we will examine differences in the minor allele frequencies of known susceptibility SNPs for each cancer between different case-subtypes that could be defined by somatic mutation signatures in known driver genes for the same cancer. Presence of heterogeneity in the minor allele

frequencies between cancer subtypes could indicate that susceptibility SNPs may pose different risks for different cancer subtypes and the mutation-signature of these subtypes can then provide clues about underlying pathways through which the risk-SNPs may be active. For susceptibility regions now known to have highly pleiotropic effects, such as the TERT-CLPTML region in Chr. 5 or the 8q.24 region, we will perform such analysis across pancancer datasets. In a second approach, we plan to model explicitly the correlation of risk SNPs and rate of somatic mutations within the aforementioned statistical framework for testing drives gene accounting for patient-level characteristics. Using this framework, for example, we will be able to test whether risk SNPs, individually or collectively, can influence the degree of selective advantage driver gene mutations provide to neoplastic cells. If such relationship can be identified, then such results can provide insights into plausible "gene-gene" interactions in multi-step carcinogenesis. Using a similar approach, we plan to also use to understand role of "gene-environment" interaction using ICGC/TCGA studies that have high-quality data on risk-factor information (e.g., smoking history).

Legacy plans

We plan to develop tools in the statistical language R, possibly interfacing with C++ and Python to increase computational speed. Upon extensive testing, we will submit our code to Bioconductor, a popular open-source library for bioinformatics software. We expect several methodological and scientific papers from this project.

Bin Zhu

Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute
9609 Medical Cener Drive, Room 7E618
Bethesda, MD 20892

Phone: (240)-276-7420
Email: bin.zhu@nih.gov

Education

University of Michigan — Ann Arbor, MI, USA
Ph.D. in Biostatistics — 2004 - 2010

Zhejiang University — Hangzhou, ZJ, China
B.SC. in Biological Sciences — 1998 - 2002

Professional Experience

Tenure-Track Principal Investigator — 2012 - Present
Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD

Postdoctoral Associate — 2010 - 2012
Department of Statistical Science and Center for Human Genetics, Duke University, Durham, NC

Publications

1. **Zhu, B.**, Song, P.X.K. and Taylor, J.M.G. (2011) Stochastic Functional Data Analysis: A Diffusion Model-based Approach. *Biometrics* 67, 1295-1304.
   Winner of 2009 **ENAR Distinguished Student Paper Award** sponsored by the *International Biometric Society* and 2009 **ASA Student Paper Competition Award** sponsored by the *Bayesian Statistical Science Section*.

2. **Zhu, B.**, Taylor, J.M.G. and Song, P.X.K. (2011) Semiparametric Stochastic Modeling of the Rate Function in Longitudinal Studies. *Journal of the American Statistical Association* 106, 1485-1495.

3. **Zhu, B.**, Dunson, D.B. and Ashley-Koch, A.E. (2011) Adverse Sub-population Regression for Multivariate Outcomes with High-dimensional Predictors. *Statistics in Medicine* 31, 4102-4113

4. **Zhu, B.**, Ashley-Koch, A.E. and Dunson, D.B. (2013) Generalized Admixture Mapping for Complex Traits. *G3: Genes, Genomes, Genetics 3, 1165-1175.*

5. **Zhu, B.**, and Dunson, D.B. (2013) Locally Adaptive Bayes Nonparametric Regression via Nested Gaussian Processes. *Journal of the American Statistical Association. In Press*.

CURRICULUM VITAE

Nov 2013

Name: Nilanjan Chatterjee
Date and Place of Birth: August 18, 1972; Calcutta, India
Citizenship: USA

Education:

1990	Graduated from High School
1993	Bachelor of Statistics (BStat), Indian Statistical Institute, Calcutta, India
1995	Master of Statistics (Mstat), Indian Statistical Institute, Calcutta, India
1999	PhD (Statistics), University of Washington, Seattle

Brief Chronology of Employment:

1995-1996	Teaching Assistant, Department of Statistics, University of Washington, Seattle
1996-1999	Research Assistant, Department of Biostatistics, University of Washington, Seattle
1999-2001	Research Fellow, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health
2001-2004	Principal Investigator (Tenure Track), Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH
2004-Pres.	Senior Investigator (Tenured), Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH
2008-Pres.	Chief, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH

Selected publications (from more than 200 publications)

1.**Chatterjee N**, Wheeler B, Sampson S, Hartge P, Chanock S and Park J. Projecting th performance of risk prediction from polygenic analyses of genome-wide association studies. *Nature Genetics* 2013, 45:400-5

2.Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P; GliomaScan Consortium, Yeager M, Chung CC, Chanock SJ, **Chatterjee N**. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet* 2012; 90:821-35

3.Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF Jr, **Chatterjee N.** Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci* 2011; 108:18026-31.

4.Park JH, Wacholder S, Gail M, Peters U, Jacobs K, Chanock S, **Chatterjee N**. Estimating effect size distribution from genome-wide association studies and implications for future discoveries.*Nat Genet* 2010; 42:570-5**.**

5.Rothman N, Garcia-Closas M, **Chatterjee N**, Maltas N, Wu Xifeng et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* 2010; 42:978-84.

CURRICULUM VITAE

NAME: Stephen J. Chanock, M.D.
Email: Chanocks@mail.nih.gov

PRESENT POSITIONS:

Director, Division of Cancer Epidemiology and Genetics

Director, Cancer Genomics Research Laboratory,
Division of Cancer Epidemiology and Genetics

EDUCATION:

1978	A.B., Princeton University, Princeton, NJ
1983	M.D., Harvard Medical School, Boston, MA

PROFESSIONAL APPOINTMENTS:

2001-07	Tenured Investigator, Head, Section on Genomic Variation, Pediatric Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD
2001-Pres.	Director, Core Genotyping Facility (now Cancer Genomics Research Laboratory), National Cancer Institute, Gaithersburg, MD and Office of the Director, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD
2005-Pres.	Co-Director, Cancer Genetic Markers of Susceptibility NCI Strategic Initiative, Office of the Director, NCI
2006-08	Appointed to the Senior Biomedical Research Service, National Institutes of Health, Bethesda, MD
2007-Pres.	Chief, Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Gaithersburg, MD
2012-2013	Acting Co-director, Center for Cancer Genomics, Office of the Director, National Cancer Institute, Bethesda, MD
2012-Pres.	Executive Committee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

NIH SERVICE RESPONSIBILITIES:

2010-Pres.	NCI Member of the NIH Senior Oversight Committee for Genomic Data Sharing
2010-Pres.	Member of NCBI Advisory Group
2011-Pres.	Provocative Questions Working Group
2012-Pres.	Steering Committee, International Cancer Genome Consortium

BIBLIOGRAPHY:

> 750 peer review publications

Simina Maria Boca, Ph.D.

Division of Cancer Epidemiology and Genetics,
National Cancer Institute
9609 Medical Center Drive, Room 7E604
Bethesda, MD 20892

Telephone: 240-276-7422
Email: simina.boca@nih.gov

Education

Ph.D. in Biostatistics	Johns Hopkins Bloomberg School of Public Health	2006-2011
M.H.S. in Bioinformatics	Johns Hopkins Bloomberg School of Public Health	2009-2011
B.S. in Mathematics	University of Illinois at Urbana-Champaign	2003-2006

Research Experience

Postdoctoral Fellow	Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute	2011-Present
Research Assistant	Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health and Sidney Kimmel Comprehensive Cancer Center	2007-2011
Visiting Researcher	Departments of Human Genetics and Biostatistics, University of Michigan	Summer 2008, Summer 2007
Research Aide	Bioinformatics Group, Mathematics and Computer Science Division, Argonne National Laboratory	Summer 2006, Summer 2005

Selected Publications

1. **Boca SM**, Sinha R, Cross AJ, Moore SC, Sampson JN. "Testing multiple biological mediators simultaneously." *Bioinformatics*, 2013. doi:10.1093/bioinformatics/btt633. [Epub ahead of print]

2. **Boca SM**, Corrada Bravo H, Caffo B, Leek JT, Parmigiani G. "A decision-theory approach to interpretable set analysis for high-dimensional data." *Biometrics*, 2013, 69(3):614-623.

3. **Boca SM+**, Rosenberg NA. "Mathematical properties of Fst between admixed populations and their parental source populations." *Theoretical Population Biology*, 2011, 80(3):208-216.

   + = SM Boca corresponding author

4. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, **Boca SM**, Carter H, Samayoa J, Bettegowda C, Gallia GL, Jallo GI, Binder ZA, Nikolsky Y, Hartigan J, Smith DR, Gerhard DS, Fults DW, VandenBerg S, Berger MS, Marie SKN, Shinjo SMO, Clara C, Phillips PC, Minturn JE, Biegel JA, Judkins AR, Resnick AC, Storm PB, Curran T, He Y, Rasheed BA, Friedman HS, Keir ST, McLendon R, Northcott PA, Taylor MD, Burger PC, Riggins GJ, Karchin R, Parmigiani G, Bigner DD, Yan H, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. "The genetic landscape of the childhood cancer medulloblastoma." *Science*, 2011, 331(6016):435-439.

5. **Boca SM**, Kinzler K, Velculescu VE, Vogelstein B, and Parmigiani G. "Patient-oriented gene set analysis for cancer mutation data." *Genome Biology*, 2010, 11: R112.

6. Wood LD, Parsons W, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, **Boca SM**, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKW, Sukumar S, Polyak C, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW , Velculescu VE, Vogelstein B. "The genomic landscapes of human breast and colorectal cancers." *Science*, 2007, 318(5853):1108-1113.