

Abstracts of papers presented
at the 2015 meeting on

GENOME INFORMATICS

October 28–October 31, 2015



Cold Spring Harbor Laboratory

1890
2015

Abstracts of papers presented
at the 2015 meeting on

GENOME INFORMATICS

October 28–October 31, 2015


Arranged by

Janet Kelso, *Max Planck Institute for Evolutionary Anthropology,
Germany*

Daniel MacArthur, *Massachusetts General Hospital*

Michael Schatz, *Cold Spring Harbor Laboratory*

This meeting was funded in part by the **National Human Genome Research Institute**, a branch of the **National Institutes of Health**.

Poster competition sponsored by  GENOME RESEARCH

A \$500 prize plus a one-year subscription to *Genome Research* will be awarded for an outstanding poster presentation.

Eligible entrants are graduate students and post-docs presenting novel unpublished work of genome-wide significance, including significant advances in large-scale genome informatics. Senior investigators and speakers who also present the work in sessions are not eligible. A panel of judges drawn from the organizing committee and the *Genome Research* Editorial Board will review all eligible posters.

Contributions from the following companies provide core support for the Cold Spring Harbor meetings program.

Corporate Sponsors

Agilent Technologies
Bristol-Myers Squibb Company
Genentech
Life Technologies (part of Thermo Fisher Scientific)
New England BioLabs

Plant Corporate Associates

Monsanto Company

The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

Cover Design: Alex Cagan, Max Planck Institute for Evolutionary Anthropology.

GENOME INFORMATICS

Wednesday, October 28 – Saturday, October 31, 2015

Wednesday	7:30 pm	1 Personal and Medical Genomics
Thursday	9:00 am	2 Transcriptomics, Alternative Splicing and Gene Predictions
Thursday	2:00 pm	3 Poster Session I
Thursday	3:30 pm	Keynote Speaker
Thursday	4:30 pm	<i>Wine and Cheese Party*</i>
Thursday	7:30 pm	4 Epigenomics and Non-coding Genome
Friday	9:00 am	5 Databases, Data Mining, Visualization, Ontologies and Curation
Friday	1:30 pm	6 Sequencing Pipelines and Assembly
Friday	4:30 pm	Keynote Speaker
Friday	5:30 pm	7 Poster Session II / Cocktail Party
Friday	7:00 pm	Banquet
Saturday	9:00 am	8 Comparative, Evolutionary and Metagenomics

* *Airlie Lawn*, weather permitting

Mealtimes at Blackford Hall are as follows:

Breakfast 7:30 am-9:00 am

Lunch 11:30 am-1:30 pm

Dinner 5:30 pm-7:00 pm

Bar is open from 5:00 pm until late

Abstracts are the responsibility of the author(s) and publication of an abstract does not imply endorsement by Cold Spring Harbor Laboratory of the studies reported in the abstract.

These abstracts should not be cited in bibliographies. Material herein should be treated as personal communications and should be cited as such only with the consent of the author.

Please note that ANY photography or video/audio recording of oral presentations or individual posters is strictly prohibited except with the advance permission of the author(s), the organizers, and Cold Spring Harbor Laboratory.

Printed on 100% recycled paper.

PROGRAM

WEDNESDAY, October 28—7:30 PM

SESSION 1 PERSONAL AND MEDICAL GENOMICS

Chairpersons: **Deanna Church**, Personalis, Inc., Menlo Park, California
Karyn Meltz Steinberg, Washington University, St. Louis, Missouri

The Rumsfeldian challenges of developing clinical sequencing tests

Deanna M. Church, Sarah Garcia, Jennifer Yen, Jason Harris, Stephen Chervitz, Aldrin Montana, Brian Linebaugh, Shujun Luo, Gabor Bartha, Elena Helman, Sean Boyle, Ravi Alla, Michael Clark, Martina Lefterova, Massimo Morra, John West, Richard Chen.
Presenter affiliation: Personalis, Inc., Menlo Park, California.

1

Haplotype-based somatic mutation calling in heterogeneous cancer samples

Daniel Cooke, Gerton Lunter.
Presenter affiliation: Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom.

2

SASE-hunter—A computational method to detect signatures of accelerated somatic evolution in non-coding regions of cancer genomes

Kyle Smith, Vinod Yadav, Brent Pedersen, Rita Shaknovich, Katherine Pollard, Subhajyoti De.
Presenter affiliation: University of Colorado, Aurora, Colorado.

3

Predictive modeling of coronary artery calcification using decision trees

Cihan Oguz, Shurjo K. Sen, Amadou Gaye, Ryan A. Neff, Adam R. Davis, Leslie G. Biesecker, Gary H. Gibbons.
Presenter affiliation: National Human Genome Research Institute, Bethesda, Maryland.

4

Identifying low frequency and rare coding variation influencing cardiometabolic traits through whole exome sequencing of 20,000 Finns—The FinMetSeq Study

Karyn Meltz Steinberg, Adam E. Locke, Sue Service, Vasily Ramensky, Matti Pirinen, Heather M. Stringham, Anne U. Jackson, Mitja Kurki, Laura J. Scott, Robert S. Fulton, Daniel C. Koboldt, Samuli Ripatti, Veikko Salomaa, Aarno Palotie, Markku Laakso, Nelson Freimer, Michael Boehnke, Richard K. Wilson.

Presenter affiliation: Washington University, St. Louis, Missouri.

5

Assessing tumor heterogeneity and tracking clonal clearance in response to therapy

Christopher A. Miller, Ha X. Dang, Gue Su Chang, Allegra Petti, Jeffrey Kico, Charles Lu, Eric J. Duncavage, Yevgeniy Gindin, Obi L. Griffith, Malachi Griffith, Meagan A. Jacoby, Geoff Uy, Christopher Maher, Matthew Ellis, Matthew Walter, Timothy Ley, Elaine Mardis, Richard Wilson.

Presenter affiliation: Washington University School of Medicine, St Louis, Missouri.

6

Identifying novel drivers of CD8+ T cell exhaustion in tumor

Meromit Singer, Chao Wang, Le Cong, Huiyuan Zhang, Sema Kurtulus, Junrong Xia, James Nevin, Orit Rozenblatt-Rosen, Vijay K. Kuchroo, Ana C. Anderson, Aviv Regev.

Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts.

7

Using Mendelian randomization to investigate association between gene expression variation and complex traits

YoSon Park, Ian McDowell, Genna Gliner, Benjamin F. Voight, Barbara E. Engelhardt, Christopher D. Brown.

Presenter affiliation: Perelman School of Medicine University of Pennsylvania, Philadelphia, Pennsylvania.

8

THURSDAY, October 29—9:00 AM

SESSION 2 TRANSCRIPTOMICS, ALTERNATIVE SPLICING AND GENE PREDICTIONS

Chairpersons: **Cole Trapnell**, University of Washington, Seattle, Washington
Alexis Battle, Johns Hopkins University, Baltimore, Maryland

Differential analysis of bifurcating single-cell gene expression trajectories

Xiaojie Qiu, Andrew Hill, Cole Trapnell.

Presenter affiliation: University of Washington, Seattle, Washington. 9

Kallisto—Near-optimal RNA-Seq quantification

Nicolas L. Bray, Harold Pimentel, Páll Melsted, Lior Pachter.

Presenter affiliation: University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, Iceland. 10

Scikit-ribo—Accurate A-site prediction and robust modeling of translation control from Riboseq and RNAseq data

Han Fang, Max Doerfel, Gholsen J. Lyon, Michael C. Schatz.

Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; Stony Brook University, Stony Brook, New York. 11

Detection and interpretation of genome structural variation in GTEx samples

Colby Chiang, Ryan M. Layer, Ryan P. Smith, Alexandra J. Scott, Amy B. Wilfert, Donald F. Conrad, Ira M. Hall.

Presenter affiliation: Washington University School of Medicine, St. Louis, Missouri. 12

Integrative models for predicting the regulatory impact of rare non-coding variation

Alexis Battle.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

An analysis of splicing variation across the sequence read archive with Rail-RNA

Abhinav Nellore, Leonardo Collado-Torres, José Alquicira-Hernández, Siruo Wang, Robert A. Phillips, Nishika Karbhari, Andrew E. Jaffe, Ben Langmead, Jeffrey T. Leek.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 13

Temporal transcriptomics reveals dysregulation of twin-peaking genes which reset the clock in a mouse model of psychiatric disease

William G. Pembroke, Arran Babbs, Kay E. Davies, Chris P. Ponting, Peter L. Oliver.

Presenter affiliation: MRC Functional Genomics Unit, Oxford, United Kingdom. 14

JunctionSeq—Detecting differential splice junction usage via RNA-Seq

Stephen W. Hartley, James C. Mullikin.

Presenter affiliation: National Human Genome Research Institute, Bethesda, Maryland. 15

THURSDAY, October 29—2:00 PM

SESSION 3

POSTER SESSION I

***Poster #s 46 – 135
See list beginning on p. xv***

THURSDAY, October 29—3:30 PM

KEYNOTE SPEAKER

Aviv Regev
Broad Institute

THURSDAY, October 29—4:30 PM

Wine and Cheese Party

SESSION 4 EPIGENOMICS AND NON-CODING GENOME

Chairpersons: **Melissa Gymrek**, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts
Michael Hoffman, University of Toronto, Toronto

Characterization of the microsatellite mutation process at every locus in the genome

Melissa Gymrek, Thomas Willems, Nick Patterson, David Reich, Yaniv Erlich.

Presenter affiliation: Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts; New York Genome Center, New York, New York.

16

Basset—Learning the regulatory code of the accessible genome with deep convolutional neural networks

David R. Kelley, Jasper Snoek, John L. Rinn.

Presenter affiliation: Harvard University, Cambridge, Massachusetts; Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts.

17

Differential nuclease sensitivity profiling of human chromatin reveals cell-type specific nucleosome positions, nucleosome sensitivity, open chromatin, and transcription factor binding

Daniel L. Vera, Eli Rodgers-Melnick, Sergiusz Wesolowski, Jonathan H. Dennis.

Presenter affiliation: Florida State University, Tallahassee, Florida.

18

Reorganization of chromosome architecture in cellular senescence

Steven W. Criscione, Marco De Cecco, Benjamin Siranosian, Yue Zhang, Jill A. Kreiling, John M. Sedivy, Nicola Neretti.

Presenter affiliation: Brown University, Providence, Rhode Island.

19

Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet

Coby Viner, James Johnson, Nicolas Walker, Marcela Sjoberg, David J. Adams, Anne C. Ferguson-Smith, Timothy L. Bailey, Michael M. Hoffman.

Presenter affiliation: University of Toronto, Toronto, Canada; Princess Margaret Cancer Centre, Toronto, Canada.

20

Unbiased discovery of cis-regulatory elements that determine mRNA post-transcriptional regulation during early development

Michal Rabani, Guo-Liang Chew, Alexander F. Schier.

Presenter affiliation: Harvard University, Cambridge, Massachusetts.

21

Hidden RNA codes revealed from the plant *in vivo* RNA structurome

Yin Tang, Philip C. Bevilacqua, Sarah M. Assmann.

Presenter affiliation: Pennsylvania State University, University Park, State College, Pennsylvania.

22

Using a bump hunting approach for genome-wide identification of novel imprinted genes

Yulia Rubanova, Andrei Turinsky, Sanaa Choufani, Rosanna Weksberg, Michael Brudno.

Presenter affiliation: University of Toronto, Toronto, Canada; The Hospital for Sick Children, Toronto, Canada.

23

FRIDAY, October 30—9:00 AM

SESSION 5 DATABASES, DATA MINING, VISUALIZATION, ONTOLOGIES AND CURATION

Chairpersons: **Suzanna Lewis**, Lawrence Berkeley National Laboratory, California
Jan Aerts, KU Leuven, Belgium

Rationally organizing the community's genotype/phenotype data

Suzanna Lewis.

Presenter affiliation: Lawrence Berkeley National Laboratory, Berkeley, California.

PCxN, the pathway co-activity map—A new approach for the unification of functional biology <u>Yered Pita-Juarez, Gabriel Altschuler, Wenbin Wei, Winston Hide.</u> Presenter affiliation: Harvard School of Public Health, Boston, Massachusetts; University of Sheffield, Sheffield, United Kingdom.	24
Genome-wide copy number variation analysis for <i>Plasmodium vivax</i> global isolates <u>Zunping Luo, Daniel N. Hupalo, Daniel E. Neafsey, Jane M. Carlton.</u> Presenter affiliation: Center for Genomics and Systems Biology, New York, New York.	25
RareVariantVis—A new tool for identification of causative variants in rare monogenic disorders from whole genome sequencing data <u>Tomasz Stokowy, Mateusz Garbulowski, Torunn Fiskerstrand, Rita Holdhus, Kornel Labun, Pawel Sztromwasser, Christian Gilissen, Alexander Hoischen, Gunnar Houge, Kjell Petersen, Inge Jonassen, Vidar Steen.</u> Presenter affiliation: University of Bergen, Bergen, Norway.	26
Visual data analysis—About discovering unknown unknowns and opening black boxes <u>Jan Aerts.</u> Presenter affiliation: KU Leuven, Leuven, Belgium.	
Discovery of genetic heterogeneity in a context of physiological homogeneity by biological distance clustering <u>Yuval Itan, Shen-Ying Zhang, Laurent Abel, Jean-Laurent Casanova.</u> Presenter affiliation: The Rockefeller University, New York, New York.	27
Ginkgo—Interactive analysis and quality assessment of single-cell CNV data <u>Robert Aboukhalil, Tyler Garvin, Jude Kendall, Timour Baslan, Gurinder S. Atwal, James Hicks, Michael Wigler, Michael C. Schatz.</u> Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.	28
Evaluating the application of sequencing data to differential coexpression using the discordant method <u>Charlotte J. Siska, Katerina Kechris.</u> Presenter affiliation: University of Colorado Anschutz Medical Campus, Aurora, Colorado.	29

SESSION 6 SEQUENCING PIPELINES AND ASSEMBLY

Chairpersons: **Gerton Lunter**, University of Oxford, United Kingdom
Aaron Quinlan, University of Utah, Salt Lake City

Basecalling from raw Oxford Nanopore data

Gerton Lunter.

Presenter affiliation: University of Oxford, Oxford, United Kingdom. 30

Diploid genome assembly and comprehensive haplotype sequence reconstruction

Jason Chin, Paul Peluso, David Rank, Maria Nettestad, Michael Schatz, Alicia Clum, Alex Copeland, Barry Kerrie.

Presenter affiliation: Pacific Biosciences, Menlo Park, California. 31

HISAT2—Graph-based alignment of next-generation sequencing reads to a population of human genomes

Daehwan Kim, Steven L. Salzberg.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 32

Comprehensive genome and transcriptome structural analysis of a breast cancer cell line using PacBio long read sequencing

Maria Nattestad, Karen Ng, Sara Goodwin, Timour Baslan, Fritz Sedlazeck, James Gurtowski, Elizabeth Hutton, Marley Alford, Elizabeth Tseng, Jason Chin, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John McPherson, James Hicks, Michael Schatz, W R. McCombie.

Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 33

Why is “querying” the genome so difficult?

Aaron Quinlan.

Presenter affiliation: University of Utah, Salt Lake City, Utah. 34

Genome scaffolding and structural variation detection from MinION Nanopore sequencing data

Zemin Ning, Louise Aigrain, James Bonfield, Robert Davies, Michael Quail, David Jackson, Thomas Keane, Richard Durbin.

Presenter affiliation: The Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom. 35

How to compare and cluster every known genome in about an hour

Adam M. Phillippy, Brian D. Ondov, Todd J. Treangen, Sergey Koren.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 36

Improved methods for NGS-based conotoxin discovery

Qing Li, Pradip K. Bandyopadhyay, Helena Safavi-Hemami, Samuel D. Robinson, Aiping Lu, Jason S. Biggs, Baldomero M. Olivera, Mark Yandell.

Presenter affiliation: University of Utah, Salt Lake City, Utah. 37

FRIDAY, October 30—4:30 PM

KEYNOTE SPEAKER

Mark Gerstein
Yale University

FRIDAY, October 30—5:30 PM

SESSION 7 POSTER SESSION II and COCKTAILS

Poster #s 136 – 211
See list beginning on p. xxviii

FRIDAY, October 30

BANQUET

Dinner 7:00 PM

SESSION 8 COMPARATIVE, EVOLUTIONARY, AND
METAGENOMICS

Chairpersons: **Aoife McLysaght**, Trinity College Dublin, Ireland
Adam Siepel, Cold Spring Harbor Laboratory, New York

Dosage sensitive genes in evolution and disease

Aoife McLysaght.

Presenter affiliation: Trinity College Dublin, Dublin, Ireland. 38

Computational analysis of disease-associated functional shifts in the periodontal microbiome

Shareef M. Dabdoub, Sukirth M. Ganesan, Purnima S. Kumar.

Presenter affiliation: The Ohio State University, Columbus, Ohio. 39

A genetic analysis of a complex trait in a “genetically intractable” gut microbe

Senā Bae, Olaf Muller, Sandi Wong, John F. Rawls, Raphael H. Valdivia.

Presenter affiliation: Duke University, Durham, North Carolina. 40

Comparative genomics over 50 newly-sequenced species of parasitic worms

Diogo M. Ribeiro, Avril Coghlan, Nancy Holroyd, Eleanor Stanley, Jason Tsai, Bhavana Harsha, Makedonka Mitreva, James Cotton, Matthew Berriman.

Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 41

Genome-wide inference of natural selection on regulatory sequences in the human genome

Brad Gulko, Ilan Gronau, Melissa J. Hubisz, Adam Siepel.

Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 42

A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins

James T. Morton, Stefan D. Freed, Shaun W. Lee, Iddo Friedberg.

Presenter affiliation: Iowa State University, Ames, Iowa. 43

Genomic assembly and analysis of highly heterozygotic polyploid parasitic protists
Aaron R. Jex, Staffan Svard, Robin B. Gasser.
Presenter affiliation: University of Melbourne, Parkville, Australia. 44

A novel software tool and pipeline for the classification of metagenomics sequencing data, and their application to the diagnosis of neuropathological infections of the nervous system
Florian P. Breitwieser, Daehwan Kim, Li Song, Carlos A. Pardo, Steven L. Salzberg.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 45

POSTER SESSION I:

Comprehensive profiling of somatic mosaicism in human brain
Taejeong Bae, Flora M. Vaccarino, Alexej Abyzov.
Presenter affiliation: Mayo Clinic, Rochester, Minnesota. 46

Decoupling array CGH sample-reference hybridization pairs for normalization of log ratio artefacts.
Aled R. Jones, Kevin J. Ryan, Adele Corrigan, Joo Wook Ahn.
Presenter affiliation: Guy's & St Thomas' NHS Foundation Trust, London, United Kingdom. 47

5-hydroxymethylcytosine in *Daphnia pulex*
Gediminas Alzbutas, Dovile Strepetkaite, Eimantas Astromskas, Rasa Sabaliauskaite, Kestutis Arbaciauskas, Arunas Lagunavicius, Juozas Lazutka.
Presenter affiliation: Thermo Fisher Scientific, Vilnius, Lithuania; Vilnius University, Vilnius, Lithuania. 48

Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets
David Amar, Tom Hait, Shai Izraeli, Ron Shamir.
Presenter affiliation: Tel Aviv University, Tel Aviv, Israel. 49

<p>Optimization of <i>de novo</i> transcriptome assembly and differential expression analysis of salt tolerance genes in the halophyte <i>Suaeda fruticosa</i> <u>Joann Diray-Arce</u>, Mark J. Clement, Bilquees Gul, M Ajmal Khan, Brent L. Nielsen. Presenter affiliation: Brigham Young University, Provo, Utah.</p>	50
<p>Molecular delineation of two major oncogenic pathways governing invasive ductal breast cancer development <u>Luay Aswad</u>, Surya P. Yenamandra, Ow Ghim Siong, Anna V. Ivshina, Vladimir A. Kuznetsov. Presenter affiliation: Bioinformatics Institute (BII), Singapore; Nanyang Technological University, Singapore.</p>	51
<p>Detection of pathogen integration sites in cancer <u>Gnanaprakash Balasubramanian</u>, Barbara Hutter, Hamza Khan, Benedikt Brors. Presenter affiliation: German Cancer Research Institute (DKFZ), Heidelberg, Germany; National Center for Tumor Diseases (NCT), Heidelberg, Germany; German Consortium for Translational Cancer Research (DKTK), Heidelberg, Germany.</p>	52
<p>AuPairWise—Biologically focused RNA-seq quality control using co-expression <u>Sara Ballouz</u>, Jesse Gillis. Presenter affiliation: CSHL, Cold Spring Harbor, New York.</p>	53
<p>Integrated genomic analysis suggests inherited predisposition to cancer in therapy-related acute lymphoblastic leukemia (tr-ALL) <u>Riyue Bao</u>, Fuhong He, Mark M. Sasaki, Jane E. Churpek, Lei Huang, Qianfei Wang, Jorge Andrade, Kenan Onel. Presenter affiliation: The University of Chicago, Chicago, Illinois.</p>	54
<p>From the ground to the cloud in just minutes—Building a customized Galaxy analysis server using only a web browser <u>Daniel Blankenberg</u>, The Galaxy Team. Presenter affiliation: Penn State University, University Park, Pennsylvania; The Galaxy Project.</p>	55
<p>A modified MAKER structural genome annotation method reveals novel gene predictions of high and low GC content in rice <u>Megan J. Bowman</u>, Jane A. Pulman, Kevin L. Childs. Presenter affiliation: Michigan State University, East Lansing, Michigan.</p>	56

A software tool for data integration in a diagnostic laboratory <u>Riccardo Brumm</u> , Sebastian H. Eck, Betina Ebert, Ina Vogl, Sandra Kuecuck, Sabine Rath, Verena Hasselbacher, Christina Sofeso, Birgit Busse, Soheyla Chahrokh-Zadeh, Christoph Marshall, Karin Mayer, Imma Rost, Hanns-Georg Klein. Presenter affiliation: Center for Human Genetics and Laboratory Diagnostics, Munich, Germany.	57
GMOD in the Cloud 2.0 <u>Scott Cain</u> , Stephen Ficklin, Colin Diesh, Lacey Sanderson, Lincoln Stein. Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada.	58
Galaxy Tool World Progression—Happier developers, happier users John Chilton, <u>Martin Cech</u> , Björn Grüning, Eric Rasche, Galaxy Team. Presenter affiliation: Penn State University, University Park, Pennsylvania.	59
<i>slncky</i>—A software and novel approach for annotation and evolutionary analysis of LncRNAs <u>Jenny Chen</u> , XiaoPeng Zhu, Alexander A. Shishkin, Sabah Kadri, Itay Maza, Jacob H. Hanna, Aviv Regev, Manuel Garber. Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Massachusetts Institute of Technology, Cambridge, Massachusetts.	60
PickArmSite, investigate the preference of using one arm for miRNAs <u>Tingwen Chen</u> , Cheng-Yang Lee, Petrus Tang. Presenter affiliation: Chang Gung University, Taoyuan, Taiwan.	61
Abundant Inverted Duplicates in the Human and Mouse Genomes as Functional Regulatory elements evolving under sex-related selection <u>Zhen-Xia Chen</u> , Yong Zhang, Ge Gao, Brian Oliver, Manyuan Long. Presenter affiliation: National Institutes of Health, Bethesda, Maryland.	62
Building more expressive Galaxy workflows <u>John M. Chilton</u> , Galaxy Team. Presenter affiliation: Pennsylvania State University, University Park, Pennsylvania.	63

<p>Assessing new tools and best practices for RNA seq data analysis and visualization with iPlant cyberinfrastructure <u>Kapeel M. Chougule</u>, Andrew Olson, Jurandir Vieira De Magalhaes, Peter Van Buren, Liya Wang, Doreen Ware. Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; iPlant Collaborative , Tucson, Arizona.</p>	64
<p>Deconvolving gene expression profiles for tumor populations with prior frequency information <u>Christopher Cremer</u>, Quaid Morris. Presenter affiliation: University of Toronto, Toronto, Canada.</p>	65
<p>Comparison of chromosome structure across conditions using a three dimensional chromosome browser <u>Steven W. Criscione</u>, Yue Zhang, Benjamin Siranosian, Alan Hwang, Marco De Cecco, John M. Sedivy, Nicola Neretti. Presenter affiliation: Brown University, Providence, Rhode Island.</p>	66
<p>Wrangling data into track hub visualizations with <i>hubward</i> <u>Ryan K. Dale</u>. Presenter affiliation: National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland.</p>	67
<p>Identification of differentially expressed genes of developing mouse tooth within their genomic locations <u>Rishi Das Roy</u>, Outi Hallikas, Elodie Renvoise, Jukka Jernvall. Presenter affiliation: Institute of Biotechnology, Helsinki, Finland.</p>	68
<p>Comparing algorithms to genotype short tandem repeats in next-generation sequencing data <u>Harriet Dashnow</u>, Alicia Oshlack. Presenter affiliation: Murdoch Childrens Research Institute, Parkville, Victoria, Australia; The University of Melbourne, Parkville, Victoria, Australia.</p>	69
<p>The impact of RNA degradation on the ability to detect fusions using TruSeq RNA library Preparation <u>Jaime I. Davila</u>, Wang Xiaoke, Numrah Fadra, Nair Asha, McDonald Amber, Crusan Barbara, Kandelaria Rumilla, Jen Jin, Klee Eric, Kipp Benjamin, Halling Kevin. Presenter affiliation: Mayo Clinic, Rochester, Minnesota.</p>	70

<p>Additional variants among the MH-GRID cohort discovered after alignment to an ancestry specific reference genome <u>Adam R. Davis</u>, Ryan A. Neff, Shurjo Sen, Cihan Oguz, Rakale Quarells, Gary H. Gibbons. Presenter affiliation: National Human Genome Institute, Bethesda, Maryland.</p>	71
<p>ESAT—A new tool for analyzing end-sequencing RNA-Seq data <u>Alan Derr</u>, Alexey A. Sergushichev, Sabah Kadri, Sebastian Kadener, Maxim N. Artyomov, Manuel Garber. Presenter affiliation: University of Massachusetts Medical School, Worcester, Massachusetts.</p>	72
<p>Reconstructing the evolutionary history of tumours <u>Amit G. Deshwar</u>, Quaid Morris. Presenter affiliation: University of Toronto, Toronto, Canada.</p>	73
<p>Precision-STAR—Unbiased allele aware mapping of RNA-seq reads to personal genomes <u>Alexander Dobin</u>, Thomas R. Gingeras. Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.</p>	74
<p>Apollo—A platform for collaborative genome curation and analysis <u>Nathan A. Dunn</u>, Monica C. Munoz-Torres, Colin Diesh, Deepak Unni, Eric Yao, Ian Holmes, Christine Elsie, Suzanna E. Lewis. Presenter affiliation: Lawrence Berkeley National Labs, Berkeley, California.</p>	75
<p>Assembly-free comparative genomics of <i>Trichomonas vaginalis</i> and three other trichomonads <u>Daniel Ence</u>, Claudia P. Marquez, Mark Yandell, Ellen J. Pritham. Presenter affiliation: University of Utah, Salt Lake City, Utah; Eccles Institute of Human Genetics, Salt Lake City, Utah.</p>	76
<p>Biologically based disease classification for childhood arthritis <u>Simon W. Eng</u>, Rae S. Yeung, Quaid Morris. Presenter affiliation: University of Toronto, Toronto, Canada; The Hospital for Sick Children, Toronto, Canada.</p>	77

TuneSim—Tunable variant set Simulator for NGS reads <u>Bertrand Escalière</u> , Sonia Van Dooren, Raphael Helaers, Gianluca Bontempi, Guillaume Smits. Presenter affiliation: IB2, Brussels, Belgium; Université Libre de Bruxelles, Brussels, Belgium.	78
Benchmarking ultrafast workflows for human genome variant calling Gloria Redon, Volodymyr Kindratenko, Victor Jongeneel, Liudmila S. Mainzer, <u>Christopher Fields</u> . Presenter affiliation: University of Illinois, Urbana, Illinois.	79
Genomic region and sample selection strategy for variant discovery and association analysis <u>Steven M. Foltz</u> , Kai Ye, Li Ding. Presenter affiliation: Washington University School of Medicine, St. Louis, Missouri.	80
Transcriptome states identified by probabilistic modeling of CLIP-seq datasets in yeast <u>Mallory A. Freeberg</u> , James Taylor. Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.	81
Orthogonal sequencing of the human exome <u>Alexander Frieden</u> , Niru Chennagiri, Eric White, Daniel Lieber, John Thompson. Presenter affiliation: Claritas Genomics, Cambridge, Massachusetts.	82
K-mer spectra filters to assemble high quality, contiguous, collapsed mosaics from non-model heterozygous genomes <u>Gonzalo A. Garcia Accinelli</u> , Darren Heavens, Jens Maintz, Diane Saunders, Matt Clark, Federica Di Palma, Bernardo J. Clavijo. Presenter affiliation: The Genome Analysis Centre, Norwich, United Kingdom.	83
DIDA—A first digenic diseases database <u>Andrea Gazzo</u> , Dorien Daneels, Elisa Cilia, Maryse Bonduelle, Marc Abramowicz, Sonia Van Dooren, Guillaume Smits, Tom Lenaerts. Presenter affiliation: Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium; Machine Learning Group, Brussels, Belgium.	84
PIPES—A tool for classifying long RNA reads <u>Sam Kovaka</u> , Alex Dobin, Thomas R. Gingeras. Presenter affiliation: Clark University, Worcester, Massachusetts.	85

Red—An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale <u>Hani Z. Girgis</u> . Presenter affiliation: University of Tulsa, Tulsa, Oklahoma.	86
RefSeq annotation of functional elements on the human and mouse reference genomes <u>Tamara Goldfarb</u> , Catherine M. Farrell, Sanjida H. Rangwala, Terence D. Murphy, Donna R. Maglott, Kim D. Pruitt. Presenter affiliation: National Institutes of Health, Bethesda, Maryland.	87
Structural alteration of transcript isoforms in human cancers <u>Leonard D. Goldstein</u> , Eric Stawiski, Thong Nguyen, James Lee, David Stokoe, Robert Gentleman, Somasekar Seshagiri. Presenter affiliation: Genentech, South San Francisco, California.	88
MeTavGen—A Taverna-based pipeline for the analysis of shotgun metagenomic data <u>Giorgio Gonnella</u> , Laura Glau, Stefan Kurtz. Presenter affiliation: University of Hamburg, Hamburg, Germany.	89
Characterization of DNA damage response and R-loops in Ewing sarcoma <u>Aparna Gorthi</u> , Yidong Chen, Alexander Bishop. Presenter affiliation: University of Texas Health Science Center at San Antonio, San Antonio, Texas.	90
Improvement of the assembly of heterozygous genomes of non-model organisms <u>Anaïs Gouin</u> , Anthony Bretaudeau, Emmanuelle d'Alençon, Claire Lemaitre, Fabrice Legeai. Presenter affiliation: Inria/IRISA Équipe GenScale, Rennes, France.	91
Concordance and contamination checker for WGS and WES matched sample studies <u>Ewa A. Grabowska</u> , Phaedra Agius, Kanika Arora, Nora C. Toussaint, Dayna M. Oswald, Vladimir Vacic. Presenter affiliation: New York Genome Center, New York, New York.	92
ADAGE—A method for the unsupervised integration of gene expression experiments applied to <i>Pseudomonas aeruginosa</i> <u>Jie Tan</u> , John H. Hammond, Deborah A. Hogan, <u>Casey S. Greene</u> . Presenter affiliation: Geisel School of Medicine at Dartmouth, Hanover, New Hampshire; University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania.	93

The genome and transcriptome of the regeneration-competent flatworm, <i>Macrostomum lignano</i>	
Kaja A. Wasik, James Gurtowski, Xin Zhou, Olivia M. Ramos, M. Joaquina Delas, Giorgia Battistoni, Osama El Demerdash, Ilaria Faciadori, Dita B. Vizoso, Peter Ladurner, Lukas Scharer, W. Richard McCombie, Gregory J. Hannon, Michael C. Schatz.	
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.	94
Systematic analysis of alternative polyadenylation during neurogenesis of murine embryonic stem cells	
<u>Kevin Ha</u> , Benjamin Blencowe, Quaid Morris.	
Presenter affiliation: University of Toronto, Toronto, Canada.	95
CLAMMS—A scalable pipeline for CNV calling and quality control, applied to over 40,000 exomes	
<u>Lukas Habegger</u> , Jonathan S. Packer, Evan K. Maxwell, Colm O'Dushlaine, Alexander Lopez, Samantha N. Fetterolf, Joseph B. Leader, David J. Carey, David H. Ledbetter, Frederick E. Dewey, Rostislav Chernomorsky, Aris Baras, John D. Overton, Jeffrey G. Reid.	
Presenter affiliation: Regeneron Pharmaceuticals, Tarrytown, New York.	96
<i>proovread-3.0</i>—PacBio hybrid error correction for di-/polyploid genomes, metagenomes and transcriptomes	
<u>Thomas Hackl</u> , Frank Förster, Matthias G. Fischer.	
Presenter affiliation: Max-Planck-Institute for Medical Research, Heidelberg, Germany; University of Wuerzburg, Würzburg, Germany.	97
Utilization of very large read depth sequencing data for the detection of variation in a father-mother-son trio	
<u>Nancy F. Hansen</u> , James C. Mullikin, Genome in a Bottle Consortium.	
Presenter affiliation: National Human Genome Research Institute, Bethesda, Maryland.	98
TEpeaks—A tool for including repetitive sequences in ChIP-seq and Clip-seq analyses	
Ying Jin, <u>Yuan Hao</u> , David Molik, Oliver Tam, Molly Hammell.	
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.	99

CNVThresher—Combining multiple lines of evidence to construct high-quality CNV call sets <u>Jason Harris</u> , Gábor Bartha, Stephen Chervitz, Deanna M. Church, Richard Chen. Presenter affiliation: Personalis, Inc., Menlo Park, California.	100
The landscape of microsatellite instability in cancer exomes <u>Ronald J. Hause</u> , Emily H. Turner, Mallory Beightol, Colin C. Pritchard, Jay Shendure, Stephen J. Salipante. Presenter affiliation: University of Washington, Seattle, Washington.	101
On the relativity of time and space of tumors—Clinical segmentation is the key to eradicate breast cancer <u>Fritz E. Hauser</u> . Presenter affiliation: Aerzte-Gesundheitszentrum Laegern AG, www.ghzl.ch, Ehrendingen AG, Switzerland.	102
Using the landscape of genetic variation in protein domains to improve functional consequence predictions <u>Jim Havrilla</u> , Aaron Quinlan. Presenter affiliation: University of Utah, Salt Lake City, Utah.	103
Encore—A comprehensive framework for cancer sequencing analysis <u>Miao He</u> , Jennifer Becq, Stefano Berri, Mark Ross, David Bentley. Presenter affiliation: Illumina, Little Chesterford, United Kingdom.	104
MetaTreeMap—Visual representation of taxonomic assignment <u>Maxime Hebrard</u> , Todd D. Taylor. Presenter affiliation: RIKEN Center for Integrative Medical Sciences (IMS), Yokohama, Japan.	105
Highlander—Variant filtering made easier <u>Raphael Helaers</u> , Miikka Vakkula. Presenter affiliation: de Duve Institute, Université Catholique de Louvain, Brussels, Belgium.	106
On the widespread and critical impact of batch effects and systematic bias in single-cell RNA-Seq data <u>Stephanie C. Hicks</u> , Mingxiang Teng, Rafael A. Irizarry. Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts; Harvard T. H. Chan School of Public Health, Boston, Massachusetts.	107

The ENCODE Uniform Analysis Pipelines—

Case studies in cloud-based data analysis and distribution

Benjamin C. Hitz, J Seth Strattan, Esther T. Chan, Tim Dreszer, Jean M. Davidson, Nikhil R. Poddurturi, Laurence D. Rowe, Cricket A. Sloan, Forrest Y. Tanaka, Carrie Davis, Alex Dobin, Sarah Djebali, Roderic Guigo, Tom Gingeras, Colin Dewey, Xintao Wei, Brenton Graveley, J M. Cherry.

Presenter affiliation: Stanford University, Palo Alto, California. 108

Integrated approach to identify clinically relevant CNVs in ClinSeq® cohort

Celine Hong, David Ng, Jennifer Johnston, Dan King, Jim Mullikin, Leslie G. Biesecker.

Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 109

De novo assembly and next-generation sequencing to analyze full-length gene variants from codon-barcoded libraries

Byungjin Hwang, Hoon Jang, Sunghoon Huh, Duhee Bang.

Presenter affiliation: Yonsei University, Seoul, South Korea. 110

The human gene damage index—A novel gene-level approach to prioritize exome variations

Yuval Itan, Lei Shang, Lluís Quintana-Murci, Shen-Ying Zhang, Laurent Abel, Jean-Laurent Casanova.

Presenter affiliation: The Rockefeller University, New York, New York. 111

The mutation significance cutoff (MSC)—A gene-specific approach to predicting the impact of human gene variants

Yuval Itan, Lei Shang, Lluís Quintana-Murci, Shen-Ying Zhang, Laurent Abel, Jean-Laurent Casanova.

Presenter affiliation: The Rockefeller University, New York, New York. 112

Large scale correlation of epigenomics data

Jonathan Laperle, Alexei Nordell-Markovits, Marc-Antoine Robert, David Bujold, David Anderson De Lima Morais, Michel Barrette, Guillaume Bourque, Pierre-Étienne Jacques.

Presenter affiliation: Université de Sherbrooke, Sherbrooke, Canada. 113

Interrogating the mechanisms of schizophrenia genetic risk in the fully characterized human brain transcriptome

Andrew E. Jaffe, Jooheon Shin, Richard E. Straub, Ran Tao, Yuan Gao, Yankai Jia, Leonardo Collado-Torres, Jeffrey T. Leek, Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger.

Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland; Johns Hopkins University, Baltimore, Maryland. 114

- Plant Reactome—A reference resource for analyzing plant metabolic and regulatory pathways**
Pankaj Jaiswal, Justin Preece, Vindhya Amarasinghe, Palitha Dharmawardhana, Peter D'Eustachio, Sushma Naithani, Guanming Wu, Antonio F. Mundo, Robin Haw, Sheldon Mckay, Joel Weiser, Lincoln Stein, Doreen Ware.
 Presenter affiliation: Oregon State University, Corvallis, Oregon. 115
- BAMQC—A quality control tool for mapped next generation sequencing datasets**
Ying Jin, Oliver Tam, David Molik, Molly Hammell.
 Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 116
- Utilization of high-throughput sequencing to detect candidate genes associated with mouse reproductive longevity**
Jyoti Joshi, Kacper Zukowski, Nehil Jain, Jeremy E. Koenig, Robert G. Beiko, Hein van der Steen.
 Presenter affiliation: Dalhousie University, Halifax, Canada. 117
- Construction of transcription factor networks for obesity using RNAseq transcriptomics**
 Ruta Skinkyte-Juskiene, Lisette J. Kogelman, Haja N. Kadarmideen.
 Presenter affiliation: University of Copenhagen, Frederiksberg C, Denmark. 118
- Genetic network method based analysis of antidepressant treatment**
 Majbritt B. Madsen, Lisette J. Kogelman, Henrik B. Rasmussen, Haja N. Kadarmideen.
 Presenter affiliation: ; University of Copenhagen, Frederiksberg C, Denmark. 119
- Building a meta-metagenome graph**
Andre Kahles, Gunnar Ratsch.
 Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York. 120
- A dynamic programming approach to the reconstruction of prokaryote gene block evolution history**
 Carly Schaeffer, David Ream, Iddo Friedberg, John Karro.
 Presenter affiliation: Miami University, Oxford, Ohio. 121

- Chop-Stitch—Targeted assembly of genes using transcriptome assembly and Bloom filter-based de Bruijn graphs**
Hamza Khan, Benjamin P. Vandervalk, René L. Warren, Inanc Birol.
 Presenter affiliation: Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada. 122
- Role of lincRNA in T helper cell differentiation**
Mohd Moin Khan, Ubaid Ullah, Omid Rasool, Sini Rautio, Zhi Chen, Riitta Lahesmaa.
 Presenter affiliation: Turku Centre for Biotechnology, Turku, Finland; Turku Doctoral Programme of Molecular Medicine (TuDMM) , Turku, Finland. 123
- Comprehensive, flexible pipelines for alternative polyadenylation analysis**
Hyunmin Kim, Jihye Kim, Nova Fong, David Bentley.
 Presenter affiliation: University of Colorado School of Medicine, Aurora, Colorado. 124
- Identification and filtration of false somatic variants caused by vector contamination**
Junho Kim, Ju Heon Maeng, Jae Seok Lim, Junehawk Lee, Jeong Ho Lee, Sangwoo Kim.
 Presenter affiliation: Yonsei University College of Medicine, Seoul, South Korea. 125
- Genomic-based 16S ribosomal RNA database web server and tools**
Seok-Won Kim, Masahira Hattori, Todd D. Taylor.
 Presenter affiliation: RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. 126
- Improving accessibility and usability of genome data at NCBI**
Paul Kitts, Michael DiCuccio, Avi Kimchi, Terence Murphy, Kim Pruitt, Tatiana Tatusova.
 Presenter affiliation: National Center for Biotechnology Information (NCBI), Bethesda, Maryland. 127
- JBAM quality score compression**
James Knight.
 Presenter affiliation: Yale University, New Haven, Connecticut. 128

- Canu—A new single-molecule sequence assembler for genomes large and small**
Sergey Koren, Brian Walenz, Adam M. Phillippy.
 Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 129
- Phylogeny of the spider mite sub-family Tetranychinae (Acari: Tetranychidae) from Japan reconstructed by their transcriptomes**
Toshinori Kozaki, Tomoko Matsuda, Kazuo Ishii, Tetsuo Gotoh.
 Presenter affiliation: Tokyo University of Agriculture and Technology, Fuchu Tokyo, Japan. 130
- Bringing genomic data into focus for studying complex diseases in specific biological contexts**
Arjun Krishnan, Ran Zhang, Victoria Yao, Chandra Theesfeld, Aaron Wong, Alicja Tadych, Alan Packer, Alex Lash, Olga G. Troyanskaya.
 Presenter affiliation: Lewis-Sigler Institute for Integrative Genomics, Princeton, New Jersey. 131
- Kollector—Transcript-guided targeted assembly of genes**
Erdi Kucuk.
 Presenter affiliation: British Columbia Cancer Agency, Vancouver, Canada; University of British Columbia, Vancouver, Canada. 132
- Dolphin—Large-scale sequencing analysis platform**
Alper Kucukural, Nicholas Merowsky, Alastair Firth, Manuel Garber.
 Presenter affiliation: UMass Medical School, Worcester, Massachusetts. 133
- Multimedia annotation using the iCLiKVAL browser extension**
Naveen Kumar, Todd D. Taylor.
 Presenter affiliation: RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. 134
- The DOE Systems Biology Knowledgebase—A system for collaborative and reproducible inference and modeling of biological function**
Vivek Kumar, Sunita Kumari, James Thomason, Mike Schatz, Doreen Ware, Sergei Maslov, Robert W. Cottingham, Rick Stevens, Adam Arkin.
 Presenter affiliation: Cold Spring Harbor Laboratory (CSHL), Cold Spring Harbor, New York. 135

POSTER SESSION II:

A comparative analysis of network mutation burdens across 21 tumor types predicts new candidate cancer genes in the tail of the mutation distribution of existing cancer genomes

Heiko Horn, Michael Lawrence, Jessica Hu, Elizabeth Worstell, Nina Ilic, Yashaswi Shresta, Eejung Kim, Atanas Kamburov, Alireza Kashani, William Hahn, Jesse Boehm, Gad Getz, [Kasper Lage](#).

Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts; Broad Institute, Cambridge, Massachusetts.

136

Large-scale prediction of pathways from GWAS and exome-sequencing projects by a systematic analysis of differential pathway architectures in diverse functional genomic networks

John Mercer, Joseph Rosenbluh, Arthur Liberzon, Dawn Thompson, Thomas Eisenhaure, Steve Carr, Jake Jaff, Jesse Boehm, Aviad Tsherniak, Aravind Subramanian, Sarah Calvo, Taibo Li, Ted Liefeld, Bang Wong, Jill Mesirov, Nir Hacohen, Aviv Regev, [Kasper Lage](#).

Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts.

137

Identification of the genetic basis underlying alternative reproductive strategies in the ruff (*Philomachus pugnax*)

[Sangeet Lamichhane](#)y, Guangyi Fan, Fredrik Widemo, Ulrika Gunnarsson, Doreen S. Thalmann, Marc Höppner, Susanne Kerje, Ulla Gustafson, BGI sequencing team, Jacob Höglund, Xin Liu, Leif Andersson.

Presenter affiliation: Science for Life Laboratory, Uppsala, Sweden.

138

The resurgence of reference quality genome

[Hayan Lee](#), James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, Richard W. McCombie, Michael C. Schatz.

Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; Stony Brook University, Stony Brook, New York.

139

Data-driven characterization of human complex diseases

[Young-suk Lee](#), Arjun Krishnan, Olga Troyanskaya.

Presenter affiliation: Princeton University, Princeton, New Jersey.

140

Forward—A bioinformatics tool to manage, execute and explore phenomic studies

Marc-André Legault, Louis-Philippe Lemieux Perreault, Jean-Claude Tardif, Marie-Pierre Dubé.

Presenter affiliation: Montreal Heart Institute, Montreal, Canada; Université de Montréal, Montreal, Canada.

141

genipe—A Python module to perform genome-wide imputation analysis

Louis-Philippe Lemieux Perreault, Marc-André Legault, Marie-Pierre Dubé.

Presenter affiliation: Beaulieu-Saucier Pharmacogenomics Centre at the Montreal Heart Institute, Montreal, Canada.

142

FastqDemultiplex—A flexible demultiplexing tool for Illumina reads

Florian Lenz.

Presenter affiliation: University Medical Center Tuebingen, Tuebingen, Germany.

143

Speeding up long-read assembly by reducing alignments overlap due to repeats

Shoudan Liang, Paul Peluso, Yinping Jiao, Doreen Ware, David Rank, Chen-Shan J. Chin.

Presenter affiliation: Pacific Biosciences, Menlo Park, California.

144

Poseidon—A highly sensitive and efficient taxonomy classifier

EunCheon Lim.

Presenter affiliation: Max-Planck Institute for Developmental Biology, Tuebingen, Germany.

145

Transcriptome and epigenome profiling of human olfactory mucosa stem cells as a model system for autism spectrum disorders

Tharvesh Moideen Liyakat Ali, Mélanie Makhlof, Lam Son Nguyen, Karine Siquier-Pernet, François Féron, Bruno Gepner, Laurence Colleaux, Céline Vallot, Claire Rougeulle.

Presenter affiliation: CNRS UMR7216 Epigenetic and Cell Fate, Paris, France.

146

- Genomsawit—A one-stop genome information portal for oil palm**
Leslie Eng-Ti Low, Kuang-Lim Chan, Mohd Amin Ab Halim, Corey Wischmeyer, Smith W. Steven, Rozana Rosli, Norazah Azizi, Nik Shazana Nik Mohd Sanusi, Nadzirah Amiruddin, Jayanthi Nagappan, Leslie Cheng-Li Ooi, Pek-Lan Chan, Ngoot-Chin Ting, Michael Hogan, Rajinder Singh, Meilina Ong-Abdullah, Robert A. Martienssen, Ravigadevi Sambanthamurthi.
 Presenter affiliation: Malaysian Palm Oil Board, Kuala Lumpur, Malaysia. 147
- Determining the hypoxic gene expression response of *S. cerevisiae* cells using RNA-seq and statistical analysis of time-course data**
Samuel Maclean, Gurmanna Kalra, Nasrine Bendjilali, Mark J. Hickman.
 Presenter affiliation: Rowan University, Glassboro, New Jersey. 148
- Microbial genome assembly using synthetic error-free reads**
Mohammed-Amin Madoui, Stefan Engelen, Corinne Cruaud, Arnaud Lemainque, Patrick Wincker, Jean-Marc Aury.
 Presenter affiliation: Genoscope-CNS, Evry, France. 149
- Targeted sequencing of FFPE ovarian cancer tumour samples on the Ion PGM platform**
Alison Meynert, Michael Churchman, Robb Hollis, Tzyvia Rye, Angie Fawkes, Lee Murphy, Colin Semple, Charlie Gourley.
 Presenter affiliation: University of Edinburgh, Edinburgh, United Kingdom. 150
- Automated transfer of workflows from Galaxy to Yabi and command line tools**
David C. Molik, Ying Jin, Molly Hammell.
 Presenter affiliation: Cold Spring Harbor Lab, Cold Spring Harbor, New York. 151
- Searching and exploring Gramene’s comparative genomics datasets on the web**
Joseph Mulvaney, Andrew Olson, James Thomason, Doreen Ware.
 Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 152

- Comparative analysis of chromatin states and gene expression profiles for various endothelium cells**
Ryuichiro Nakato, Yuki Katou, Toutai Mituyama, Hiroshi Kimura, Youichiro Wada, Hiroyuki Aburatani, Katsuhiko Shirahige.
 Presenter affiliation: University of Tokyo, Bunkyo-ku, Japan. 153
- Probing transcriptional regulation in tumor specimens yields hallmarks of prostate cancer outcome**
Ekaterina Nevedomskaya, Suzan Stelloo, Henk G. van der Poel, Jeroen de Jong, Geert J. van Leenders, Guido Jenster, Lodewyk F. Wessels, Andre M. Bergman, Wilbert Zwart.
 Presenter affiliation: Netherlands Cancer Institute, Amsterdam, Netherlands. 154
- Computational analysis of target specificity of double-stranded RNA binding protein Staufen**
Kun Nie, Quaid D. Morris.
 Presenter affiliation: Terrence Donnelly Center for Cellular and Biomolecular Research, Toronto, Canada; University of Toronto, Toronto, Canada. 155
- Comparative genome analysis of members of the Magnaporthaceae family of fungi**
Laura H. Okagaki, Joshua K. Sailsbry, Alexander W. Eyer, Titus John, Cristiano C. Nunes, Ralph A. Dean.
 Presenter affiliation: North Carolina State University, Raleigh, North Carolina. 156
- Integrated web services supporting search and interactive analysis tools at Gramene**
Andrew Olson, Kapeel Chougule, Joseph Mulvaney, Justin Preece, James Thomason, Doreen Ware.
 Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 157
- The identification of the gene expression signatures of tissue-specific effects of pioglitazone treatments in a murine model of type 2 diabetes**
Meeyoung Park, Amy Rumora, Lucy Hinder, Junguk Hur, Felix Eichinger, Matthias Kretzler, Eva L. Feldman.
 Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 158

Effects of hormonal changes and cigarette smoking on oral microbiome	
<u>Purnima Kumar</u> , Binnaz Leblebicioglu, <u>Akshay Paropkari</u> .	
Presenter affiliation: The Ohio State University, Columbus, Ohio.	159
Enhancing the utility and usability of gemini for rare and common disease research	
<u>Brent S. Pedersen</u> , Aaron R. Quinlan.	
Presenter affiliation: University of Utah, Salt Lake City, Utah.	160
Accurate and efficient transcript identification and quantification using RNA-Seq data	
<u>Mihaela Perte</u> a, Geo M. Pertea, Steven L. Salzberg.	
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.	161
NCBI's Genetic Variation Resources	
<u>Lon D. Phan</u> , NCBI Variation Working Group.	
Presenter affiliation: National Institutes of Health, Bethesda, Maryland.	162
Mixture models that estimate gene-expression activation on a single-sample basis for any expression platform	
<u>Stephen R. Piccolo</u> , Evan Johnson.	
Presenter affiliation: Boston University School of Medicine, Boston, Massachusetts.	163
Transcript differential analysis of RNA-Seq data with sleuth	
<u>Harold Pimentel</u> , Nicolas Bray, Pall Melsted, Lior Pachter.	
Presenter affiliation: UC Berkeley, Berkeley, California.	164
Genome-wide characterization of chromatin state plasticity	
<u>Luca Pinello</u> , Alexander Gusev, Hilary Finucane, Jialiang Huang, Alkes Price, Guo-Cheng Yuan.	
Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts.	165
New genetic approaches in patients with transposition of the great arteries	
<u>Alex Postma</u> , Fleur Tjong, Julien Barc, Barbara Mulder, Connie Bezzina.	
Presenter affiliation: Academic Medical Center, Amsterdam, Netherlands.	166

An automated data management system for hereditary cancer analysis in clinical diagnostics	
<u>Meera Prasad</u> , Aijazuddin Syed, Yan Wang, Mustafa Syed, Zhen Y. Liu, Donovan T. Cheng, Marc Ladanyi, Liying Zhang, Michael F. Berger, Ahmet Zehir.	
Presenter affiliation: Memorial Sloan-Kettering Cancer , New York, New York.	167
Boiler—A compression tool for BAM files supporting fast, accurate queries	
<u>Jacob Pritt</u> , Ben Langmead.	
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.	168
Annotating, maintaining, and curating RefSeq prokaryotic genomes	
<u>Kim D. Pruitt</u> , Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Stacy Ciufu, Daniel Haft, Wenjun Li, Kathleen O'Neill, Tatiana Tatusova.	
Presenter affiliation: National Institutes of Health, Bethesda, Maryland.	169
Variation in tissue-specific codon usage across four members of the Poaceae	
<u>Jane A. Pulman</u> , Megan J. Bowman, Kevin L. Childs.	
Presenter affiliation: Michigan State University, East Lansing, Michigan.	170
Using ERCC spike-ins and erccdashboard R package to assess performance of differential gene expression detection by Ion AmpliSeq™ Transcriptome assays	
<u>Rongsu Qi</u> , Srinka Ghosh, Tommie Lincecum.	
Presenter affiliation: Thermo Fisher Scientific, South San Francisco, California.	171
The role of alternative splicing and gene expression in diffuse intrinsic pontine gliomas	
<u>Arun K. Ramani</u> , Pawel Buczkowicz, Robert Siddaway, Man Yu, Yue Jiang, Patricia Rakopoulos, Cynthia Hawkins, Michael Brudno.	
Presenter affiliation: Hospital for Sick Children, Toronto, Canada.	172
microRNA target sites act as regulatory hotspots in 3'UTRs	
<u>Simon H. Rasmussen</u> , Mireya Plass, Anders Krogh.	
Presenter affiliation: University of Copenhagen, Copenhagen, Denmark.	173

Accurate prediction of breakpoints in sequences

Uma D. Paila, Chun-Song Yang, Bryce Paschal, [Aakrosh Ratan](#).
Presenter affiliation: University of Virginia, Charlottesville, Virginia.

174

Iso-Seq bioinformatics analysis with PacBio long reads

[Meisam Razaviyayn](#), David Tse.

Presenter affiliation: Stanford University, Stanford, California.

175

Highly accurate read mapping of third generation sequencing reads for improved structural variation analysis

[Philipp Rescheneder](#), Fritz J. Sedlazeck, Maria Nattestad, Arndt von Haeseler, Michael C. Schatz.

Presenter affiliation: Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Vienna, Austria.

176

Clarify and quantifying mechanisms of DSB formation using mathematical modelling and BLESS sequencing

Norbert Dojer, Jules Nde, Abhishek Mitra, Ji Li, Yea-Lih Lin, Anna Kubicka, Magdalena Skrzypczak, Nicola Crosetto, Magda Bienko, Ivan Dikic, Krzysztof Ginalski, Philippe Pasero, [Maga Rowicka](#).

Presenter affiliation: University of Texas Medical Branch, Galveston, Texas.

177

Normalizing variation between open-access variation datasets

[Gary I. Saunders](#), Dylan Spalding, Franciso J. Lopez, Jag Y. Kandasamy, Cristina Y. Gonzalez, Justin Paschall.

Presenter affiliation: European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge, United Kingdom.

178

The Galaxy HiC browser—An interactive multi-dimensional genome topology browser and data repository

[Michael E. Sauria](#), Carl Eberhard, James Taylor.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

179

Quantitative proteogenomics application to personalised proteomics

[Christoph Schlaffner](#), Graham Ritchie, Theodoros Roumeliotis, Julia Steinberg, Christine Le Maitre, Mark Wilkinson, Eleftheria Zeggini, Jyoti Choudhary.

Presenter affiliation: Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

180

Updating approaches to reference assembly curation

Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Paul Flicek, Richard Durbin.

Presenter affiliation: NCBI, Bethesda, Maryland. 181

Regulatory variation in mice with diverse responses to environmental stimuli is driven by transposable element variation and environmentally induced chromatin remodeling at tissue specific transcription factor binding sites

Juan Du, Amy Leung, Candi Trac, Brian Parks, Aldons J. Lusis, Rama Natarajan, Dustin E. Schones.

Presenter affiliation: City of Hope, Duarte, California. 182

Detection of structural variants using third generation sequencing

Fritz J. Sedlazeck, Maria Nattestad, Michael C. Schatz.

Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 183

Reference-masked RNA-Seq assembly in human tissues relevant for cardiovascular disease and Type 2 diabetes reveals unannotated transcripts at the chr9p21 GWAS candidate locus

Shurjo K. Sen, Anna E. Sappington, Cihan Oguz, Adam R. Davis, Gary H. Gibbons.

Presenter affiliation: NIH, Bethesda, Maryland. 184

Optimizations of physical genome map contiguity by *in silico* ligation

Palak Sheth, Eva Chan, Alex Hastie, Andy Pang, Thomas Anantharaman, Erik Holmlin, Zeljko Dzakula, Xiang Zhou, Edward Cho, Vanessa Hayes, Han Cao.

Presenter affiliation: BioNano Genomics, San Diego, California. 185

Highly dynamic expansions of antimicrobial loci among *Medicago truncatula* accessions are revealed by inclusion of SMRT sequencing

Jason R. Miller, Peng Zhou, Joann Mudge, Thiru Ramaraj, Brian Walenz, Peter Tiffin, Nevin D. Young, Kevin A. Silverstein.

Presenter affiliation: University of Minnesota, Minneapolis, Minnesota. 186

Duplex sequencing for low allele frequency detection

Angad P. Singh, Matt Hims, Alina Raza, Rebecca Leary, Wendy Winckler, Derek Chiang.

Presenter affiliation: Novartis Institutes for Biomedical Research, Cambridge, Massachusetts. 187

- SomVarIUS—Somatic variant identification from unpaired tissue samples**
Kyle S. Smith, Vinod K. Yadav, Shanshan Pei, Daniel A. Pollyea, Craig T. Jordan, Subhrajyoti De.
 Presenter affiliation: University of Colorado, Aurora, Colorado. 188
- Gramene—Comparative plant genomics and pathway resources**
Joshua Stein, Wei Sharon, Justin Preece, Sushma Naithani, Andrew Olson, Yinping Jiao, Joseph Mulvaney, Sunita Kumari, Kapeel Chougule, Justin Elser, Bo Wang, James Thomason, Marcela K. Tello-Ruiz, Peter D'Eustachio, Robert Petryszak, Paul Kersey, Pankaj Jaiswal, Doreen Ware.
 Presenter affiliation: CSHL, Cold Spring Harbor, New York. 189
- Profiling protein occupants of the genome—Is TF footprinting ready for prime time?**
Myong-Hee Sung, Songjoon Baek, Gordon Hager.
 Presenter affiliation: National Institutes of Health, Baltimore, Maryland. 190
- De novo mutations induced by multiple DNA double strand breaks are revealed by whole-genome sequencing of *Arabidopsis thaliana***
Hidenori Tanaka, Nobuhiko Muramoto, Kazuto Kugou, Arisa Oda, Takahiro Nakamura, Kunihiro Ohta, Norihiro Mitsukawa.
 Presenter affiliation: Toyota Central R&D Labs., Inc., Nagakute, Aichi, Japan. 191
- Increasing discoverability and connectivity of scientific media through annotation with iCLiKVAL**
Todd D. Taylor, Naveen Kumar.
 Presenter affiliation: RIKEN, Yokohama, Japan. 192
- Exploring database options for storage of diverse DNA sequence variation datasets**
Jamie K. Teer, Richard Z. Liu, Guillermo Gonzalez-Calderon, Rodrigo Carvajal-Pelaez.
 Presenter affiliation: H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida. 193
- A comparative study of metagenomic analysis pipelines for accurate quantification of relative species abundance**
Yee Voan Teo, Alice Chu, Ian Pan, Andy Ly, Nicola Neretti.
 Presenter affiliation: Brown University, Providence, Rhode Island. 194

Galaxy Methylation Toolkit as a galaxy flavor

Nitesh Turaga, Enis Afgan, Benjamin Berman, James Taylor, Galaxy Team.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 195

200 Mammals—Sequence conservation at the single basepair level

Jason Turner-Maier, Jessica Alföldi, 200 Mammals Consortium, Jeremy Johnson, Voichita Marinescu, Hyun Ji Noh, Ross Swofford, Eva Murén, Chris P. Ponting, Gill Bejerano, Jussi Taipale, Oliver Ryder, Kerstin Lindblad-Toh.

Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 196

Identification of global regulators of T-helper cell lineages specification

Kartiek Kanduri, Subhash Tripathi, Antti Larjo, Henrik Mannerström, Ubaid Ullah, Riikka Lund, David Hawkins, Bing Ren, Harri Lähdesmäki, Riitta Lahesmaa.

Presenter affiliation: University of Turku, Turku, Finland. 197

An integrated analysis of the transcriptional response of human monocyte-derived macrophages to LPS

Annalaura Vacca, Stuart Aitken, Kenneth J. Baillie, David A. Hume, Colin A. Semple.

Presenter affiliation: MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom. 198

CRISPRscan—Designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*

Charles E. Vejnár, Miguel A. Moreno-Mateos, Jean-Denis Beaudoin, Juan P. Fernandez, Emily K. Mis, Mustafa K. Khokha, Antonio J. Giraldez.

Presenter affiliation: Yale University School of Medicine, New Haven, Connecticut. 199

SAPLING—A tool for customized network analysis focusing on psychiatric genetics

Wim Verleyen, Jesse Gillis.

Presenter affiliation: Cold Spring Harbor Laboratory, Woodbury, New York. 200

Unveiling the complexity of maize B73 transcriptome by single molecule long read sequencing <u>Bo Wang</u> , Elizabeth Tseng, Michael Regulski, Tyson Clark, Ting Hon, Yiping Jiao, Andrew Olson, Joshua Stein, Doreen Ware. Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.	201
Building a distributed system for sequence analysis <u>Liya Wang</u> , Peter Van Buren, Doreen Ware. Presenter affiliation: Cold Spring Harbor Labs, Cold Spring Harbor, New York; iPlant Collaborative, Tucson, Arizona.	202
Rapid, dynamic, and interactive visualization framework for pathogen identification in complex respiratory specimens from unexplained respiratory disease outbreak responses <u>Yuanbo Wang</u> , S S. Morrison, H P. Desai, J M. Winchell. Presenter affiliation: APHL, Silver Spring, Maryland; CDC, Atlanta, Georgia; GaTech, Atlanta, Georgia.	203
A novel approach to determining null models and controls for co-expression networks <u>Melanie Weber</u> , Jesse Gillis. Presenter affiliation: Cold Spring Harbor Laboratory, Woodbury, New York.	204
Decision tree-based method for integrating multi-domain data to identify childhood obesity disease endotypes <u>ClarLynda R. Williams-DeVane</u> , Michele Josey. Presenter affiliation: North Carolina Central University, Durham, North Carolina.	205
Scaling cancer subpopulation phylogeny reconstruction to thousands of tumors <u>Jeff A. Wintersinger</u> , Amit G. Deshwar, Quaid Morris. Presenter affiliation: University of Toronto, Toronto, Canada.	206
Genetic insights into juvenile idiopathic arthritis derived from deep whole genome sequencing <u>Laiping Wong</u> , Kaiyu Jiang, James N. Jarvis. Presenter affiliation: University at Buffalo, Buffalo, New York.	207
Kraken 2—Faster and more sensitive metagenomic classification <u>Derrick E. Wood</u> , Ben Langmead. Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.	208

NanoSim—Nanopore sequencing read simulator based on statistical characterization

Chen Yang, Justin Chu, René L. Warren, Inanç Birol.

Presenter affiliation: University of British Columbia, Vancouver, Canada; British Columbia Cancer Agency, Vancouver, Canada.

209

Cloud-based variant discovery using GenomeVIP

R. Jay Mashl, Kai Ye, Li Ding.

Presenter affiliation: Washington University in St Louis, St Louis, Missouri.

210

Probabilistic model for detecting mRNA translation efficiency changes from ribosome profiling

Yi Zhong, Theofanis Karaletsos, Philipp Drewe, Vipin Sreedharan, Kamini Singh, Hans-Guido Wendel, Gunnar Rätsch.

Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York.

211

AUTHOR INDEX

- Ab Halim, Mohd Amin, 147
 Abel, Laurent, 27, 111, 112
 Aboukhalil, Robert, 28
 Abramowicz, Marc, 84
 Aburatani, Hiroyuki, 153
 Abyzov, Alexej, 46
 Adams, David J., 20
 Afgan, Enis, 195
 Agius, Phaedra, 92
 Ahn, Joo Wook, 47
 Aigrain, Louise, 35
 Aitken, Stuart, 198
 Alföldi, Jessica, 196
 Alford, Marley, 33
 Alla, Ravi, 1
 Alquicira-Hernández, José, 13
 Altschuler, Gabriel, 24
 Alzbutas, Gediminas, 48
 Amar, David, 49
 Amarasinghe, Vindhya, 115
 Amber, McDonald, 70
 Amiruddin, Nadzirah, 147
 Anantharaman, Thomas, 185
 Anderson De Lima Morais,
 David, 113
 Anderson, Ana C., 7
 Andersson, Leif, 138
 Andrade, Jorge, 54
 Antoniou, Eric, 33
 Arbaciauskas, Kestutis, 48
 Arkin, Adam, 135
 Arora, Kanika, 92
 Artyomov, Maxim N., 72
 Asha, Nair, 70
 Assmann, Sarah M., 22
 Astromskas, Eimantas, 48
 Aswad, Luay, 51
 Atwal, Gurinder S., 28
 Aury, Jean-Marc, 149
 Azizi, Norazah, 147
- Babbs, Arran, 14
 Badretdin, Azat, 169
 Bae, Sena, 40
 Bae, Taejeong, 46
- Baek, Songjoon, 190
 Bailey, Timothy L., 20
 Baillie, Kenneth J., 198
 Balasubramanian,
 Gnanaprakash, 52
 Ballouz, Sara, 53
 Bandyopadhyay, Pradip K., 37
 Bang, Duhee, 110
 Bao, Riyue, 54
 Baras, Aris, 96
 Barbara, Crusan, 70
 Barc, Julien, 166
 Barrette, Michel, 113
 Bartha, Gabor, 1, 100
 Baslan, Timour, 28, 33
 Battistoni, Giorgia, 94
 Beaudoin, Jean-Denis, 199
 Beck, Timothy, 33
 Becq, Jennifer, 104
 Beightol, Mallory, 101
 Beiko, Robert G., 117
 Bejerano, Gill, 196
 Bendjilali, Nasrine, 148
 Benjamin, Kipp, 70
 Bentley, David, 104, 124
 Berger, Michael F., 167
 Bergman, Andre M., 154
 Berman, Benjamin, 195
 Berri, Stefano, 104
 Berriman, Matthew, 41
 Bevilacqua, Philip C., 22
 Bezzina, Connie, 166
 Bienko, Magda, 177
 Biesecker, Leslie G., 4, 109
 Biggs, Jason S., 37
 Birol, Inanç, 122, 209
 Bishop, Alexander, 90
 Blankenberg, Daniel, 55
 Blencowe, Benjamin, 95
 Boehm, Jesse, 136, 137
 Boehnke, Michael, 5
 Bonduelle, Maryse, 84
 Bonfield, James, 35
 Bontempi, Gianluca, 78
 Bourque, Guillaume, 113

Bowman, Megan J., 56, 170
 Boyle, Sean, 1
 Bray, Nicolas, 10, 164
 Breitwieser, Florian P., 45
 Bretaudeau, Anthony, 91
 Brors, Benedikt, 52
 Brover, Vyacheslav, 169
 Brown, Christopher D., 8
 Brudno, Michael, 23, 172
 Brumm, Riccardo, 57
 Buczkowicz, Pawel, 172
 Bujold, David, 113
 Busse, Birgit, 57

 Cain, Scott, 58
 Calvo, Sarah, 137
 Cao, Han, 185
 Carey, David J., 96
 Carlton, Jane M., 25
 Carr, Steve, 137
 Carvajal-Pelaez, Rodrigo, 193
 Casanova, Jean-Laurent, 27,
 111, 112
 Cech, Martin, 59
 Chahrokh-Zadeh, Soheyla, 57
 Chan, Esther T., 108
 Chan, Eva, 185
 Chan, Kuang-Lim, 147
 Chan, Pek-Lan, 147
 Chang, Gue Su, 6
 Chen, Jenny, 60
 Chen, Richard, 1, 100
 Chen, Tingwen, 61
 Chen, Yidong, 90
 Chen, Zhen-Xia, 62
 Chen, Zhi, 123
 Cheng, Donovan T., 167
 Chennagiri, Niru, 82
 Chernomorsky, Rostislav, 96
 Cherry, J M., 108
 Chervitz, Stephen, 1, 100
 Chetvernin, Vyacheslav, 169
 Chew, Guo-Liang, 21
 Chiang, Colby, 12
 Chiang, Derek, 187
 Childs, Kevin L., 56, 170
 Chilton, John, 59, 63
 Chin, Jason, 31, 33, 144

 Cho, Edward, 185
 Choudhary, Jyoti, 180
 Choufani, Sanaa, 23
 Chougule, Kapeel, 64, 157, 189
 Chu, Alice, 194
 Chu, Justin, 209
 Church, Deanna M., 1, 100
 Churchman, Michael, 150
 Churpek, Jane E., 54
 Cilia, Elisa, 84
 Ciufu, Stacy, 169
 Clark, Matt, 83
 Clark, Michael, 1
 Clark, Tyson, 201
 Clavijo, Bernardo J., 83
 Clement, Mark J., 50
 Clum, Alicia, 31
 Coghlan, Avril, 41
 Collado-Torres, Leonardo, 13,
 114
 Colleaux, Laurence, 146
 Cong, Le, 7
 Conrad, Donald F., 12
 Cooke, Daniel, 2
 Copeland, Alex, 31
 Corrigan, Adele, 47
 Cottingham, Robert W., 135
 Cotton, James, 41
 Cremer, Christopher, 65
 Criscione, Steven W., 19, 66
 Crosetto, Nicola, 177
 Cruaud, Corinne, 149

 D'Eustachio, Peter, 189
 Dabdoub, Shareef M., 39
 Dale, Ryan K., 67
 d'Alençon, Emmanuelle, 91
 Daneels, Dorian, 84
 Dang, Ha X., 6
 Das Roy, Rishi, 68
 Dashnow, Harriet, 69
 Davidson, Jean M., 108
 Davies, Kay E., 14
 Davies, Robert, 35
 Davila, Jaime I., 70
 Davis, Adam R., 4, 71, 184
 Davis, Carrie, 108
 De Cecco, Marco, 19, 66

de Jong, Jeroen, 154
 De, Subhajyoti, 3, 188
 Dean, Ralph A., 156
 Delas, M. Joaquina, 94
 Dennis, Jonathan H., 18
 Derr, Alan, 72
 Desai, H P., 203
 Deshwar, Amit G., 73, 206
 D'Eustachio, Peter, 115
 Dewey, Colin, 108
 Dewey, Frederick E., 96
 Dharmawardhana, Palitha, 115
 Di Palma, Federica, 83
 DiCuccio, Michael, 127, 169
 Diesh, Colin, 58, 75
 Dikic, Ivan, 177
 Ding, Li, 80, 210
 Diray-Arce, Joann, 50
 Djebali, Sarah, 108
 Dobin, Alex, 74, 85, 108
 Doerfel, Max, 11
 Dojer, Norbert, 177
 Dreszer, Tim, 108
 Drewe, Philipp, 211
 Du, Juan, 182
 Dubé, Marie-Pierre, 141, 142
 Duncavage, Eric J., 6
 Dunn, Nathan A., 75
 Durbin, Richard, 35, 181
 Dzakula, Zeljko, 185

Eberhard, Carl, 179
 Ebert, Betina, 57
 Eck, Sebastian H., 57
 Eichinger, Felix, 158
 Eisenhaure, Thomas, 137
 El Demerdash, Osama, 94
 Ellis, Matthew, 6
 Elser, Justin, 189
 Elsik, Christine, 75
 Ence, Daniel, 76
 Eng, Simon W., 77
 Engelen, Stefan, 149
 Engelhardt, Barbara E., 8
 Eric, Klee, 70
 Erlich, Yaniv, 16
 Escalière, Bertrand, 78
 Eyer, Alexander W., 156

Faciatori, Ilaria, 94
 Fadra, Numrah, 70
 Fan, Guangyi, 138
 Fang, Han, 11
 Farrell, Catherine M., 87
 Fawkes, Angie, 150
 Feldman, Eva L., 158
 Ferguson-Smith, Anne C., 20
 Fernandez, Juan P., 199
 Féron, François, 146
 Fetterolf, Samantha N., 96
 Ficklin, Stephen, 58
 Finucane, Hilary, 165
 Firth, Alastair, 133
 Fischer, Matthias G., 97
 Fiskerstrand, Torunn, 26
 Flicek, Paul, 181
 Foltz, Steven M., 80
 Fong, Nova, 124
 Förster, Frank, 97
 Freeberg, Mallory A., 81
 Freed, Stefan D., 43
 Freimer, Nelson, 5
 Friedberg, Iddo, 43, 121
 Frieden, Alexander, 82
 Fulton, Robert S., 5

Ganesan, Sukirth M., 39
 Gao, Ge, 62
 Gao, Yuan, 114
 Garber, Manuel, 60, 72, 133
 Garbulowski, Mateusz, 26
 Garcia Accinelli, Gonzalo A., 83
 Garcia, Sarah, 1
 Garvin, Tyler, 28
 Gasser, Robin B., 44
 Gaye, Amadou, 4
 Gazzo, Andrea, 84
 Gentleman, Robert, 88
 Gepner, Bruno, 146
 Getz, Gad, 136
 Ghim Siong, Ow, 51
 Ghosh, Srinka, 171
 Gibbons, Gary H., 4, 71, 184
 Gilissen, Christian, 26
 Gillis, Jesse, 53, 200, 204
 Ginalski, Krzysztof, 177
 Gindin, Yevgeniy, 6

Gingeras, Thomas R., 74, 85,
 108
 Giraldez, Antonio J., 199
 Girgis, Hani Z., 86
 Glau, Laura, 89
 Gliner, Genna, 8
 Goldfarb, Tamara, 87
 Goldstein, Leonard D., 88
 Gonnella, Giorgio, 89
 Gonzalez, Cristina Y., 178
 Gonzalez-Calderon, Guillermo,
 193
 Goodwin, Sara, 33, 139
 Gorthi, Aparna, 90
 Gotoh, Tetsuo, 130
 Gouin, Anaïs, 91
 Gourley, Charlie, 150
 Grabowska, Ewa A., 92
 Graveley, Brenton, 108
 Graves-Lindsay, Tina, 181
 Greene, Casey S., 93
 Griffith, Malachi, 6
 Griffith, Obi L., 6
 Gronau, Ilan, 42
 Grüning, Björn, 59
 Guigo, Roderic, 108
 Gul, Bilquees, 50
 Gulko, Brad, 42
 Gunnarsson, Ulrika, 138
 Gurtowski, James, 33, 94, 139
 Gusev, Alexander, 165
 Gustafson, Ulla, 138
 Gymrek, Melissa, 16

 Ha, Kevin, 95
 Habegger, Lukas, 96
 Hackl, Thomas, 97
 Hacohen, Nir, 137
 Haft, Daniel, 169
 Hager, Gordon, 190
 Hahn, William, 136
 Hait, Tom, 49
 Hall, Ira M., 12
 Hallikas, Outi, 68
 Hammell, Molly, 99, 116, 151
 Hammond, John H., 93
 Hanna, Jacob H., 60
 Hannon, Gregory J., 94

 Hansen, Nancy F., 98
 Hao, Yuan, 99
 Harris, Jason, 1, 100
 Harsha, Bhavana, 41
 Hartley, Stephen W., 15
 Hasselbacher, Verena, 57
 Hastie, Alex, 185
 Hattori, Masahira, 126
 Hause, Ronald J., 101
 Hauser, Fritz E., 102
 Havrilla, Jim, 103
 Haw, Robin, 115
 Hawkins, Cynthia, 172
 Hawkins, David, 197
 Hayes, Vanessa, 185
 He, Fuhong, 54
 He, Miao, 104
 Heavens, Darren, 83
 Hebrard, Maxime, 105
 Helaers, Raphael, 78, 106
 Helman, Elena, 1
 Hickman, Mark J., 148
 Hicks, James, 28, 33
 Hicks, Stephanie C., 107
 Hide, Winston, 24
 Hill, Andrew, 9
 Hims, Matt, 187
 Hinder, Lucy, 158
 Hitz, Benjamin C., 108
 Hoffman, Michael M., 20
 Hogan, Deborah A., 93
 Hogan, Michael, 147
 Höglund, Jacob, 138
 Hoischen, Alexander, 26
 Holdhus, Rita, 26
 Hollis, Robb, 150
 Holmes, Ian, 75
 Holmlin, Erik, 185
 Holroyd, Nancy, 41
 Hon, Ting, 201
 Hong, Celine, 109
 Höppner, Marc, 138
 Horn, Heiko, 136
 Houge, Gunnar, 26
 Howe, Kerstin, 181
 Hu, Jessica, 136
 Huang, Jiali, 165
 Huang, Lei, 54

Hubisz, Melissa J., 42
 Huh, Sunghoon, 110
 Hume, David A., 198
 Hupalo, Daniel N., 25
 Hur, Junguk, 158
 Hutter, Barbara, 52
 Hutton, Elizabeth, 33
 Hwang, Alan, 66
 Hwang, Byungjin, 110
 Hyde, Thomas M., 114

Ilic, Nina, 136
 Irizarry, Rafael A., 107
 Ishii, Kazuo, 130
 Itan, Yuval, 27, 111, 112
 Ivshina, Anna V., 51
 Izraeli, Shai, 49

Jackson, Anne U., 5
 Jackson, David, 35
 Jacoby, Meagan A., 6
 Jacques, Pierre-Étienne, 113
 Jaff, Jake, 137
 Jaffe, Andrew E., 13, 114
 Jain, Nehil, 117
 Jaiswal, Pankaj, 115, 189
 Jang, Hoon, 110
 Jarvis, James N., 207
 Jenster, Guido, 154
 Jernvall, Jukka, 68
 Jex, Aaron R., 44
 Jia, Yankai, 114
 Jiang, Kaiyu, 207
 Jiang, Yue, 172
 Jiao, Yinping, 144, 189, 201
 Jin, Jen, 70
 Jin, Ying, 99, 116, 151
 John, Titus, 156
 Johnson, Evan, 163
 Johnson, James, 20
 Johnson, Jeremy, 196
 Johnston, Jennifer, 109
 Jonassen, Inge, 26
 Jones, Aled R., 47
 Jongeneel, Victor, 79
 Jordan, Craig T., 188
 Josey, Michele, 205
 Joshi, Jyoti, 117

Kadarmideen, Haja N., 118, 119
 Kadener, Sebastian, 72
 Kadri, Sabah, 60, 72
 Kahles, Andre, 120
 Kalra, Gurmannat, 148
 Kamburov, Atanas, 136
 Kandasamy, Jag Y., 178
 Kanduri, Kartiek, 197
 Karaletsos, Theofanis, 211
 Karbhari, Nishika, 13
 Karro, John, 121
 Kashani, Alireza, 136
 Katou, Yuki, 153
 Keane, Thomas, 35
 Kechris, Katerina, 29
 Kelley, David R., 17
 Kendall, Jude, 28
 Kerje, Susanne, 138
 Kerrie, Barry, 31
 Kersey, Paul, 189
 Kevin, Halling, 70
 Khan, Hamza, 52, 122
 Khan, M Ajmal, 50
 Khokha, Mustafa K., 199
 Kim, Daehwan, 32, 45
 Kim, Eejung, 136
 Kim, Hyunmin, 124
 Kim, Jihye, 124
 Kim, Junho, 125
 Kim, Sangwoo, 125
 Kim, Seok-Won, 126
 Kimchi, Avi, 127
 Kimura, Hiroshi, 153
 Kindratenko, Volodymyr, 79
 King, Dan, 109
 Kitts, Paul, 127
 Kico, Jeffrey, 6
 Klein, Hanns-Georg, 57
 Kleinman, Joel E., 114
 Knight, James, 128
 Koboldt, Daniel C., 5
 Koenig, Jeremy E., 117
 Kogelman, Lisette J., 118, 119
 Koren, Sergey, 36, 129
 Kovaka, Sam, 85
 Kozaki, Toshinori, 130
 Kramer, Melissa, 33
 Kreiling, Jill A., 19

Kretzler, Matthias, 158
 Krishnan, Arjun, 131, 140
 Krogh, Anders, 173
 Kubicka, Anna, 177
 Kuchroo, Vijay K., 7
 Kucuk, Erdi, 132
 Kucukural, Alper, 133
 Kuecuk, Sandra, 57
 Kugou, Kazuo, 191
 Kumar, Naveen, 134, 192
 Kumar, Purnima, 39, 159
 Kumar, Vivek, 135
 Kumari, Sunita, 135, 189
 Kurki, Mitja, 5
 Kurtulus, Sema, 7
 Kurtz, Stefan, 89
 Kuznetsov, Vladimir A., 51

Laakso, Markku, 5
 Labun, Kornel, 26
 Ladanyi, Marc, 167
 Ladurner, Peter, 94
 Lage, Kasper, 136, 137
 Lagunavicius, Arunas, 48
 Lähdesmäki, Harri, 197
 Lahesmaa, Riitta, 123, 197
 Lamichhane, Sangeet, 138
 Langmead, Ben, 13, 168, 208
 Laperle, Jonathan, 113
 Larjo, Antti, 197
 Lash, Alex, 131
 Lawrence, Michael, 136
 Layer, Ryan M., 12
 Lazutka, Juozas, 48
 Le Maitre, Christine, 180
 Leader, Joseph B., 96
 Leary, Rebecca, 187
 Leblebicioglu, Binnaz, 159
 Ledbetter, David H., 96
 Lee, Cheng-Yang, 61
 Lee, Hayan, 139
 Lee, James, 88
 Lee, Jeong Ho, 125
 Lee, Junehawk, 125
 Lee, Shaun W., 43
 Lee, Young-suk, 140
 Leek, Jeffrey T., 13, 114
 Lefterova, Martina, 1

Legault, Marc-André, 141, 142
 Legeai, Fabrice, 91
 Lemainque, Arnaud, 149
 Lemaitre, Claire, 91
 Lemieux Perreault, Louis-Philippe, 141, 142
 Lenaerts, Tom, 84
 Lenz, Florian, 143
 Leung, Amy, 182
 Lewis, Suzanna E., 75
 Ley, Timothy, 6
 Li, Ji, 177
 Li, Qing, 37
 Li, Taibo, 137
 Li, Wenjun, 169
 Liang, Shoudan, 144
 Liberzon, Arthur, 137
 Lieber, Daniel, 82
 Liefeld, Ted, 137
 Lim, EunCheon, 145
 Lim, Jae Seok, 125
 Lin, Yea-Lih, 177
 Lincecum, Tommie, 171
 Lindblad-Toh, Kerstin, 196
 Linebaugh, Brian, 1
 Liu, Richard Z., 193
 Liu, Xin, 138
 Liu, Zhen Y., 167
 Liyakat Ali, Tharvesh Moideen, 146
 Locke, Adam E., 5
 Long, Manyuan, 62
 Lopez, Alexander, 96
 Lopez, Francis J., 178
 Low, Leslie Eng-Ti, 147
 Lu, Aiping, 37
 Lu, Charles, 6
 Lund, Riikka, 197
 Lunter, Gerton, 2, 30
 Luo, Shujun, 1
 Luo, Zunping, 25
 Lusion, Aldons J., 182
 Ly, Andy, 194
 Lyon, Gholson J., 11

Macleod, Samuel, 148
 Madoui, Mohammed-Amin, 149
 Madsen, Majbritt B., 119

Maeng, Ju Heon, 125
 Maglott, Donna R., 87
 Maher, Christopher, 6
 Maintz, Jens, 83
 Mainzer, Liudmila S., 79
 Makhlouf, Mélanie, 146
 Mannerström, Henrik, 197
 Marcus, Shoshana, 139
 Mardis, Elaine, 6
 Marinescu, Voichita, 196
 Marquez, Claudia P., 76
 Marshall, Christoph, 57
 Martienssen, Robert A., 147
 Mashl, R. Jay, 210
 Maslov, Sergei, 135
 Matsuda, Tomoko, 130
 Maxwell, Evan K., 96
 Mayer, Karin, 57
 Maza, Itay, 60
 McCombie, W. Richard, 33, 94,
 139
 McDowell, Ian, 8
 Mckay, Sheldon, 115
 McLysaght, Aoife, 38
 McPherson, John, 33
 Melsted, Páll, 10, 164
 Meltz Steinberg, Karyn, 5
 Mercer, John, 137
 Merowsky, Nicholas, 133
 Mesirov, Jill, 137
 Meynert, Alison, 150
 Miller, Christopher A., 6
 Miller, Jason R., 186
 Mis, Emily K., 199
 Mitra, Abhishek, 177
 Mitreva, Makedonka, 41
 Mitsukawa, Norihiro, 191
 Mituyama, Toutai, 153
 Moin, Mohd M., 123
 Molik, David, 99, 116, 151
 Montana, Aldrin, 1
 Moreno-Mateos, Miguel A., 199
 Morra, Massimo, 1
 Morris, Quaid, 65, 73, 77, 95,
 155, 206
 Morrison, S S., 203
 Morton, James T., 43
 Mudge, Joann, 186
 Mulder, Barbara, 166
 Muller, Olaf, 40
 Mullikin, James C., 15, 98, 109
 Mulvaney, Joseph, 152, 157, 189
 Mundo, Antonio F., 115
 Munoz-Torres, Monica C., 75
 Muramoto, Nobuhiko, 191
 Murén, Eva, 196
 Murphy, Lee, 150
 Murphy, Terence, 87, 127
 Nagappan, Jayanthi, 147
 Naithani, Sushma, 115, 189
 Nakamura, Takahiro, 191
 Nakato, Ryuichiro, 153
 Natarajan, Rama, 182
 Nattestad, Maria, 33, 139, 176,
 183
 Nde, Jules, 177
 Neafsey, Daniel E., 25
 Neff, Ryan A., 4, 71
 Nellore, Abhinav, 13
 Neretti, Nicola, 19, 66, 194
 Nettetstad, Maria, 31
 Nevedomskaya, Ekaterina, 154
 Nevin, James, 7
 Ng, David, 109
 Ng, Karen, 33
 Nguyen, Lam Son, 146
 Nguyen, Thong, 88
 Nie, Kun, 155
 Nielsen, Brent L., 50
 Nik Mohd Sanusi, Nik Shazana,
 147
 Ning, Zemin, 35
 Noh, Hyun Ji, 196
 Nordell-Markovits, Alexei, 113
 Nunes, Cristiano C., 156
 Oda, Arisa, 191
 O'Dushlaine, Colm, 96
 Oguz, Cihan, 4, 71, 184
 Ohta, Kunihiro, 191
 Okagaki, Laura H., 156
 Oliver, Brian, 62
 Oliver, Peter L., 14

Olivera, Baldomero M., 37
 Olson, Andrew, 64, 152, 157, 189, 201
 Ondov, Brian D., 36
 O'Neill, Kathleen, 169
 Onel, Kenan, 54
 Ong-Abdullah, Meilina, 147
 Ooi, Leslie Cheng-Li, 147
 Oswald, Dayna M., 92
 Oshlack, Alicia, 69
 Overton, John D., 96

Pachter, Lior, 10, 164
 Packer, Alan, 131
 Packer, Jonathan S., 96
 Paila, Uma D., 174
 Palotie, Aarno, 5
 Pan, Ian, 194
 Pang, Andy, 185
 Pardo, Carlos A., 45
 Park, Meeyoung, 158
 Park, YoSon, 8
 Parks, Brian, 182
 Paropkari, Akshay, 159
 Paschal, Bryce, 174
 Paschall, Justin, 178
 Pasero, Philippe, 177
 Patterson, Nick, 16
 Pedersen, Brent, 3, 160
 Pei, Shanshan, 188
 Peluso, Paul, 31, 144
 Pembroke, William G., 14
 Perteau, Geo M., 161
 Perteau, Mihaela, 161
 Petersen, Kjell, 26
 Petryszak, Robert, 189
 Petti, Allegra, 6
 Phan, Lon D., 162
 Phillippy, Adam M., 36, 129
 Phillips, Robert A., 13
 Piccolo, Stephen R., 163
 Pimentel, Harold, 10, 164
 Pinello, Luca, 165
 Pirinen, Matti, 5
 Pita-Juarez, Yered, 24
 Plass, Mireya, 173
 Poddurturi, Nikhil R., 108
 Pollard, Katherine, 3

Pollyea, Daniel A., 188
 Ponting, Chris P., 14, 196
 Postma, Alex, 166
 Prasad, Meera, 167
 Preece, Justin, 115, 157, 189
 Price, Alkes, 165
 Pritchard, Colin C., 101
 Pritham, Ellen J., 76
 Pritt, Jacob, 168
 Pruitt, Kim D., 87, 127, 169
 Pulman, Jane A., 56, 170

Qi, Rongsu, 171
 Qiu, Xiaojie, 9
 Quail, Michael, 35
 Quarells, Rakale, 71
 Quinlan, Aaron, 34, 103, 160
 Quintana-Murci, Lluís, 111, 112

Rabani, Michal, 21
 Rakopoulos, Patricia, 172
 Ramani, Arun K., 172
 Ramaraj, Thiru, 186
 Ramensky, Vasily, 5
 Ramos, Olivia M., 94
 Rangwala, Sanjida H., 87
 Rank, David, 31, 144
 Rasche, Eric, 59
 Rasmussen, Henrik B., 119
 Rasmussen, Simon H., 173
 Rasool, Omid, 123
 Ratan, Aakrosh, 174
 Rath, Sabine, 57
 Rättsch, Gunnar, 120, 211
 Rautio, Sini, 123
 Rawls, John F., 40
 Raza, Alina, 187
 Razaviyayn, Meisam, 175
 Ream, David, 121
 Redon, Gloria, 79
 Regev, Aviv, 7, 60, 137
 Regulski, Michael, 201
 Reich, David, 16
 Reid, Jeffrey G., 96
 Ren, Bing, 197
 Renvoise, Elodie, 68
 Rescheneder, Philipp, 176
 Ribeiro, Diogo M., 41

Rinn, John L., 17
 Ripatti, Samuli, 5
 Ritchie, Graham, 180
 Robert, Marc-Antoine, 113
 Robinson, Samuel D., 37
 Rodgers-Melnick, Eli, 18
 Rosenbluh, Joseph, 137
 Rosli, Rozana, 147
 Ross, Mark, 104
 Rost, Imma, 57
 Rougeulle, Claire, 146
 Roumeliotis, Theodoros, 180
 Rowe, Laurence D., 108
 Rowicka, Maga, 177
 Rozenblatt-Rosen, Orit, 7
 Rubanova, Yulia, 23
 Rumilla, Kandelaria, 70
 Rumora, Amy, 158
 Ryan, Kevin J., 47
 Ryder, Oliver, 196
 Rye, Tzyvia, 150

Sabaliauskaite, Rasa, 48
 Safavi-Hemami, Helena, 37
 Sailsbry, Joshua K., 156
 Salipante, Stephen J., 101
 Salomaa, Veikko, 5
 Salzberg, Steven L., 32, 45, 161
 Sambanthamurthi, Ravigadevi, 147
 Sanderson, Lacey, 58
 Sappington, Anna E., 184
 Sasaki, Mark M., 54
 Saunders, Diane, 83
 Saunders, Gary I., 178
 Sauria, Michael E., 179
 Schaeffer, Carly, 121
 Scharer, Lukas, 94
 Schatz, Michael C., 11, 28, 31, 33, 94, 135, 139, 176, 183
 Schier, Alexander F., 21
 Schlaffner, Christoph, 180
 Schneider, Valerie A., 181
 Schones, Dustin E., 182
 Scott, Alexandra J., 12
 Scott, Laura J., 5
 Sedivy, John M., 19, 66
 Sedlazeck, Fritz J., 33, 176, 183

Semple, Colin, 150, 198
 Sen, Shurjo K., 4, 71, 184
 Sergushichev, Alexey A., 72
 Service, Sue, 5
 Seshagiri, Somasekar, 88
 Shaknovich, Rita, 3
 Shamir, Ron, 49
 Shang, Lei, 111, 112
 Sharon, Wei, 189
 Shendure, Jay, 101
 Sheth, Palak, 185
 Shin, Jooheon, 114
 Shirahige, Katsuhiko, 153
 Shishkin, Alexander A., 60
 Shresta, Yashaswi, 136
 Siddaway, Robert, 172
 Siepel, Adam, 42
 Silverstein, Kevin A., 186
 Singer, Meromit, 7
 Singh, Angad P., 187
 Singh, Kamini, 211
 Singh, Rajinder, 147
 Siquier-Pernet, Karine, 146
 Siranosian, Benjamin, 19, 66
 Siska, Charlotte J., 29
 Sjoberg, Marcela, 20
 Skinkyte-Juskiene, Ruta, 118
 Skrzypczak, Magdalena, 177
 Sloan, Cricket A., 108
 Smith, Kyle, 3, 188
 Smith, Ryan P., 12
 Smits, Guillaume, 84
 Snoek, Jasper, 17
 Sofeso, Christina, 57
 Song, Li, 45
 Spalding, Dylan, 178
 Sreedharan, Vipin, 211
 Stanley, Eleanor, 41
 Stawiski, Eric, 88
 Steen, Hein van der, 117
 Steen, Vidar, 26
 Stein, Joshua, 189, 201
 Stein, Lincoln, 58, 115
 Steinberg, Julia, 180
 Stelloo, Suzan, 154
 Steven, Smith W., 147
 Stevens, Rick, 135
 Stokoe, David, 88

Stokowy, Tomasz, 26
 Strattan, J Seth, 108
 Straub, Richard E., 114
 Strepetkaite, Dovile, 48
 Stringham, Heather M., 5
 Subramanian, Aravind, 137
 Sundaravadanam, Yogi, 33
 Sung, Myong-Hee, 190
 Svard, Staffan, 44
 Swofford, Ross, 196
 Syed, Aijazuddin, 167
 Syed, Mustafa, 167
 Sztromwasser, Pawel, 26

Tadych, Alicja, 131
 Taipale, Jussi, 196
 Tam, Oliver, 99, 116
 Tan, Jie, 93
 Tanaka, Forrest Y., 108
 Tanaka, Hidenori, 191
 Tang, Petrus, 61
 Tang, Yin, 22
 Tao, Ran, 114
 Tardif, Jean-Claude, 141
 Tatusova, Tatiana, 127, 169
 Taylor, James, 81, 179, 195
 Taylor, Todd D., 105, 126, 134, 192
 Teer, Jamie K., 193
 Tello-Ruiz, Marcela K., 189
 Teng, Mingxiang, 107
 Teo, Yee Voan, 194
 Thalmann, Doreen S., 138
 Theesfeld, Chandra, 131
 Thomason, James, 135, 152, 157, 189
 Thompson, Dawn, 137
 Thompson, John, 82
 Tiffin, Peter, 186
 Ting, Ngoot-Chin, 147
 Tjong, Fleur, 166
 Toussaint, Nora C., 92
 Trac, Candi, 182
 Trapnell, Cole, 9
 Treangen, Todd J., 36
 Tripathi, Subhash, 197
 Troyanskaya, Olga, 131, 140
 Tsai, Jason, 41

Tse, David, 175
 Tseng, Elizabeth, 33, 201
 Tsherniak, Aviad, 137
 Turaga, Nitesh, 195
 Turinsky, Andrei, 23
 Turner, Emily H., 101
 Turner-Maier, Jason, 196

Ullah, Ubaid, 123, 197
 Unni, Deepak, 75
 Uy, Geoff, 6

Vacca, Annalaura, 198
 Vaccarino, Flora M., 46
 Vacic, Vladimir, 92
 Valdivia, Raphael H., 40
 Vallot, Céline, 146
 Van Buren, Peter, 202
 Van Buren, Peter, 64
 van der Poel, Henk G., 154
 Van Dooren, Sonia, 78, 84
 van Leenders, Geert J., 154
 Vandervalk, Benjamin P., 122
 Vejnar, Charles E., 199
 Vera, Daniel L., 18
 Verleyen, Wim, 200
 Vieira De Magalhaes, Jurandir, 64
 Vikkula, Miikka, 106
 Viner, Coby, 20
 Vizoso, Dita B., 94
 Vogl, Ina, 57
 Voight, Benjamin F., 8
 von Haeseler, Arndt, 176

Wada, Youichiro, 153
 Walenz, Brian, 129, 186
 Walker, Nicolas, 20
 Walter, Matthew, 6
 Wang, Bo, 189, 201
 Wang, Chao, 7
 Wang, Liya, 64, 202
 Wang, Qianfei, 54
 Wang, Siruo, 13
 Wang, Yan, 167
 Wang, Yuanbo, 203
 Ware, Doreen, 64, 115, 135, 144, 152, 157, 189, 201, 202

Warren, René L., 122, 209
 Wasik, Kaja A., 94
 Weber, Melanie, 204
 Wei, Wenbin, 24
 Wei, Xintao, 108
 Weinberger, Daniel R., 114
 Weiser, Joel, 115
 Weksberg, Rosanna, 23
 Wendel, Hans-Guido, 211
 Wesolowski, Sergiusz, 18
 Wessels, Lodewyk F., 154
 West, John, 1
 White, Eric, 82
 Widemo, Fredrik, 138
 Wigler, Michael, 28
 Wilfert, Amy B., 12
 Wilkinson, Mark, 180
 Willems, Thomas, 16
 Williams-DeVane, ClarLynda R.,
 205
 Wilson, Richard, 5, 6
 Winchell, J M., 203
 Wincker, Patrick, 149
 Winckler, Wendy, 187
 Wintersinger, Jeff A., 206
 Wischmeyer, Corey, 147
 Wong, Aaron, 131
 Wong, Bang, 137
 Wong, Laiping, 207
 Wong, Sandi, 40
 Wood, Derrick E., 208
 Worstell, Elizabeth, 136
 Wu, Guanming, 115

 Xia, Junrong, 7
 Xiaoke, Wang, 70

 Yadav, Vinod, 3, 188
 Yandell, Mark, 37, 76
 Yang, Chen, 209
 Yang, Chun-Song, 174
 Yao, Eric, 75
 Yao, Victoria, 131
 Ye, Kai, 80, 210
 Yen, Jennifer, 1
 Yenamandra, Surya P., 51
 Yeung, Rae S., 77
 Yoo, Shinjae, 139

 Young, Nevin D., 186
 Yu, Man, 172
 Yuan, Guo-Cheng, 165

 Zeggini, Eleftheria, 180
 Zehir, Ahmet, 167
 Zhang, Huiyuan, 7
 Zhang, Liying, 167
 Zhang, Ran, 131
 Zhang, Shen-Ying, 27, 111, 112
 Zhang, Yong, 62
 Zhang, Yue, 19, 66
 Zhong, Yi, 211
 Zhou, Peng, 186
 Zhou, Xiang, 185
 Zhou, Xin, 94
 Zhu, XiaoPeng, 60
 Zukowski, Kacper, 117
 Zwart, Wilbert, 154

THE RUMSFELDIAN CHALLENGES OF DEVELOPING CLINICAL SEQUENCING TESTS.

Deanna M Church, Sarah Garcia, Jennifer Yen, Jason Harris, Stephen Chervitz, Aldrin Montana, Brian Linebaugh, Shujun Luo, Gabor Bartha, Elena Helman, Sean Boyle, Ravi Alla, Michael Clark, Martina Lefterova, Massimo Morra, John West, Richard Chen

Personalis, Inc., Menlo Park, CA

Diagnostic sequencing is on the rise as costs decrease and clinical utility increases. While assays, tools and data sets developed for research provide a reasonable starting point; additional effort is necessary to achieve clinical-grade sensitivity and robustness for these approaches. Improvements are needed across the entire process, including assay design, bioinformatics analysis and validation.

One notably persistent challenge is the lack of comprehensive ‘gold standard’ data sets. While groups such as Genome in a Bottle (GIAB) have made valuable resources available to the genomics community, these data sets are not without some shortcomings. In addition to not covering the entire genome, these reference sets have been developed on healthy individuals and the provided samples are of high quality. These reference standards are not representative of the variant types commonly seen in a clinical testing lab. Additionally, provided samples, especially in cancer testing, are not always of high quality. Thus, clinical labs need to reach beyond these samples and data sets in order to provide high quality validation that is reflective of real-world performance. Annotation and analysis pipelines must also be validated as these can have a significant impact on test results. A variant filtered out by a stringent scheme will not be reported. Validating variant annotation is particularly challenging, as there are no ‘gold standard’ validation sets. Using a test set of over 100 hand curated variants as well as data from ClinVar, we found several differences between 3 annotation tools, even when using the same underlying set of transcripts. During this talk I will discuss the approaches we have taken to develop and validate our clinical Mendelian and cancer tests.

HAPLOTYPE-BASED SOMATIC MUTATION CALLING IN HETEROGENEOUS CANCER SAMPLES

Daniel Cooke, Gerton Lunter

Wellcome Trust Centre for Human Genetics, Statistical and Population Genetics, Oxford, United Kingdom

Variant detection algorithms can usually be divided into two types - those that focus on common variants, and those that target somatic mutations. This is partly due to the greater challenge of somatic cancer calling, which in addition to detection and classification of germline mutations, must also be sensitive to low-frequency somatic mutations that occur in tumour samples. The task is made more difficult by the complex nature of cancer genomes, which may contain inter-tumour heterogeneity, intra-tumour heterogeneity (e.g. subclonality), and frequent gene duplications.

Recently, many common variant detection algorithms have converged on sophisticated haplotype-based methods, which have better behaviour around complex loci than single-site methods. In contrast, somatic callers often employ simpler coverage-based hypothesis testing or model comparison methods, and as a result often focus on either SNV or indel detection, or internally employ distinct methods for each. This can lead to confusion in which approach is best applied to real data and how to interpret the inferences made, which could explain why cancer researchers often prefer custom pipelines which use more familiar common variant callers instead of dedicated somatic callers.

We present a new program that implements the first haplotype-based somatic mutation detection algorithm. The haplotype-based model allows explicit modelling of sequencing and alignment errors. Our algorithm uses a Bayesian model and EM to jointly infer somatic and common SNV and indel variants in tumour-normal samples, which may include multiple dependent tumour samples. Additionally, the algorithm uses a novel heuristic 'haplotype tree' to propose candidate haplotypes, which helps avoid windowing artefacts which can be a problem in current haplotype based methods. The program has a common interface for common variant and somatic cancer calling, and inferences are presented in a uniform way.

We demonstrate the algorithm on 1000G data, and present comparative benchmarks against other somatic callers using the ICGC-TCGA DREAM Somatic Mutation Calling Challenge data.

SASE-HUNTER: A COMPUTATIONAL METHOD TO DETECT SIGNATURES OF ACCELERATED SOMATIC EVOLUTION IN NON-CODING REGIONS OF CANCER GENOMES

Kyle Smith¹, Vinod Yadav¹, Brent Pedersen^{1,2}, Rita Shakhovich³, Katherine Pollard⁴, Subhajyoti De¹

¹University of Colorado, Department of Medicine, Aurora, CO, ²University of Utah, Eccles Institute for Human Genetics, Salt Lake City, UT, ³Weill Cornell Medical College, Department of Medicine, New York, NY, ⁴University of California San Francisco, Gladstone Institutes and Department of Epidemiology and Biostatistics, San Francisco, CA

Oncogenic mutations outside protein-coding regions remain largely unexplored. Analyses of the TERT locus have indicated that non-coding regulatory mutations can be more frequent than previously suspected and play important roles in oncogenesis. So far, limited studies are under-way to identify recurrent mutations in promoters of known genes. And yet, functional mutations need not always be recurrent at the same base position (e.g. TP53 mutations are distributed throughout the gene), and non-coding mutations are no exceptions. Recurrence based detection methods are not designed to detect these alternative mutation signatures. Signature of accelerated somatic evolution (SASE) is one such novel mutation signature in non-coding regions that we recently reported. Genomic regions under accelerated evolution are those that accumulate an excess of mutations compared to that expected based on the background mutation rates. In mammalian evolution, human accelerated regions (HARs; regions that acquired significantly more substitutions than expected after divergence from the common ancestor with chimpanzees) were often found to have regulatory functions associated with human-specific attributes. We applied the concept to cancer, and developed a computational method, SASE-hunter to identify the signature of accelerated somatic evolution (SASE) in a genomic locus, and prioritized those loci that carried the signature in multiple cancer patients. Interestingly, even when an affected locus carried the signature in multiple individuals, the mutations contributing to SASE themselves were not necessarily recurrent at the base-pair resolution. In a pan-cancer analysis of 12 tumor types, we detected SASE in the promoters of known cancer genes such as MYC, BCL2, RBM5 and WWOX. SASEs in selected cancer gene promoters were associated with over-expression of these genes, and also correlated with the age of onset of cancer, aggressiveness of the disease and survival. Taken together, our work detects a hitherto under-appreciated and clinically important class of regulatory changes in cancer genomes.

PREDICTIVE MODELING OF CORONARY ARTERY CALCIFICATION USING DECISION TREES

Cihan Oguz¹, Shurjo K Sen¹, Amadou Gaye¹, Ryan A Neff¹, Adam R Davis¹, Leslie G Biesecker², Gary H Gibbons¹

¹National Human Genome Research Institute, Cardiovascular Disease Section, Bethesda, MD, ²National Human Genome Research Institute, Clinical Genomics Section, Bethesda, MD

Personalized medicine aims at using clinical and multi-omic data to enable informed medical decision making for disease treatment and prevention. This aim resonates with systems biology that advocates a holistic approach for studying biological systems. Instead of focusing on individual biological components, this holistic approach requires the integration of multiple biological components and disparate data sets into predictive models. The complexity of modeling disease phenotypes using data coming from different sources (e.g., clinical measurements, bioinformatics analysis, and demographic data) requires advanced computational tools. To this end, we took a systems approach and used decision tree ensembles for predictive modeling of an atherosclerosis marker, namely coronary artery calcification (CAC), utilizing multi-omic data from the ClinSeq® study. Using data from 16 control subjects with zero CAC score (based on the Agatston method) and 16 cases with high CAC scores (average score of 1411), we first built classification models by utilizing 43 clinical variables and the genotypes of 57 SNPs compiled from previous GWAS on CAC. The predictive accuracy was quantified by generating receiver operating characteristics (ROC) curves (for samples not used in model training) and computing the area under each curve (AUC). Models built using a combination of clinical and genotype data for predicting CAC reached an average AUC of 0.84, whereas only clinical (genotype) data based models reached an AUC of 0.79 (0.75) demonstrating the predictive improvement using combined data sets. Next, we identified 56 additional SNPs that predicted the CAC state with extremely high accuracy (AUC=0.9957). Interestingly, these SNPs have been previously associated with several CAC risk factors including BMI, waist-hip ratio, type 2 diabetes, serum cholesterol, fibrinogen as well as glycosylated hemoglobin levels. Finally, we utilized the RNA-Seq expression levels of a panel of gene markers of CAC and achieved an AUC of 0.95. By systematically perturbing our decision tree based models, we identified that CALU (a regulator of the vitamin K-dependent gamma-carboxylation system) and DDR1 (a regulator of cell proliferation, migration, and differentiation) are the only strong predictors of CAC for both extremely high and lower CAC levels. Our finding suggests a precursor role (leading to extremely high CAC levels) for CALU and DDR1. This finding and additional biological insights generated by our analysis illustrate the potential of using decision tree based models as accurate and informative multi-omic diagnostic tools for personalized medicine.

IDENTIFYING LOW FREQUENCY AND RARE CODING VARIATION INFLUENCING CARDIOMETABOLIC TRAITS THROUGH WHOLE EXOME SEQUENCING OF 20,000 FINNS: THE FINMETSEQ STUDY

Karyn Meltz Steinberg¹, Adam E Locke², Sue Service³, Vasily Ramensky³, Matti Pirinen⁴, Heather M Stringham², Anne U Jackson², Mitja Kurki⁵, Laura J Scott², Robert S Fulton¹, Daniel C Koboldt¹, Samuli Ripatti⁴, Veikko Salomaa⁴, Aarno Palotie^{4,5}, Markku Laakso⁶, Nelson Freimer³, Michael Boehnke², Richard K Wilson¹

¹Washington University, McDonnell Genome Institute, St. Louis, MO, ²University of Michigan, Department of Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, ³University of California Los Angeles, Los Angeles, CA, ⁴Finnish Institute for Molecular Medicine, Helsinki, Finland, ⁵Broad Institute of MIT and Harvard, Cambridge, MA, ⁶Kuopio University Hospital and University of Eastern Finland, Department of Medicine, Kuopio, Finland

Due to recent population bottlenecks, low frequency, deleterious variants are enriched in the Finnish population. The resulting increase in power and the on-going large-scale longitudinal population studies make them an excellent resource for studying the genetics of complex traits. To examine the genetic architecture of cardiometabolic traits in Finland, we sequenced the exomes of 20,000 individuals from two large population-based longitudinal studies: the METSIM and FINRISK studies. Using an integrated variant discovery and genotyping approach across three sites, we identified more over 1.5 million variable sites. After adjusting for relatedness, we performed single-variant and gene-based association tests for >50 phenotypes including lipids, glucose and insulin, anthropometrics, and blood pressure measures. We identified 46 associations at exome-wide significance ($p < 5 \times 10^{-7}$) at 26 loci in 9 different traits. We replicated low frequency non-synonymous variants in *PCSK9* (rs11591147) associated with decreased levels of total and LDL cholesterol and ApoB levels, in *HNF4A* (rs1800961) associated with lower HDL cholesterol, and a non-synonymous variant in *LPA* (rs3798220) at ~1% frequency associated with total cholesterol. We also identified a non-synonymous variant at <0.5% allele frequency in *LCAT* associated with significantly lower HDL cholesterol. Mutations in *LCAT* are also known to cause Fish-eye and Norum disease, both of which are characterized by significantly altered lipid profiles and corneal occlusions. The large-scale, but focused nature of this study allows us the unique opportunity to identify coding variation influencing cardiometabolic traits, and potentially to follow up candidate variants. For example, although the particular variant identified in *LCAT* has not been implicated in either Fish-eye or Norum disease, we can re-contact carriers to identify vision problems and further examine their lipid profiles.

ASSESSING TUMOR HETEROGENEITY AND TRACKING CLONAL CLEARANCE IN RESPONSE TO THERAPY

Christopher A Miller^{1,2}, Ha X Dang^{1,2}, Gue Su Chang¹, Allegra Petti^{1,2}, Jeffrey Klco³, Charles Lu¹, Eric J Duncavage⁴, Yevgeniy Gindin¹, Obi L Griffith^{1,2}, Malachi Griffith^{1,5}, Meagan A Jacoby², Geoff Uy², Christopher Maher^{1,2}, Matthew Ellis⁶, Matthew Walter², Timothy Ley^{2,5}, Elaine Mardis^{1,2,5}, Richard Wilson^{1,2,5}

¹Washington University School of Medicine, McDonnell Genome Institute, St Louis, MO, ²Washington University School of Medicine, Dept of Medicine, St Louis, MO, ³St Jude Children's Research Hospital, Dept of Pathology, Memphis, TN, ⁴Washington University School of Medicine, Dept of Pathology & Immunology, St Louis, MO, ⁵Washington University School of Medicine, Dept of Genetics, St Louis, MO, ⁶Baylor College of Medicine, Lester and Sue Smith Breast Center, Houston, TX

Most tumors contain subclonal populations that may expand or diminish in response to evolutionary pressures, such as cancer therapies. Accurately detecting and tracking these populations over time requires combining deep sequencing of DNA from multiple timepoints with computational approaches for variant detection, subclonal inference, and reconstruction of tumor phylogeny. We have developed a collection of tools for performing such analyses and present results from three studies that give insight into their application, effectiveness, and prognostic value.

First we demonstrate that the genomic landscape and clonal architecture of 22 primary breast tumors evolves during neoadjuvant aromatase inhibitor therapy. By sequencing pre- and post-treatment samples, we were able to resolve subclonal populations with resistance or response in 77% of patients. We then developed a quantitative metric of clonal instability that was used to categorize the cohort into three distinct response groups.

Next, we sequenced a cohort of 51 patients with myeloid malignancies treated with epigenetic modifiers sampled at up to 10 distinct timepoints over 1 year. This allowed multi-dimensional clustering to be used to resolve distinct subclonal populations. Ultra-deep sequencing using barcoded reads provided more accurate variant allele fractions, and also allowed us to track residual disease and clonal clearance with very high sensitivity.

Finally, we used a set of 50 patients with Acute Myeloid Leukemia to show that different subclonal populations may be more or less susceptible to treatment and these phenotypes are associated with specific driver mutations. We also used deep sequencing to assay tumor clearance after induction chemotherapy, and show that this measure was both prognostic and more sensitive than traditional morphologic assays of disease clearance.

Together, these results demonstrate the power of pairing deep sequencing with algorithms for clonal inference and highlight the necessity of understanding cancer in an evolutionary context.

IDENTIFYING NOVEL DRIVERS OF CD8+ T CELL EXHAUSTION IN TUMOR

Meromit Singer*¹, Chao Wang*², Le Cong*¹, Huiyuan Zhang², Sema Kurtulus², Junrong Xia², James Nevin², Orit Rozenblatt-Rosen¹, Vijay K Kuchroo^{#2}, Ana C Anderson^{#2}, Aviv Regev^{#1,3}

¹Broad Institute of MIT and Harvard, ., Cambridge, MA, ²Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, ³Howard Hughes Medical Institute, Massachusetts Institute of Technology, Department of Biology, Cambridge, MA

* - Co-first author

- Co-senior author

New cancer immunotherapies that stimulate dysfunctional (exhausted) T cells in the tumor environment hold great promise at inducing and maintaining cancer remission. However, there is still much variation in response that is not understood between individual patients and across cancer types, only a small portion of the system that regulates T cell dysfunction has been targeted clinically, and many of the components and interactions in the system remain poorly understood. The complexity of the underlying system and the many unknowns call for a systems-level approach, but studies of the molecular mechanisms underlying T cell exhaustion in tumors have been limited by the low input amounts available from in vivo models and clinical samples.

Here, we leveraged high-throughput small-input and single cell RNA-seq and computational analysis of tumor infiltrating lymphocytes (TILs) from in vivo tumors to pinpoint novel drivers and characteristics of dysfunctional CD8+ TILs. We defined a novel tumor-specific CD8+ exhaustion signature, discovering a role for a novel gene in CD8+ exhaustion, which we validated using mouse models. Subsequent analysis of RNA-Seq profiles of CD8+ T cells from tumors with and without dysfunctional TILs further refined the exhaustion signature, highlighting several transcription factors as novel candidate drivers of CD8+ exhaustion in tumor. We then adapted the CRISPR-Cas9 genome engineering system for editing of primary T cells, applied it to knockout these factors in CD8+ cells and used in vivo transfers of the Cas9-edited CD8+ cells to show that perturbation of the TFs significantly reduced tumor growth, validating their role as drivers of CD8+ exhaustion in tumor.

Our work identifies novel transcription factors that drive CD8+ exhaustion in mouse tumor models and highlights new potential targets for alteration with immunotherapies. Our general genomic, computational and CRISPR tools can be used to pinpoint key molecular mechanisms in other immune cells from in vivo tumor models.

USING MENDELIAN RANDOMIZATION TO INVESTIGATE ASSOCIATION BETWEEN GENE EXPRESSION VARIATION AND COMPLEX TRAITS

YoSon Park¹, Ian McDowell², Genna Gliner³, Benjamin F Voight^{1,4,5}, Barbara E Engelhardt⁶, Christopher D Brown¹

¹Perelman School of Medicine University of Pennsylvania, Dept. of Genetics, Philadelphia, PA, ²Duke University, Dept. of Computational Biology and Bioinformatics, Durham, NC, ³Princeton University, Dept. of Operations Research and Financial Engineering, Princeton, NJ, ⁴Perelman School of Medicine University of Pennsylvania, Dept. of Systems Pharmacology and Translational Therapeutics, Philadelphia, PA, ⁵Perelman School of Medicine University of Pennsylvania, Institute of Bioinformatics, Philadelphia, PA, ⁶Princeton University, Dept. of Computer Science and Center for Statistics and Machine Learning, Princeton, NJ

Genome-wide association studies (GWAS) have been immensely successful in identifying genetic variants associated with complex phenotypes. However such studies are unable to explain causality, nor do they provide insight into functional mechanisms as necessary for targeted therapeutic intervention. The majority of these variants is noncoding and may lead to changes in disease risk via changes in gene expression. Mendelian Randomization (MR) allows testing of causal relationships between biomarkers and disease while reducing the effects of reverse causality or confounding due to hidden covariates. However, to date, MR has largely been used in limited epidemiological settings. To facilitate causal inference for thousands of variants associated with gene expression (eQTL), we implemented a tool to analyze genome-wide eQTL data and GWAS summary statistics in an MR framework. As a proof of principle, we applied MR to a meta-analysis of blood serum metabolites from the Global Lipid Genetics Consortium using liver eQTLs previously identified by our group. We identified several known and novel associations with LDL-C levels including *SORT1*, *HMGCR*, *SLC44A2*, *ST3GAL4* and *ANGPTL3*. For example, one standard deviation (SD) change of *SORT1* expression increases LDL-C levels by 6.5 mg/dl ($p=1 \times 10^{-100}$), resulting in an increased risk for coronary heart disease ($OR=1.14$, $p=1 \times 10^{-9}$), replicating known mechanisms of *SORT1*. We also applied MR to eQTLs identified by the Genotype-Tissue Expression (GTEx) consortium across 44 cell types. This expanded catalog of cell types improves causal gene and cell type identification. Application to GTEx pilot data identified several additional genes meriting further investigation including *APOB* (LDL-C levels, $p=4 \times 10^{-11}$) and *G6PC2* (fasting glucose levels, $p=1 \times 10^{-36}$) in subcutaneous adipose tissues. In conclusion, we propose that GWAS and eQTL integration through MR is able to prioritize candidate disease genes and quantify the sensitivity of organismal phenotypes to changes in gene expression.

DIFFERENTIAL ANALYSIS OF BIFURCATING SINGLE-CELL GENE EXPRESSION TRAJECTORIES

Xiaojie Qiu, Andrew Hill, Cole Trapnell

University of Washington, Genome Sciences, Seattle, WA

Single-cell trajectory analysis is a powerful approach for studying gene regulatory changes during cell differentiation and other dynamic processes. Recently, we showed that individual cells can be ordered according to progress through differentiation by analyzing their transcriptomes with unsupervised algorithms. Previous studies by our group and others have been limited to linear trajectories tracking unipotent progenitor cells. Such cellular trajectories have only one outcome. However, during development, cells make fate decisions that lead to one of several mutually exclusive states in the adult. How to reconstruct and analyze single-cell trajectories that include and span fate decisions is an open problem.

Here, we describe an approach for reconstructing single-cell trajectories that include bifurcations corresponding to cell fate decisions. We then describe statistical methods for identifying genes that are differentially expressed between trajectory outcomes. We illustrate the power of this technique by analyzing differentiating bronchoalveolar progenitor cells undergoing specification into type I and type II pneumocytes. This analysis reveals hundreds of genes with lineage-dependent expression. Our approach, which encodes a topological description of the trajectory as continuous predictors in a generalized linear model, can distinguish, for example, genes that become lineage-dependent proximal to the fate specification from those that are restricted to a lineage later in differentiation. We conclude with an analysis of bifurcations in settings other than development to argue that single cell trajectory analysis can help pinpoint the genes that drive a process from those more downstream.

KALLISTO: NEAR-OPTIMAL RNA-SEQ QUANTIFICATION

Nicolas L Bray^{1,2,3}, Harold Pimentel⁴, Páll Melsted⁵, Lior Pachter^{3,4,6}

¹UC Berkeley, Innovative Genomics Initiative, Berkeley, CA, ²UC Berkeley, Center for RNA Systems Biology, Berkeley, CA, ³UC Berkeley, Department of Molecular & Cell Biology, Berkeley, CA, ⁴UC Berkeley, Department of Computer Science, Berkeley, CA, ⁵University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, Reykjavík, Iceland, ⁶UC Berkeley, Department of Mathematics, Berkeley, CA

In the past year, RNA-Seq projects have expanded in scope to include hundreds or even thousands of samples per experiment. The scale of data production is straining existing analysis pipelines; even the mapping of reads to transcripts takes hours or days on standard computing infrastructure.

With the program Sailfish, Patro and Kingsford introduced a novel k-mer based approach to quantification which is much faster than existing methods. However replacing reads with their constituent k-mers as in Sailfish leads to a significant loss in accuracy. We introduce a new approach that does not require mapping reads, but instead is based on the idea of "pseudoalignment". Instead of determining the exact matching locations of (paired) reads in transcripts, we show that it suffices to determine for each (paired) read which transcripts it is compatible with. Efficient hashing allows pseudoalignment to be ultrafast.

We have implemented these ideas in a program called kallisto, with which we are able to accurately estimate transcript abundances directly from 30 million human RNA-Seq reads in less than 5 minutes on a single core. Kallisto's speed and use of pseudoalignments rather than alignments is not only convenient, it enables new types of computations and applications. For example, with kallisto bootstrap estimates can be used to determine uncertainty of abundance estimates in genes with complicated isoform structure. Kallisto can also learn and correct for sequence specific bias that arises due to non-uniform priming. When evaluated on metagenomic reads, we found that kallisto can quantify abundances of genomes with more accuracy and equal speed compared to existing methods, and can scale to handle a reference set of over 2000 bacterial genomes.

To take advantage of these possibilities we have developed a companion analysis tool in R called sleuth that can fully utilize biological replicates and the bootstrap of kallisto to improve statistical accuracy in differential analysis. Together, kallisto and sleuth completely transform RNA-Seq analysis from a cumbersome, computationally intensive task requiring complex "pipelines", to a simple computation that can be performed in minutes on a laptop.

SCIKIT-RIBO: ACCURATE A-SITE PREDICTION AND ROBUST MODELING OF TRANSLATION CONTROL FROM RIBOSEQ AND RNASEQ DATA

Han Fang^{1,2,3}, Max Doerfel², Gholson J Lyon², Michael C Schatz¹

¹Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, ²Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY, ³Stony Brook University, Dept. of Applied Math and Statistics, Stony Brook, NY

Ribosome profiling (Riboseq) is a powerful technique for monitoring protein translation in vivo. Some have suggested that expression measurements generated by Riboseq better explain the variance of protein abundance, compared to RNAseq data alone. However, there are very few methods available to jointly analyze Riboseq and RNAseq data in a systematic and standardized fashion. It is also unclear how to determine the A-site location on a read since Riboseq does not provide this information directly.

Here, we present scikit-ribo, a statistical learning framework for joint analysis of Riboseq and RNAseq data. We provide modules for ribosome A-site prediction, shrinkage estimation of translation efficiency (TE), and stop-codon read-through detection. Based on reads occupying start codons, we used a SVM classifier to directly learn key features of A-site location along Riboseq reads. We observed strong dependencies of A-site location on both read length and codon offset due to variable effects of the digestion enzyme on individual ribosomes. We also show that after optimization of mapping, Cufflinks and Salmon are highly concordant on gene-level expression quantification on Riboseq data.

To demonstrate its effectiveness, we used scikit-ribo to analyze data from wild-type and knockout yeast strains involving *Naa10*, which is bound to the ribosome as part of the NatA complex. Our prediction method indeed has much higher accuracy for identifying A-site location than previous methods (0.86 vs. 0.64, 10-fold CV). The predicted 61-sense codon usage has a significant correlation with published estimates ($\rho=0.55$, $p\text{-value}=9\times 10^{-6}$). Both Riboseq and RNAseq data showed that the mutant has significant reduction in expression of conjugation/mating genes (BH adjusted $p\text{-value}: 4\times 10^{-13}$). Genes involved in multi-organism processes have a significant reduction of TE (BH adjusted $p\text{-value}: 5\times 10^{-6}$), and their transcripts are mostly down regulated. Scikit-ribo also identified 18 genes that have two-fold or more stop codon read-through in the mutant, relative to wild-type. GO analysis suggested these read-through events are enriched in cytosolic large ribosomal subunit genes (BH adjusted $p\text{-value}: 0.02$). Together, these results show that scikit-ribo provides novel insights into the role of ARD1 in translational control. Ongoing work includes joint modeling of the translation initiation and elongation rates, which could yield a robust inference of protein synthesis rate.

DETECTION AND INTERPRETATION OF GENOME STRUCTURAL VARIATION IN GTEx SAMPLES

Colby Chiang¹, Ryan M Layer², Ryan P Smith¹, Alexandra J Scott¹, Amy B Wilfert³, Donald F Conrad³, Ira M Hall^{1,4}

¹Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, ²University of Utah, Department of Human Genetics, Salt Lake City, UT, ³Washington University School of Medicine, Department of Genetics, St. Louis, MO, ⁴Washington University School of Medicine, Department of Medicine, St. Louis, MO

Structural variation (SV) is a broad class of genome variation that includes copy number variants (CNVs), balanced rearrangements and mobile element insertions. SV is recognized to be an important source of human genetic diversity – 5,000-10,000 SVs are detectable in the typical human genome using short-read DNA sequencing technologies – but little is known about the mechanisms through which SVs affect gene expression and phenotypic variation. The availability of deep whole genome sequencing (WGS) and RNA expression data in the GTEx cohort offers an unprecedented opportunity to address this question.

Here, we describe our work aimed at comprehensive detection and interpretation of structural variation in GTEx samples. We first developed an improved pipeline for SV detection in large WGS cohorts that is fast, accurate, scalable to thousands of genomes, and produces multi-sample callsets that are on par with traditional joint variant calling approaches. This pipeline is loosely based on our SpeedSeq software and is composed of four stages: 1) SV discovery on each individual genome using the LUMPY algorithm, 2) SV integration across all samples to produce a unified, cohort-level VCF of spatially refined breakpoints, 3) SV breakpoint genotyping with SVTyper, and 4) read-depth copy number annotation with CNVnator.

We applied these methods to 147 GTEx WGS datasets to generate a unified, cohort-level VCF of 49,730 SVs along with copy number and genotype annotations. Using this dataset, we mapped SV eQTLs in 13 tissues analyzed by RNA-seq, resulting in 1,596 SVs affecting the expression of 1,801 genes, including 506 genes that were not identified by eQTL mapping using SNVs and indels alone. We will present the results of our ongoing analysis of the tissue specificity and directionality of these expression effects with respect to SV class and overlap with known regulatory elements, and our efforts to identify and study causal SVs. We further describe analyses aimed at discerning the contribution of difficult-to-identify SVs that are not typically included in functional studies including rare variants, complex variants, multi-allelic CNVs, and mobile element insertions.

AN ANALYSIS OF SPLICING VARIATION ACROSS THE SEQUENCE READ ARCHIVE WITH RAIL-RNA

Abhinav Nellore^{1,2,3}, Leonardo Collado-Torres^{1,3,4}, José Alquicira-Hernández⁶, Siruo Wang^{1,3,7}, Robert A Phillips^{1,3,8}, Nishika Karbhari^{1,3,9}, Andrew E Jaffe^{1,3,4,5}, Ben Langmead^{1,2,3}, Jeffrey T Leek^{1,3}

¹Johns Hopkins University, Department of Biostatistics, Baltimore, MD, ²Johns Hopkins University, Department of Computer Science, Baltimore, MD, ³Johns Hopkins University, Center for Computational Biology, Baltimore, MD, ⁴Johns Hopkins Medical Campus, Lieber Institute for Brain Development, Baltimore, MD, ⁵Johns Hopkins University, Department of Mental Health, Baltimore, MD, ⁶National Autonomous University of Mexico, Undergraduate Program on Genomic Sciences, Mexico City, Mexico, ⁷Centre College, Department of Mathematics and Computer Science, Danville, KY, ⁸Salisbury University, Department of Biological Sciences, Salisbury, MD, ⁹University of Texas, College of Natural Sciences, Austin, TX

We exhibit Rail-RNA, software that analyzes many RNA sequencing (RNA-seq) samples and is easy to deploy on rented computer clusters in the cloud spanning thousands of processing cores using Amazon Elastic MapReduce. With one or two commands, Rail-RNA downloads, preprocesses, and aligns hundreds to thousands of RNA-seq samples in the cloud for under a dollar per sample, facilitating rapid and reproducible analysis. Rail-RNA's outputs are compatible with downstream Bioconductor packages such as derfinder, DESeq, and edgeR. We use Rail-RNA to study the breadth of splicing variation across approximately 20,000 RNA-seq samples sequenced on Illumina platforms and publicly available on the Sequence Read Archive (SRA). Some highlights of our analysis are that (1) canonical gene annotations like RefSeq and GENCODE capture exon-exon junctions overlapped by over 95% of spliced reads in most SRA samples, but only ~70% of exon-exon junctions, where each is found in at least 1,000 SRA samples, are annotated; (2) a random sample of K samples from SRA has junction content appearing in at least $R\%$ of the K samples that converges rapidly as K increases, suggesting a simple technique for automated curation of junctions input to modern alignment software that does not rely on current annotation; (3) up to 10% of samples on SRA have metadata fields (e.g., sex) that are incorrect. Rail-RNA is open-source software that can be run in the cloud by anyone who signs up for Amazon Web Services or on local hardware. It is available at <http://rail.bio> .

TEMPORAL TRANSCRIPTOMICS REVEALS DYSREGULATION OF TWIN-PEAKING GENES WHICH RESET THE CLOCK IN A MOUSE MODEL OF PSYCHIATRIC DISEASE

William G Pembroke, Arran Babbs, Kay E Davies, Chris P Ponting, Peter L Oliver

MRC Functional Genomics Unit, Department of Physiology Anatomy and Genetics, Oxford, United Kingdom

The mammalian suprachiasmatic nucleus (SCN) drives daily rhythmic behaviour and physiology, yet a comprehensive understanding of its coordinated transcriptional programmes is lacking. Furthermore there has been a long standing association between circadian disruption and psychiatric disease, for which the exact mechanisms linking the two are unknown. To analyse in detail the circadian transcriptome of the mammalian SCN and its variation in psychiatric disease, we carried out RNA-seq in mice over a 24-hour light / dark cycle in both wild-type and blind-drunk (*Bdr*), a mutant strain for Snap-25 that shows both schizophrenic endophenotypes and abnormal circadian behaviours. In addition to identifying 4,569 genes and 194 novel intergenic non-coding RNAs which exhibited a classic sinusoidal expression signature, co-expression network analysis identified a group of 766 genes, which unexpectedly peaked twice, near both the start and end of the dark phase; this twin-peaking group is significantly enriched for synaptic transmission genes that are crucial for light-induced phase-shifting of the circadian clock. This twin-peaking module is also dysregulated in *Bdr*, which may provide clues to the molecular pathways linking circadian disruption with psychiatric disease. Overall, our data allow us to propose that transcriptional timing in the SCN is gating clock resetting mechanisms and dysregulation of these genes may play a role in psychiatric disease.

JUNCTIONSEQ: DETECTING DIFFERENTIAL SPLICE JUNCTION USAGE VIA RNA-SEQ

Stephen W Hartley, James C Mullikin

National Human Genome Research Institute, Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD

Next generation RNA sequencing (RNA-Seq) can, in theory, provide unparalleled access to isoform-resolution expression information. However, the development of robust and efficient tools for detecting isoform-level differentials remains challenging.

Differential isoform regulation is a broad and inclusive term that encompasses a wide variety of different phenomena including the use of alternative promoters, alternative polyadenylation sites, cassette exons, mutually exclusive exons, alternative donors/acceptors, nonsense mediated decay, and intron retentions. As a result, even when the statistical methods are robust and accurate, the interpretation of the results is non-trivial. While numerous tools have been developed that claim to detect such differentials, the output of these tools is often counterintuitive and provides little to assist in the interpretation of the results. With the notable exception of DEXSeq, most tools provide little useful information to the end-user beyond the locus and the p-value.

We introduce JunctionSeq, which uses a variation of the already-well-established methods used by DEXSeq and DESeq2, extended and modified to allow testing for differential usage of both known and novel splice junction loci.

JunctionSeq is designed to be transparent and intuitive, and provides a number of powerful visualization tools designed to allow researchers to assess, evaluate, and characterize differential isoform regulation. JunctionSeq produces highly-customizable, publication-quality plots for individual genes of interest, and standard browser summary tracks for use with genome-wide viewers such as IGV or the UCSC genome browser.

We demonstrate the efficacy of JunctionSeq on an RNA-Seq experiment intended to detect neurally-regulated genes in the rat pineal gland. This experiment consisted of both in vivo comparisons between day and night conditions as well as in vitro comparisons between untreated pineal glands and glands treated with norepinephrine or dibutyryl cyclic AMP.

JunctionSeq detects several isoforms that exhibit consistent and replicable differential regulation. The differential isoform regulation of one of these genes is confirmed by qPCR, and the existence of several novel transcripts and splice variants is confirmed via long-read SMRT sequencing via the PACBIO RS II platform.

CHARACTERIZATION OF THE MICROSATELLITE MUTATION PROCESS AT EVERY LOCUS IN THE GENOME

Melissa Gymrek^{1,2,3,4}, Thomas Willems^{4,5}, Nick Patterson², David Reich^{2,6,7}, Yaniv Erlich^{4,8}

¹Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, ²Broad Institute of MIT and Harvard, Cambridge, MA, ³Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, ⁴New York Genome Center, New York, NY, ⁵Computational and Systems Biology Program, MIT, Cambridge, MA, ⁶Department of Genetics, Harvard Medical School, Boston, MA, ⁷Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, ⁸Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY

Microsatellites, or short tandem repeats (STRs), are comprised of repeating motifs of 1-6bp that span over 1% of the human genome and contribute to over 40 Mendelian disorders. These loci are prone to polymerase slippage during replication, resulting in mutation rates that are orders of magnitude higher than those of point mutations, conferring high evolvability. Revealing the mutation process of STRs is important to address questions in medical genetics, forensics, and population genetics. However, most studies of STR polymorphism have focused on a highly ascertained set of loci that are extraordinarily polymorphic and easy to genotype, providing biased mutation models. Here, we harnessed novel bioinformatics tools and an analytical framework to estimate the mutation rate at each STR in the human genome. First, we used our lobSTR algorithm to generate the most comprehensive STR polymorphism dataset to date, consisting of 1.5 million loci across 300 samples sequenced to high coverage by the Simons Genome Diversity Project. These samples originate from diverse genetic backgrounds and maximize our power to observe STR evolution across a wide range of time scales. Next, we developed an analytical model of the STR mutation process that imposes a length constraint on allele size and shows greater consistency than traditional stepwise models with empirical data. Our model allows us to obtain maximum likelihood estimates of mutation rate and length constraint parameters at each STR in the dataset by correlating genotypes with local sequence heterozygosity. We extensively tested this model on simulated and observed genotypes and accurately recovered mutation rates for markers with published *de novo* rates. Applying our method at each STR locus in the genome, we find that nearly all loci show evidence of a length constraint, and a subset of STRs show extremely high mutation rates of 10^{-2} /locus/generation. We used this call-set to analyze sequence determinants of STR variation, assess patterns of variation in coding regions, and scan for STRs under selective pressure.

BASSET—LEARNING THE REGULATORY CODE OF THE ACCESSIBLE GENOME WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

David R Kelley^{1,2}, Jasper Snoek³, John L Rinn^{1,2}

¹Harvard University, Stem Cell and Regenerative Biology, Cambridge, MA,

²Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, ³Harvard University, School of Engineering and Applied Science, Cambridge, MA

Practical applications for genomic science require research to develop several capacities. Importantly, researchers continue to map individual human genomes and statistically identify variants that influence phenotypes. However, our inability to effectively interpret variation in the noncoding genome limits the reach of this progress. Here, we demonstrate the power of machine learning to annotate every nucleotide variant of the genome with its influence on chromatin accessibility in any cell or tissue that has available data.

Machine learning approaches have proven effective for determining the sequence preferences of DNA binding proteins, histone modifications, and DNA accessibility. Recently, deep convolutional neural networks have achieved groundbreaking accuracies on many classic machine learning problems in image analysis and natural language processing. This technique maps easily to DNA sequence analysis; initial layers of the model learn to recognize sequence motifs and subsequent layers synthesize their spatial interactions into regulatory codes.

We applied deep convolutional neural networks to predict cell-specific DNA accessibility in chromatin from the underlying DNA sequence. Training and testing on a dataset of DNaseI hypersensitivity peaks mapped in 125 cells by the ENCODE consortium, we substantially exceeded the accuracy of the state-of-the-art support vector machines for this task.

Using the model, we can accurately predict the change in cell-specific accessibility conferred by any mutation in the genome. In accessible regions, we pinpoint the specific nucleotides most critical to maintaining an open chromatin state; these nucleotides have greater conservation than surrounding sequence. This approach also identifies mutations that would significantly increase accessibility, which can open repressed functional elements and affect nearby gene regulation.

We now have the ability to edit any nucleotide in the genome, but models to predict the outcome of the vast majority of those edits have fallen behind. Using our approach, researchers can perform a single genome-mapping experiment and simultaneously annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility. Functional genomics continues to hold great promise. To realize this potential, machine learning approaches like deep convolutional neural networks are needed to unravel the complexity of eukaryotic gene regulation and enable the use of the genome for personalized medicine.

DIFFERENTIAL NUCLEASE SENSITIVITY PROFILING OF HUMAN CHROMATIN REVEALS CELL-TYPE SPECIFIC NUCLEOSOME POSITIONS, NUCLEOSOME SENSITIVITY, OPEN CHROMATIN, AND TRANSCRIPTION FACTOR BINDING.

Daniel L. Vera¹, Eli Rodgers-Melnick², Sergiusz Wesolowski³, Jonathan H Dennis^{4,1}

¹Florida State University, Center for Genomics and Personalized Medicine, Tallahassee, FL, ²Cornell University, Institute for Genomic Diversity, Ithaca, NJ, ³Florida State University, Department of Mathematics, Tallahassee, FL, ⁴Florida State University, Department of Biological Science, Tallahassee, FL

The eukaryotic genome is packaged into the fundamental unit of chromatin structure, the nucleosome. The positions of nucleosomes on DNA regulate protein-DNA interactions and in turn influence DNA-templated events. High-throughput sequencing of DNA from chromatin digested with micrococcal nuclease (MNase-seq) allows for a global examination of nucleosome landscapes. One surprising finding from studies employing MNase-seq is that the global nucleosome positioning landscape is unexpectedly stable across different cell types and physiological conditions, despite potentially large differences in gene expression levels. On the other hand, the binding of transcription factors that mediate the regulation of transcription shows detectable variation across cell types both by ChIP-seq of transcription factors and by DNase Hypersensitive Site (DHS) mapping. Here we show that by performing MNase-seq using two different degrees of digestion (Differential Nuclease Sensitivity, DNS-seq), we are able to simultaneously identify nucleosome positions, nucleosome-depleted MNase Hypersensitive Sites (MHS), transcription factor binding footprints, and nucleosomes that are hypersensitive ("labile") or hyper-resistant ("resistant") to MNase digestion. Combined with targeted-enrichment of MNase-seq libraries for transcription start sites (TSSs), DNS-seq enables the robust identification of multiple dimensions of chromatin structure simultaneously at high resolution with a limited number of reads. By performing DNS-seq in two distinct human cell types, we identified differences in nucleosome positions and the presence of labile nucleosomes and MHSs between cell types, and that these differences are linked to differential transcription factor binding and gene expression levels. We also identified resistant nucleosomes that show little variation between cell types but are associated with paused RNA polymerase II. By retaining subnucleosomal footprints during library preparation, DNS-seq also allows the detection transcription factor footprints in the absence of ChIP-seq data. DNS-seq, combined with targeted enrichment for TSSs, subverts the need for massive amounts of reads to map nucleosomes in large-genome organisms, and enables the characterizations of multiple layers of chromatin structure information simultaneously.

REORGANIZATION OF CHROMOSOME ARCHITECTURE IN CELLULAR SENESCENCE

Steven W Criscione¹, Marco De Cecco¹, Benjamin Siranosian¹, Yue Zhang¹, Jill A Kreiling¹, John M Sedivy¹, Nicola Neretti^{1,2}

¹Brown University, Department of Molecular Biology, Cell Biology and Biochemistry, Providence, RI, ²Brown University, Center for Computational Molecular Biology, Providence, RI

Replicative cellular senescence is a fundamental biological process characterized by an irreversible arrest of proliferation. Senescent cells accumulate a variety of epigenetic changes but the 3-dimensional (3D) organization of their chromatin is not known. We applied a combination of whole-genome chromosome conformation capture (Hi-C), fluorescence in situ hybridization (FISH), and in silico modeling methods to characterize the 3D architecture of interphase chromosomes in proliferating, quiescent and senescent cells. While the overall organization of the chromatin into active (A) and repressive (B) compartments and topologically associated domains (TADs) is conserved between the three conditions, a subset (~15%) of TADs switches between compartments. On a global level, the Hi-C interaction matrices of senescent cells are characterized by a relative loss of long-range, and gain of short-range interactions within chromosomes. Direct measurements of distances between genetic loci, chromosome volumes, and chromatin accessibility suggest that the Hi-C interaction changes are caused by a significant reduction of the volumes occupied by individual chromosome arms. In contrast, centromeres oppose this overall compaction trend and increase in volume. The structural model arising from our study provides a unique high-resolution view of the complex chromosomal architecture in senescent cells.

MODELING METHYL-SENSITIVE TRANSCRIPTION FACTOR MOTIFS WITH AN EXPANDED EPIGENETIC ALPHABET

Coby Viner^{1,2}, James Johnson³, Nicolas Walker⁴, Marcela Sjoberg⁵, David J Adams⁵, Anne C Ferguson-Smith⁴, Timothy L Bailey³, Michael M Hoffman^{1,2,6}

¹University of Toronto, Department of Computer Science, Toronto, Canada, ²Princess Margaret Cancer Centre, Toronto, Canada, ³University of Queensland, Institute for Molecular Bioscience, Brisbane, Australia, ⁴University of Cambridge, Department of Genetics, Cambridge, United Kingdom, ⁵Wellcome Trust Sanger Institute, Cambridge, United Kingdom, ⁶University of Toronto, Department of Medical Biophysics, Toronto, Canada

Introduction. Many transcription factors (TFs) initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a TF might encounter. Cytosine is often present in the modified forms: 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC). TFs have been shown to distinguish unmodified from modified bases. Modification-sensitive TFs provide a mechanism through which widespread changes in DNA methylation and hydroxymethylation, found in many cancers, can dramatically shift active gene expression programs.

Methods. To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites in DNA with covalent modifications. Our models expand the standard A/C/G/T alphabet, adding m (5mC) and h (5hmC). We adapted the well-established position weight matrix formulation of TF binding affinity to this expanded alphabet.

We engineered several tools to work with expanded-alphabet sequence and position weight matrices. First, we developed a program, Cytomod, to create a modified sequence, using data from bisulfite and oxidative bisulfite sequencing experiments. Cytomod decides between multiple modifications at a single locus, using a configurable evidence model. Second, new versions of MEME (Multiple EM for Motif Elicitation), DREME (Discriminative Regular Expression Motif Elicitation), and MEME-ChIP enable *de novo* discovery of modification-sensitive motifs. A new version of CentriMo enables central motif enrichment analysis to infer direct DNA binding in an expanded-alphabet context. These versions permit users to specify new alphabets, anticipating future alphabet expansions.

Results. We created an expanded-alphabet genome sequence using whole-genome maps of 5mC and 5hmC in naive *ex vivo* mouse T cells from BLUEPRINT. Using this sequence, expanded-alphabet position weight matrices, and ChIP-seq data from Mouse ENCODE and others, we identified cis-regulatory modules active only in the presence or absence of cytosine modifications. We reproduced various known methylation binding preferences, including the preference of ZFP57 and C/EBP β for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.

UNBIASED DISCOVERY OF CIS-REGULATORY ELEMENTS THAT DETERMINE MRNA POST-TRANSCRIPTIONAL REGULATION DURING EARLY DEVELOPMENT

Michal Rabani, Guo-Liang Chew, Alexander F Schier

Harvard University, Molecular and Cellular Biology, Cambridge, MA

Early embryonic development is directed by the post-transcriptional control of maternally provided mRNAs that are present in the oocyte at the time of fertilization, while the embryonic genome is transcriptionally silent. Three hours post fertilization, the maternal-to-zygotic transition (MZT) marks the onset of zygotic transcription and transition into transcription-based regulation. To establish this transfer of control from maternal to zygotically encoded developmental programs, many maternal mRNAs are silenced and cleared during MZT. Although the control over mRNA translation and degradation is pivotal during early development, it is poorly understood how these processes are regulated.

To address this question, we developed UTR-Seq, an unbiased reporter system to simultaneously assess the role of tens of thousands of short UTR sequences in mRNA stability and translation. Using massive parallel oligonucleotide synthesis, we synthesized fragments of 3'UTR sequences of early expressed maternal and zygotic transcripts, and cloned them into an RNA pool of tens of thousands mRNAs that differ only in their 3'UTRs. We introduced this RNA library into developing zebrafish embryos via microinjection, and used high-throughput sequencing to assess dynamic changes in RNA stability and translation. We developed a novel computational scheme that combines exponential decay kinetics with generative sequence models in order to build a catalog of functional regulatory elements embedded in UTRs of mRNAs and direct their translation and clearance.

By comparing RNA clearance profiles across multiple versions of this library that have different polyA tail lengths and ORF structures, we identify classes of sequences that are under distinct modes of regulation. The majority of UTR sequences produce a “default” decay profile that depends on the mRNA’s polyA tail length and ORF structure. But smaller classes of sequences share common regulatory motifs that determine their unique regulation. We find that destabilizing elements and micro-RNA seeds expedite clearance during MZT, cytoplasmic cleavage and polyadenylation signals stabilize messages maternally, and signals for maternally induced deadenylation are associated with early degradation.

Through this principled and unbiased approach, we lay the foundation for further investigation and aim to elucidate the regulatory pathways that mediate mRNA fate during early embryogenesis. Our approach also provides a novel high throughput experimental and computational technology to decode sequence-to-activity relationships in mRNA regulation across many systems.

HIDDEN RNA CODES REVEALED FROM THE PLANT *IN VIVO* RNA STRUCTUROME

Yin Tang^{1,2}, Philip C Bevilacqua^{1,3,4}, Sarah M Assmann^{1,2,4}

¹Pennsylvania State University, University Park, Bioinformatics and Genomics Graduate Program, State College, PA, ²Pennsylvania State University, University Park, Department of Biology, State College, PA, ³Pennsylvania State University, University Park, Department of Chemistry, State College, PA, ⁴Pennsylvania State University, University Park, Center for RNA Molecular Biology, State College, PA

RNA can fold into secondary and tertiary structures, which are important for RNA regulation of gene expression. We recently developed a method to perform genome-wide RNA structure profiling *in vivo* employing high-throughput sequencing techniques, and applied this methodology to the model plant species *Arabidopsis* (Ding *et al.*, 2014; Ding *et al.*, 2015). This method makes it possible to probe thousands of RNA structures at one time in living cells. Hidden RNA codes have been revealed by bioinformatic and biostatistical analyses of our *in vivo* RNA structuromes including RNA secondary structures related to alternative polyadenylation and splicing (Ding *et al.*, 2014). Recently, our further analysis of this dataset revealed a correlation between mRNA structure and the structure of the encoded protein (Tang *et al.*). We have found that regions of individual mRNAs that code for protein domains generally have significantly higher structural reactivity than regions that encode protein domain junctions. This relationship is especially prominent for proteins annotated for catalytic activity but is reversed in proteins annotated for binding and transcription regulatory activity. We found a similar pattern for disordered regions of proteins as compared to their ordered regions, wherein mRNA segments that code for ordered regions have significantly higher structural reactivity than those that encode disordered regions. These results indicate the vital roles of RNA structures in regulation of gene expression and protein folding.

We also developed a new computational platform, StructureFold, to facilitate the bioinformatic analysis of high throughput RNA structure profiling data. As a component of the Galaxy platform (<https://usegalaxy.org>), StructureFold (Tang *et al.*, 2015) integrates reads mapping, RT stop count calculation, structural reactivity derivation, and RNA structure prediction in a user-friendly web-based interface or via local installation. It is an efficient and easy tool for the analysis of high-throughput RNA structural probing data, which otherwise requires great bioinformatic efforts.

References

- Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C. and Assmann, S.M. *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 2014;505(7485):696-700.
- Ding, Y., Kwok, C.K., Tang, Y., Bevilacqua, P.C. and Assmann, S.M. Genome-wide profiling of *in vivo* RNA structure at single-nucleotide resolution using structure-seq. *Nat Protoc* 2015;10(7):1050-1066.
- Tang, Y., Assmann, S.M. and Bevilacqua, P.C. Protein structure is related to RNA structural reactivity *in vivo*. submitted.
- Tang, Y., Bouvier, E., Kwok, C.K., Ding, Y., Nekrutenko, A., Bevilacqua, P.C. and Assmann, S.M. StructureFold: genome-wide RNA secondary structure mapping and reconstruction *in vivo*. *Bioinformatics* 2015.

USING A BUMP HUNTING APPROACH FOR GENOME-WIDE IDENTIFICATION OF NOVEL IMPRINTED GENES.

Yulia Rubanova^{1,2}, Andrei Turinsky², Sanaa Choufani³, Rosanna Weksberg^{3,4}, Michael Brudno^{1,2,3}

¹Department of Computer Science, University of Toronto, Toronto, Canada, ²Centre for Computational Medicine, The Hospital for Sick Children, Toronto, Canada, ³Genetics and Genome Biology Program, Hospital for Sick Children Research Institute, University of Toronto, Toronto, Canada, ⁴Division of Clinical and Metabolic Genetics, Department of Paediatrics, Hospital for Sick Children, University of Toronto, Toronto, Canada

Imprinting is an epigenetic mechanism that plays a crucial role in fetal development and growth. Imprinted regions are expressed in a parent-of-origin manner due to differences in methylation between paternal and maternal alleles. In our study we search for novel imprinted regions leveraging such methylation differences.

We use the Infinium HumanMethylation450 BeadChip data to perform a genome-wide search for new candidate imprinted regions. Compared to previous studies based on the Infinium HumanMethylation27 BeadChip data, our work covers almost 20 times more CpG sites (485,577 CpGs genome-wide compared to 27,578 CpGs in ~13,000 genes), which allows a more thorough analysis of methylation data.

We compare methylation values in samples from normal tissues, containing alleles from both parents, to methylation in samples of uniparental origin. The latter includes samples with uniparental disomy (UPD) (alleles from only one parent in single chromosomes), paternally-derived androgenetic complete hydatidiform moles (AnCHMs) and maternally-derived mature cystic ovarian teratoma (MCT).

We use a statistical bump hunting approach. Bump hunting detects microarray probes where differences between cohorts exceed a certain threshold; merges the corresponding probe sites into extended genomic regions based on several criteria; and statistically validates the results via permutation analysis. In our study, we utilize the bump hunting approach on the genome-wide DNA methylation data to discover new imprinted genes and validate imprinted gene candidates from previous studies.

Using a the epigenome-wide bump hunting approach we were able to verify 21 known imprinted region and candidates and reveal 26 new candidate imprinted regions. Across the novel imprinted regions detected by our method, we observe large methylation differences of 25% on average between paternal and maternal alleles compared to an average difference of only 9% in non-imprinted regions. In particular, the found candidates overlap the promoter and body of several genes with diverse function, such as KRT12, implicated in corneal dystrophy; LDHAL6A, which is involved in metabolic pathways; and TBPL2, required for the differentiation of myoblasts into myocytes. Our results demonstrate how statistical techniques for epigenome-wide analysis may be able to detect many new imprinted genome regions as targets for future biomedical validation and functional analysis.

PCxN: THE PATHWAY CO-ACTIVITY MAP: A NEW APPROACH FOR THE UNIFICATION OF FUNCTIONAL BIOLOGY

Yered Pita-Juarez¹, Gabriel Altschuler², Wenbin Wei², Winston Hide^{1,2}

¹Harvard School of Public Health, Department of Biostatistics, Boston, MA,

²University of Sheffield, Sheffield Institute for Translational Neuroscience, Sheffield, United Kingdom

Cellular processes and pathways are often represented as independent sets of genes. As such, analysis of genomic data for pathway enrichment is conducted using tools such as GSEA that analyze gene sets as independent units. In contrast, recent functional genomic experiments demonstrate ubiquitous interaction amongst regulatory and signaling molecules, enzymes, structural proteins and DNA regulatory events, directly challenging the traditional view of pathways as relatively independent functional units. Approaches to expose potential interactions have relied primarily on shared genes between processes, networks and pathways, combining information from databases and interaction networks, or using direct physical interaction between genes and gene products to determine likely interaction. These approaches are popular but provide inadequate inference of global interactions between functional activities.

With the aim of discovery of global interaction between pathways we have created a pathway co-activity map, PCxN, based on the co-expression between canonical pathways and functional modules. Using a background of 58,000 microarrays from the Gene Expression Omnibus as a common reference, we have estimated the co-activity relationships between canonical pathways from KEGG, Reactome, Wikipathways and Netpath. We have also mapped activity between terms in the GO ontology and in 144 static modules - clusters of genes derived from a functional interaction network. We characterise pairwise pathway relationships by taking the partial correlation between the summary statistics for pathway activity, accounting for contribution from shared genes.

PCxN provides a high level map of related cellular functions by systematically quantifying the relationship between pathways. It is a weighted undirected network in which the nodes represent pathways and the edges represent correlation coefficients. It provides the opportunity to understand the relationship between biological functions not just by their shared member genes, but by their interactions.

This pathway co-activity map makes it possible, for the first time, to explore known and novel relationships between pathways of interest, extending the study of a pathway to its co-activated neighbours across functional space, revealing associations between specific pathways or processes and enabling analysis of the influence of previously unknown pathway relationships. Global clustering reveals four classes of co-activated pathways within functional biology: RNA processing, protein trafficking/folding, immune regulation and growth and development. We apply PCxN to Alzheimer's variant signatures to reveal key known and novel interacting disease pathways. We provide an online resource of PCxN., and an interface that extends GSEA results to visualize relationships between significant pathways and reveal novel pathways seeded by enrichment results.

GENOME-WIDE COPY NUMBER VARIATION ANALYSIS FOR *PLASMODIUM VIVAX* GLOBAL ISOLATES.

Zunping Luo¹, Daniel N Hupalo¹, Daniel E Neafsey², Jane M Carlton¹

¹Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, ²Broad Institute of MIT and Harvard, Cambridge, MA

Of the four *Plasmodium* species that routinely cause human malaria, *Plasmodium vivax* is the most widespread species outside Africa, causing ~15.8 million cases in 2013. We collected and sequenced 171 blood samples from patients infected with *P. vivax* from eight regions of the world, utilizing a hybrid selection approach to reduce human DNA contamination. Copy number variation (CNV) has been associated with differing susceptibility to disease, and copy number variants of important genes have been identified in a few *P. vivax* patient samples; however, a genome-wide analysis of CNV in a global population of *P. vivax* has never been done. We compared two structural variation discovery programs, Pindel (Ye *et al.*, Bioinformatics 2009) and DELLY (Rausch *et al.*, Bioinformatics 2012) to investigate genome-wide CNV in 102 *P. vivax* isolates. These isolates consisted of single genotype infections, with high quality reads, and are from six geographical locations: China, Thailand, Papua New Guinea, Peru, Colombia and Mexico. First, we compared the number of copy number variants discovered by Pindel and DELLY, using custom Python scripts to parse the Pindel and DELLY outputs; and then we mined the results for specific genes of interest. Copy number variants were subsequently validated by PCR. We present here our analysis and focus on several genes of interest. (1) Duffy binding protein (DBP) is involved in red blood cell invasion, and duplication of the gene has been identified in patients from Madagascar (Menard *et al.*, PLoS NTD 2013). No DBP CNV was observed in any of our global isolates, a result validated by PCR. (2) Drug resistance genes including multi-drug resistance gene (MDR1) and dihydrofolate reductase-thymidylate synthase (DHFR-TS) are involved in antimalarial drug resistance. We found putative CNV of these genes, confirming previous reports of MDR1 gene duplication in patients failing drug treatment in Thailand and Laos (Imwong, *et al.*, Antimicrobial Agents Chemotherapy 2008); CNV of DHFR has never before been reported in *P. vivax*. By using two different algorithms to characterize genome-wide CNV in our *P. vivax* global isolates, in addition to wet-lab validation, we have generated a map of gene duplications in this important malaria parasite species.

RAREVARIANTVIS: A NEW TOOL FOR IDENTIFICATION OF CAUSATIVE VARIANTS IN RARE MONOGENIC DISORDERS FROM WHOLE GENOME SEQUENCING DATA

Tomasz Stokowy¹, Mateusz Garbulowski², Torunn Fiskerstrand¹, Rita Holdhus¹, Kornel Labun³, Pawel Sztromwasser¹, Christian Gilissen⁴, Alexander Hoischen⁴, Gunnar Houge¹, Kjell Petersen⁵, Inge Jonassen⁵, Vidar Steen¹

¹University of Bergen, Department of Clinical Science, Bergen, Norway, ²Silesian University of Technology, Department of Informatics, Gliwice, Poland, ³Silesian University of Technology, Department of Automatic Control, Gliwice, Poland, ⁴Radboud University Medical Centre, Department of Human Genetics, Nijmegen, Netherlands, ⁵University of Bergen, Department of Informatics, Bergen, Norway

The search for causative genetic variants in rare diseases with unknown etiology of presumed monogenic inheritance has been boosted by the implementation of whole exome (WES) and whole genome (WGS) sequencing. WGS seems to be superior to WES thanks to equally distributed coverage and possibilities of non-coding variant analysis, but the analysis and visualization of the vast amounts of data is demanding. To meet this challenge, we have developed a new tool – RareVariantVis – for analysis of genome sequence data (including non-coding regions) for both germ line and somatic variants. It visualizes variants along their respective chromosomes, providing information about exact chromosomal position, zygosity (i.e. percentage of variant reads) and frequency, with click-able information regarding dbSNP IDs (if relevant), gene association (if relevant) and whether the variant is inherited from the mother, father or both. Rare variants with no dbSNP ID or frequency below a certain user-defined threshold as well as de novo variants can be flagged and visualized in different colors. We have tested extensively the RareVariantVis tool in two WGS data sets, Genome in a Bottle Ashkenazim Trio data (Complete Genomics) and about 30 in-house samples (Illumina X Ten) obtained from families with rare inherited disorders. This work has clearly demonstrated the usefulness of RareVariantVis in the screening and identification of possible causative variants for monogenic disorders.

The RareVariantVis tool accepts vcf files and annotated variant tables. It can be efficiently run on a desktop computer - whole genome is loaded, filtered and visualized in about 10 minutes. The tool with its documentation is available for download under the following link:
<http://bioconductor.jp/packages/3.2/bioc/html/RareVariantVis.html>

DISCOVERY OF GENETIC HETEROGENEITY IN A CONTEXT OF PHYSIOLOGICAL HOMOGENEITY BY BIOLOGICAL DISTANCE CLUSTERING

Yuval Itan¹, Shen-Ying Zhang^{1,2}, Laurent Abel^{2,1}, Jean-Laurent Casanova^{1,2,3}

¹The Rockefeller University, Human Genetics of Infectious Diseases, New York, NY, ²INSERM, Human Genetics of Infectious Diseases, Paris, France, ³Howard Hughes Medical Institute, Howard Hughes Medical Institute, New York, NY

To determine the disease-causing allele(s) underlying primary human inborn errors, high-throughput genomic methods are applied and provide thousands of gene variants per patient. We recently reported a novel approach, the “human gene connectome” (HGC) – the set of all in silico-predicted biologically plausible routes and distances between all pairs of human genes, effective for prioritizing gene variants by biological distance from known disease-causing genes. However, there is currently no available method for automating the selection of candidate disease-causing mutant alleles in the absence of a known morbid gene in at least one patient with the disease of interest, posing a major bottleneck in the field in high-throughput clinical genomics. We hypothesized that within a cohort of patients with the same Mendelian disease, the cluster that contains the key disease-causing gene for each patient is the HGC-predicted biologically smallest cluster. We then developed and applied a Mendelian clustering algorithm, which estimates the biologically smallest HGC-predicted cluster that contains one allele per patient. By that we (i) approximated a solution for an NP-complete algorithmic problem (i.e. not possible to solve on a large scale by a computer), and (ii) estimated and statistically validated a set of disease-causing alleles in whole exome sequencing cohorts of Mendelian disease patients. The unbiased approach described above should facilitate the discovery of morbid alleles in patients with primary inborn errors that lack a genetic etiology.

GINKGO: INTERACTIVE ANALYSIS AND QUALITY ASSESSMENT OF SINGLE-CELL CNV DATA

Robert Aboukhalil, Tyler Garvin, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, Michael C Schatz

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

In recent years, single-cell sequencing has become an important tool for unraveling the genomic heterogeneity of biological samples, and has enabled the study of tumor evolution, circulating tumor cells, and neuronal mosaicism. One important application of single-cell sequencing is to identify large-scale (>10kb) copy-number variations, which are known to play important roles in several diseases.

Here we introduce Ginkgo, a web-based platform for the interactive analysis and quality assessment of single-cell copy-number alterations. Ginkgo automates and standardizes the computation required to go from mapped reads to copy-number profiles of individual cells, to phylogenetic trees of entire cell populations. Ginkgo also enables users to navigate within a cell's copy number profile, zoom into regions of interest, annotate profiles with genes of interest, and automatically export amplification/deletion tracks to the UCSC browser for further inspection. Ginkgo is available online at <http://qb.cshl.edu/ginkgo>.

To validate Ginkgo, we reproduce the major findings of six human datasets across five recent single-cell studies. These datasets address vastly different scientific questions, were collected from a variety of tissue types, and make use of different experimental and computational approaches at different institutions.

Next, we use Ginkgo's quality assessment tools to examine the data characteristics of three commonly used single-cell amplification techniques (MDA, MALBAC, and DOP-PCR) through comparative analysis of 9 different single-cell datasets. We find that both MALBAC and DOP-PCR outperform MDA in terms of data quality. As previously reported, MDA displays poor coverage uniformity and low signal-to-noise ratios. Coupled with high GC biases, MDA is unreliable for accurately determining CNVs compared to the other two techniques. Furthermore, while both DOP-PCR and MALBAC data can be used to generate CNV profiles and identify large variants, we find that DOP-PCR data exhibits lower coverage dispersion and smaller GC biases when compared to MALBAC data. Given the same level of coverage, our results indicate that data prepared using DOP-PCR can reliably call CNVs at higher resolution with better signal-to-noise ratios.

EVALUATING THE APPLICATION OF SEQUENCING DATA TO DIFFERENTIAL COEXPRESSION USING THE DISCORDANT METHOD

Charlotte J Siska¹, Katerina Kechris²

¹University of Colorado Anschutz Medical Campus, Computational Bioscience Program, Aurora, CO, ²University of Colorado Anschutz Medical Campus, Department of Biostatistics and Informatics, Aurora, CO

Differential coexpression of two molecular features (i.e., gene-gene, gene-miRNA, protein-metabolite) occurs when associations between features are different between biological groups, such as disease vs control. These pairs can be important in identifying biochemical pathways or interactions that are unique. We have developed a method called Discordant, which determines differentially coexpressed molecular features by using mixture models and the EM algorithm. We have shown through previous work that it has greater power than other leading methods and also is able to identify novel and known interactions. When constructing the Discordant method, the only type of data we considered was continuous, such as transcriptomic data from microarrays. Microarrays are still widely used in research, but RNA-Seq is becoming the more common platform. In this work, we extend and evaluate whether the Discordant method can be applied to discrete data such as RNA-Seq, by using non-parametric rank correlation alternatives. We investigated the application of discrete vs. continuous data by using a breast cancer dataset from the Cancer Genome Atlas (TCGA) that had RNA-Seq and microarray data from the same samples. We observed miRNA-mRNA pairs with two Discordant results – miRNA-mRNA pairs based on miRNASeq and RNA-Seq, and miRNA-mRNA pairs based on miRNASeq and microarray. We validated and compared results by determining the number of molecular feature pairs that had an experimentally validated breast cancer miRNA. We found that breast-related miRNAs were found at about the same rate with continuous or discrete transcriptomic data, that known miRNA were highly ranked with both platforms, and that the rank-based correlations suitably met the assumptions of the Discordant model. From these findings, we show evidence that our method is applicable to discrete data with similar performance.

BASECALLING FROM RAW OXFORD NANOPORE DATA

Gerton Lunter

University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford,
United Kingdom

Oxford Nanopore's MinION platform's many interesting features including its small size (it is operated from an ordinary laptop), its ability to sequence very long molecules, and its ability to physically select or reject molecules to sequence in real time. However, as expected from a single-molecule platform, the reads are fairly noisy, and this affects many applications.

The MinION device measures the current of H^+ ions through a pore that is partially blocked by a single-stranded DNA molecule. The specific configuration of DNA bases within the pore modulates this H^+ current. A special "motor" enzyme periodically and stochastically moves the DNA molecule through the pore, one base at a time, and the resulting stepwise changes in the H^+ current are decoded to reveal the original DNA sequence. The raw signal consists of current measurements at 3-5 kHz, for each of the 500 pores. To deal with this large data volume, this signal is processed in real time to identify the intervals of constant current, referred to as "events". Event sequences are then uploaded to the cloud and converted to DNA sequences Nanopore's cloud basecaller, Metrichor.

Inevitably, the event caller makes mistakes, and in the absence of raw data these are difficult to correct downstream, leading to indel errors. Here I present a base caller that works directly off the raw signal. In addition to improved base qualities, the model in principle allows identification of deviations from the expected signals, to identify base modifications. I will present some initial results of this method comparing methylated and non-methylated *E. Coli* sequences.

DIPLOID GENOME ASSEMBLY AND COMPREHENSIVE HAPLOTYPE SEQUENCE RECONSTRUCTION

Jason Chin¹, Paul Peluso¹, David Rank¹, Maria Nettetstad², Michael Schatz², Alicia Clum³, Alex Copeland³, Barry Kerrie³

¹Pacific Biosciences, Menlo Park, CA, ²Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ³Joint Genome Institute, Walnut Creek, CA

Genome sequence represents a comprehensive blueprint that codes for the form and function of all living organisms; describing the differences of individuals, species, and all levels of taxonomic hierarchy. Outside of the simplest cases (haploid, bacteria, or inbreds), genomic information is not carried in a single reference per individual, but rather has higher ploidy ($n \geq 2$) for almost all organisms. The existence of two or more highly related sequences within an individual makes it extremely difficult to build high-quality, highly contiguous genome assemblies from short DNA fragments. Long DNA sequence reads already provide invaluable information for reconstructing haploid genomes from cell lines or inbred strains to high contiguity and high consensus accuracy. With the longer read lengths and largely un-biased error profiles, it is possible to reconstruct comprehensive haplotype specific contigs (“haplotigs”). Genomic researchers can identify haplotype-specific genomic features beyond simple SNPs from these haplotigs, including large indels, inversions, and translocations. Based on the earlier work on a polyploidy aware assembler, FALCON (<https://github.com/PacificBiosciences/FALCON>), we developed new algorithms and software (“FALCON-unzip”) for comprehensive haplotype reconstructions from SMRT[®] sequencing data. Moreover, an “augmented alignment process” that matches phased reads and haplotigs are utilized for achieving high haplotype consensus accuracy.

We apply the algorithms and the prototype software to a synthetic cross of the sequencing data from two inbred Arabidopsis strains and a highly repetitive diploid fungal genome (*Clavicornia pyxidata*). For the fungal genome, an independent sequencing data set from an orthogonal platform is available from the Sequence Read Archive. These datasets allow us to evaluate the performance of FALCON-unzip. We achieve an N50 of 1.53 Mb (of the 1n assembly contigs) of the ~42 Mb 1n genome and an N50 of the haplotigs of 872 kb. The assembly contig sizes are significantly higher than the previous attempt with short-read data that resulted in a contig N50 of 87kb. The consensus accuracy and base concordance is evaluated with the reads from the Illumina[®] platform. Given that the resulting contigs and haplotigs are more contiguous and the heterozygous variants are captured in phase with less crossing, we can determine the differences between haplotypes at gene and larger size scales better. The more complete view of both haplotypes can help in generating better annotation and provide useful biological insights. Finally, we apply this method to a couple of recently collected human data sets to demonstrate the potential for resolving complicated but important, biomedically relevant regions in human genomes.

HISAT2: GRAPH-BASED ALIGNMENT OF NEXT-GENERATION SEQUENCING READS TO A POPULATION OF HUMAN GENOMES

Daehwan Kim¹, Steven L Salzberg^{1,2}

¹Johns Hopkins University, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, ²Johns Hopkins University, Departments of Biomedical Engineering, Computer Science, and Biostatistics, Baltimore, MD

Since the introduction of next-generation sequencing (NGS) technologies, multiple large-scale human sequencing projects have been launched, including the 1000 Genomes Project, GTEx, and GEUVADIS. These projects have already yielded a large and growing amount of information about human genetic variation, including >65 million SNPs (in dbSNP) and >10 million structural variants (in dbVar). Although these variants represent a valuable resource for genetic analysis, computational tools do not adequately incorporate the variants into genetic analysis. Most analyses begin by aligning reads against the human reference genome, which was assembled using only a few individuals of northern European origin and thus does not reflect genetic variations across individuals and populations. Alignment based on this single reference leads to two forms of error, or bias: (1) reads from regions of a source genome that are substantially different from the reference genome will simply fail to align, and (2) reads from regions with small differences can be aligned to the incorrect location. These discrepancies in mapping can introduce significant biases in downstream analyses, and may cause important genetic variants to be missed entirely. Although some programs have attempted to provide variant-aware alignment, solutions to date have been either extremely slow, highly memory intensive, or both.

To address these challenges, we have developed a novel indexing scheme that captures a wide representation of the human population with low memory requirements. We have combined this index with a mapping algorithm that enables fast, accurate, and sensitive alignment of NGS reads. Building an index incorporating the complete genomes of individuals would introduce an enormous amount of redundancy, requiring a huge amount of memory. Because *Homo sapiens* has relatively low diversity as a species (0.1% difference between individuals) and genomic differences are widely shared among human populations, we instead created an index that focuses on incorporating the differences into the reference genome. We have extended the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index to incorporate genomic differences among individuals into the reference genome, while keeping memory requirements low enough to fit the entire index onto a desktop computer. Using this novel Hierarchical Graph FM index (HGFM), we have built a new alignment system, HISAT2, with an index that incorporates ~12.3M common SNPs from the dbSNP database. This new scheme shows promise, with an index size of just 6.2 GB and only 30~80% additional CPU time compared to HISAT, among the fastest alignment programs, and greater alignment accuracy for reads containing SNPs. We will soon extend the index to incorporate structural variants. HISAT2 is open-source software available at www.ccb.jhu.edu/software/hisat2.

COMPREHENSIVE GENOME AND TRANSCRIPTOME STRUCTURAL ANALYSIS OF A BREAST CANCER CELL LINE USING PACBIO LONG READ SEQUENCING

Maria Nattestad¹, Karen Ng², Sara Goodwin¹, Timour Baslan¹, Fritz Sedlazeck¹, James Gurtowski¹, Elizabeth Hutton¹, Marley Alford¹, Elizabeth Tseng³, Jason Chin³, Timothy Beck², Yogi Sundaravadanam², Melissa Kramer¹, Eric Antoniou¹, John McPherson², James Hicks¹, Michael Schatz¹, W R McCombie¹

¹Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY, ²Ontario Institute for Cancer Research, Cancer Genomics, Toronto, Canada, ³Pacific Biosciences, Bioinformatics, Menlo Park, CA

Genomic instability is one of the hallmarks of cancer, leading to widespread copy number variations, chromosomal fusions, and other structural variations. The breast cancer cell line SK-BR-3 is an important model for HER2+ breast cancers, which are among the most aggressive forms of the disease and affect one in five cases. Through short read sequencing, copy number arrays, and other technologies, the genome of SK-BR-3 is known to be highly rearranged with many copy number variations, including an approximately twenty-fold amplification of the HER2 oncogene. However, these technologies cannot precisely characterize the nature and context of the identified genomic events and other important mutations may be missed altogether because of repeats, multi-mapping reads, and the failure to reliably anchor alignments to both sides of a variation.

To address these challenges, we have sequenced SK-BR-3 using PacBio long read technology. Using the new P6-C4 chemistry, we generated more than 70X coverage of the genome with average read lengths of 9-13kb (max: 71kb). Using Lumpy for split-read alignment analysis, as well as our novel assembly-based algorithms for finding complex variants, we have developed a detailed map of structural variations in this cell line. Taking advantage of the newly identified breakpoints and combining these with copy number assignments, we have developed an algorithm to reconstruct the mutational history of this cancer genome. From this we have discovered a complex series of nested duplications and translocations between chr17 and chr8, two of the most frequent translocation partners in primary breast cancers, resulting in amplification of HER2. We have also carried out full-length transcriptome sequencing using PacBio's Iso-Seq technology, which has revealed a number of previously unrecognized gene fusions and isoforms. Combining long-read genome and transcriptome sequencing technologies enables an in-depth analysis of how changes in the genome affect the transcriptome, including how gene fusions are created across multiple chromosomes. This analysis has established the most complete cancer reference genome available to date, and is already opening the door to applying long-read sequencing to patient samples with complex genome structures.

WHY IS “QUERYING” THE GENOME SO DIFFICULT?

Aaron Quinlan

University of Utah, Departments of Human Genetics and Biomedical Informatics, Salt Lake City, UT

Segregating the minority of genetic variants that underlie a phenotype from the millions of irrelevant variants is a fundamental informatics challenge that researchers routinely face. This difficulty is becoming ever more acute as disease studies involve thousands and eventually millions of human genome. On the one hand, this is a problem of scale - millions of genomes yield hundreds of millions of variants and trillions of genotypes and existing algorithms are largely ill-suited to datasets of this magnitude. On the other, it is a problem of context: in many cases it is either very difficult or impossible to collect sufficient genome annotations to provide the necessary context to stratify the genetic variation observed. I will present substantial progress in the development of multiple technologies that we are actively developing to address these challenges. Our ultimate goal is the development of a highly scalable and interactive system for querying genomes in diverse disease and organismal contexts.

GENOME SCAFFOLDING AND STRUCTURAL VARIATION DETECTION FROM MINION NANOPORE SEQUENCING DATA

Zemin Ning, Louise Aigrain, James Bonfield, Robert Davies, Michael Quail, David Jackson, Thomas Keane, Richard Durbin

The Wellcome Trust Sanger Institute, Wellcome Genome Campus,
Hinxton, Cambridge, United Kingdom

With its long length profile and speedy data production on a portable device, the MinION sequencers provide promising prospects for diagnostics, genome assembly, phasing and many more applications. We explored the newly generated data with two applications – MinION assisted genome scaffolding and detection of structural variations on a eukaryotic genome – *S. cerevisiae*.

To make the most use of the read length, we started with the making of fake mate pairs by shredding the MinION reads into 1 or 2kb segments. Here the insert size is the cutting distance between the separated pairs. Mate pair sequences with known insert sizes were aligned against the target assembly and alignments were processed to build up a connection graph in which contigs were treated as nodes and inter-contig links were stored as edges. Scaffolding was performed based on the connection graph using insert sizes to estimate the gap length. We have developed a pipeline SMIS - Single Molecular Integrated Scaffolding [1] for this application. First, an initial assembly using Illumina short reads [2] was generated and MinION assisted scaffolding was carried out to increase scaffold length followed by gap closure on contigs. Using the Illumina MiSeq reads [[2], the final assembly has N50 stats of scaffold 858kb and contig 330kb, respectively with 11 of 17 chromosomes assembled into essentially one piece.

We also report our current progress in identification of structural variations from the MinION reads. Comparing two yeast strains S288C and W303, we present the results of sensitivity and specificity on deletions, insertions and inversions using the available data. Finally, we discuss the effects of chimeric reads, a common and interesting feature from single molecular sequencing platforms.

[1] SMIS: Single Molecular Integrated Scaffolding

<http://sourceforge.net/projects/phusion2/files/smis/>

[2] <http://www.ncbi.nlm.nih.gov/biosample/839783/>

[3] <http://labshare.cshl.edu/shares/schatzlab/www-data/nanocorr/>

HOW TO COMPARE AND CLUSTER EVERY KNOWN GENOME IN ABOUT AN HOUR

Adam M Phillippy¹, Brian D Ondov², Todd J Treangen², Sergey Koren¹

¹National Institutes of Health, National Human Genome Research Institute, Bethesda, MD, ²National Biodefense Analysis and Countermeasures Center, Genomics, Frederick, MD

The rapid growth of genomic data has begun to outpace traditional methods for sequence clustering and search. Given a massive collection of genomes, it is infeasible to perform pairwise alignment for basic tasks like quality control, clustering, and species identification. To address this, we demonstrate that the MinHash technique, first applied to clustering the World Wide Web, can be applied to biological sequences with similar effect, and extend this idea to include biologically relevant distance and significance measures. Our new tool, Mash, uses locality-sensitive hashing to perform dimensionality reduction and rapidly approximate the distance between two genomes or metagenomes. Using Mash, we explored several use cases, including a thousand-fold size reduction and clustering of all 55,000 NCBI RefSeq genomes in about an hour. The resulting 107 MB Mash database includes all RefSeq microbial and eukaryotic genomes, effectively delineates species boundaries, reconstructs approximate phylogenies, and can be searched in seconds using assembled genomes or raw sequencing runs from any technology, including Illumina, Pacific Biosciences, and Oxford Nanopore. For metagenomics, Mash scales to thousands of samples and can replicate Human Microbiome Project and Global Ocean Survey results in a fraction of the time previously required. Thus, this technique can be applied to any problem that requires a fast distance approximation, e.g. to triage and cluster sequence data, assign species labels to unknown genomes, quickly identify mis-tracked samples, and search massive genomic databases. To facilitate integration with other software, Mash is implemented as a lightweight C++11 toolkit and freely released under a BSD license at <https://github.com/marbl/mash>.

IMPROVED METHODS FOR NGS-BASED CONOTOXIN DISCOVERY

Qing Li¹, Pradip K Bandyopadhyay², Helena Safavi-Hemami², Samuel D Robinson², Aiping Lu³, Jason S Biggs⁴, Baldomero M Olivera², Mark Yandell^{1,5}

¹University of Utah, Eccles institute of Human Genetics, Salt Lake City, UT, ²University of Utah, Dept. of Biology, Salt Lake City, UT, ³Tongji University, School of life sciences and technology, Shanghai, China, ⁴University of Guam Marine Laboratory, UOG Station, Mangilao, GA, ⁵University of Utah, The Utah Science Technology and Research Initiative Center for Genetic Discovery, Salt Lake City, UT

Cone snails (genus *Conus*) have attracted scientific interest for the great neuropharmacological potential of their venom, which consists of a complex mixture of peptides known as conotoxins. For discovery purposes, we have carried out a survey of the venom-ducts of 16 *Conus* species using next-generation high-throughput RNA-seq. In silico analyses of these data are complicated because paralogous conotoxin precursors display both highly conserved, as well as hyper-varied regions. As a result, NGS-based discovery involves an inherent trade off between the fidelity of transcript assembly and sensitivity towards novel discovery. On the one hand, overly lenient assembly parameters create a few, long, but misassembled chimeric transcripts, which lessen the true discovery potential of NGS. On the other hand, overly stringent assembly parameters can mistake sequencing artifacts as novel discoveries. Another issue is contamination from co-processed samples. Distinguishing these cases from low abundance transcripts is challenging as well. With these caveats in mind, we have developed a set of best practice tools, procedures, and protocols designed to maximize sensitivity and specificity for NGS-based conotoxin discovery. Our approach employs a novel kmerization tool called Taxonomer which can very rapidly cluster and taxonomically classify reads prior to assembly. Taxonomer is able to pre-classify millions of reads in minutes, enabling targeted and precise micro-assemblies of complete, high-fidelity transcripts and thus maximizes the discovery potential of RNA-seq-based conotoxin discovery. In fact, this pipeline discovers, on average ~30% more full-length toxins than current best practices. We have also begun to explore the feasibility of this approach for genomic sequencing data. Despite the present hurdles in assembling *Conus* genomes, we have been able to recover and assemble a number of conotoxin-encoding genomic regions. Finally we demonstrate the power of our approach NGS-based conotoxin discovery for investigations of how conotoxin repertory and expression correlate with taxonomy and life history traits for our 16 *Conus* species.

DOSAGE SENSITIVE GENES IN EVOLUTION AND DISEASE

Aoife McLysaght

Trinity College Dublin, Genetics, Dublin, Ireland

Copy number variants (CNVs) are very common, and are the largest contributor to human genetic variation when considered per base-pair. Most of this variation is neutral, but a substantial fraction is pathogenic, contributing to neuropsychiatric conditions, heart disease, cancers, and others. The pathogenicity is generally considered to be due to dosage sensitivity of one or more genes within the CNV region, such that dosage imbalances perturb function.

We previously discovered a link between the patterns of evolution by gene duplication and whether a gene is present in pathogenic CNVs. We conclude that this is probably due to ancient and persistent dosage-sensitivity of the genes.

In this talk I will present our recent work on refining the analysis of vertebrate genome evolution in order to better detect these evolutionary patterns. I will also place gene evolutionary patterns in the context of observed benign and pathogenic CNVs.

COMPUTATIONAL ANALYSIS OF DISEASE-ASSOCIATED FUNCTIONAL SHIFTS IN THE PERIODONTAL MICROBIOME

Shareef M Dabdoub*, Sukirth M Ganesan*, Purnima S Kumar

The Ohio State University, Periodontology, Columbus, OH

It is well known that periodontitis, a polymicrobial disease that destroys tooth-supporting structures, is taxonomically heterogenous. However, little is known about the functional idiosyncrasies of this ecosystem. We examined 73 microbial assemblages from 25 individuals with generalized chronic periodontitis (25 deep and 23 shallow-site samples) and 25 periodontally healthy individuals using comparative metagenomics. Core metabolic networks were computed, and abundances of functional genes examined within this framework.

Whole-genome shotgun sequencing was performed using the Illumina MiSeq platform, generating 22.65 million high-quality sequences. Filtered sequences were processed using the MG-RAST pipeline (augmented by computational resources available at the Ohio Supercomputer Center) for subsystem classification and characterization. Significant differential abundance was determined using DESeq2, and additional analysis and visualization tools were developed with Python and are available on GitHub (PyMGRASST).

Over two-thirds of the core metagenome, especially genes encoding for energy metabolism, stress response, iron transport, metal and antibiotic resistance, flagella and lipopolysaccharide, diverged significantly between health and disease (both deep and shallow sites). Virulence-enhancing functional synergisms were observed in disease between the virome, archeome and bacteriome. Bacteria in the healthy ecosystem functioned as 'generalists', with all species contributing equally to the functional requirements of the system, while the disease-associated microbiome was dominated by 'specialists', with each species contributing a defined set of functions. Even though the communities were phylogenetically heterogeneous at both subject and site levels, they were functionally congruent. Several genes, but not species, demonstrated robust discriminating power between health and disease.

The periodontal microbiome shifts from an energy efficient ecosystem in health to a highly entropic system in disease. Global functional dysbiosis is seen in disease; with every site capable of initiating a pro-inflammatory host response. Importantly, shallow sites in individuals with disease appear to be at greater risk for harm than previously believed. The identification of taxonomically idiosyncratic but functionally similar communities supports a gene-centric rather than a species-centric theory of disease causation.

All authors contributed equally.

A GENETIC ANALYSIS OF A COMPLEX TRAIT IN A “GENETICALLY INTRACTABLE” GUT MICROBE

Sena Bae¹, Olaf Muller², Sandi Wong³, John F Rawls^{2,3}, Raphael H Valdivia³

¹Duke University, Biomedical Engineering, Durham, NC, ²Duke University, Center for the Genomics of Microbial Systems, Durham, NC, ³Duke University, Molecular Genetics and Microbiology, Durham, NC

Microbes mediate immune and nutrient homeostasis in the vertebrate gastrointestinal tract. The molecular basis for these host-microbe interaction is poorly understood as many gut microbes are not amenable to molecular genetic manipulation. We combined phenotypic selection after chemical mutagenesis with population-based whole genome sequencing to identify genes that are required for motility in the firmicute *Exiguobacterium*, a component of the vertebrate gut microbiota that contributes to lipid uptake. We derived strong associations between the loss of motility and mutations in predicted *Exiguobacterium* motility genes and genes of unknown function. We confirmed the genetic linkage between the predicted causative mutations and loss of motility by identifying suppressor mutations that restored motility. These results indicate that a genetic dissection of complex traits in microbes can be readily accomplished without the need to develop molecular genetic tools.

COMPARATIVE GENOMICS OVER 50 NEWLY-SEQUENCED SPECIES OF PARASITIC WORMS

Diogo M Ribeiro¹, Avril Coghlan¹, Nancy Holroyd¹, Eleanor Stanley¹, Jason Tsai¹, Bhavana Harsha¹, Makedonka Mitreva², James Cotton¹, Matthew Berriman¹

¹Wellcome Trust Sanger Institute, Hinxton, United Kingdom, ²McDonnell Genome Institute, Washington University, St. Louis, MO

Parasitic worms (a.k.a. helminths) are a group of Nematodes (roundworms) and Platyhelminthes (flatworms) that infect as much as 24% of the human population worldwide, as well as devastate crops and livestock. Helminths are therefore a major cause for human morbidity (>100 million disability-adjusted life years) and of vast socio-economic importance.

The 50 Helminth Genome Initiative aims to create a genomics framework to expand study into this field, by assembling draft genomes of over 50 relevant parasitic helminthes for which genomic data was previously unavailable.

Here, we compare the genomes of free-living nematodes and Platyhelminthes to the newly sequenced parasitic species and other previously published helminths. Together, our dataset encompass the most life-impairing helminth diseases, such as the causing agents for river blindness, schistosomiasis (major cause of reduced growth and cognitive development in children), cystercercosis, elephantiasis and trichuriasis, as well as the most important nematode plant-pathogens and livestock pathogens.

Exploring a dataset of 81 helminth genomes (which range from 40 Mbp to 1250 Mbp) and 10 metazoan outgroups, comprising over 1.5 million genes, we aim to deepen our understanding on how parasites adapted to parasitism (e.g. development of host infection, tissue invasion, evasion of immune system, feeding) and find biological particularities that can be exploited for the development of new drugs. Analysis of such a large and novel dataset is riddled with computational and biological challenges which we tackle through several bioinformatic approaches, including development of pipelines for filtering noisy data, data mining, and employment of large-scale comparative and functional genomics frameworks.

We scanned our huge dataset for evolutionary insights regarding the origins of parasitism by 1) clustering genes into (>100.000) families and recreate their evolutionary history using the EnsemblCompara pipeline; 2) develop custom scripts to detect gene expansions in specific subclasses, clades and in specific species, as well as gene families with dissimilar gene distributions; 3) functionally annotate the genes and address the evolution and potential function of those families, in the light of parasitism. Moreover, we defined sets of nematode and platyhelminth core and phylum-specific genes and identified the presence or absence of the machinery involved in cytosine and adenine DNA methylation and RNA interference, mechanisms largely unknown in most helminth species.

GENOME-WIDE INFERENCE OF NATURAL SELECTION ON REGULATORY SEQUENCES IN THE HUMAN GENOME.

Brad Gulko¹, Ilan Gronau², Melissa J Hubisz⁴, Adam Siepel³

¹Cornell University, Graduate Field of Computer Science, Ithaca, NY, ²Interdisciplinary Center (IDC) Herzliya, Efi Arazi School for Computer Science, Herzliya, Israel, ³Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, ⁴Cornell University, Department of Biological Statistics & Computational Biology, Ithaca, NY

For decades, it has been hypothesized that gene regulation has played a central role in human evolution, yet much remains unknown about the genome-wide impact of regulatory mutations. In this talk, I will describe work my group has been pursuing over the last few years to better characterize the evolution of gene regulatory elements in humans and nonhuman primates, based on patterns of polymorphism and divergence in complete genome sequences. First, I will review a new probabilistic method, called INSIGHT, that we developed to measure the influence of selection on collections of short, interspersed noncoding elements across the genome. Using INSIGHT, we showed that natural selection has profoundly influenced human transcription factor binding sites since the divergence of humans from chimpanzees 4-6 million years ago, and that regulatory elements contribute substantially to both adaptive substitutions and deleterious polymorphisms, with key implications for human evolution and disease. Next, I will describe how we have adapted the INSIGHT framework for use in estimating the probability that a point mutation at each position in a genome will influence fitness. These fitness consequence (fitCons) scores serve as evolution-based measures of potential genomic function. We have generated fitCons scores for three human cell types based on public data from ENCODE. Compared with conventional conservation scores, fitCons scores show considerably improved prediction power for cis-regulatory elements. In addition, they indicate that 4.2–7.5% of nucleotides in the human genome have influenced fitness since the human-chimpanzee divergence, and they suggest only modest impact from recent evolutionary turnover on the functional content of the genome. Finally, I will describe our recent work on extending fitCons to accommodate much larger collections of genomic covariates. These methods are based on an information theoretic measure of the additional information obtained from each candidate functional genomic data type, which serves as an objective function in a hierarchical divisive clustering algorithm. Preliminary results indicate that these methods substantially improve the genomic resolution and predictive power of the fitCons scores.

A LARGE SCALE PREDICTION OF BACTERIOCIN GENE BLOCKS SUGGESTS A WIDE FUNCTIONAL SPECTRUM FOR BACTERIOCINS

James T Morton¹, Stefan D Freed², Shaun W Lee³, Iddo Friedberg⁴

¹Miami University, Department of Computer Science and Software Engineering, Oxford, OH, ²University of Notre Dame, Department of Biological Sciences, South Bend, IN, ³University of Notre Dame, Department of Biological Sciences, South Bend, IN, ⁴Iowa State University, Veterinary Microbiology and Preventive Medicine, Ames, IA

Bacteriocins are peptide-derived molecules produced by bacteria, whose recently-discovered functions include virulence factors and signalling molecules as well as antibiotics. Their large spectrum of roles makes them molecules of primary importance in basic and medical microbiology, and of strong interest as possible drugs in the post-antibiotic era. To date, close to five hundred bacteriocins have been identified and classified. Recent discoveries have shown that bacteriocins are highly diverse and widely distributed among bacterial species. Given the heterogeneity of bacteriocin compounds, many tools struggle with identifying novel bacteriocins due to their vast sequence and structural diversity. Many bacteriocins undergo numerous post-translational processing or modifications necessary for the biosynthesis of the final mature form. Enzymatic modification of bacteriocins as well as their export from the cell is achieved by proteins whose genes are often located in a discrete gene neighborhoods proximal to the bacteriocin precursor gene, referred to as *context genes* in this study. Although bacteriocins themselves are structurally diverse, context genes have been shown to be largely conserved across unrelated species. Using this knowledge, we set out to identify new candidates for context genes which may clarify how bacteriocins are synthesized, and to identify new candidates for bacteriocins that bear no sequence similarity to known toxins. To achieve these goals, we have developed a novel method to predict bacteriocin locations, the **B**acteriocin **O**peron and Gene Block **A**ssociator (BOA) that can identify homologous bacteriocin associated gene clusters and predict novel ones. We discover that several phyla have a strong preference for bacteriocin genes, suggesting distinct functions for this group of molecules.

GENOMIC ASSEMBLY AND ANALYSIS OF HIGHLY HETEROZYGOTIC POLYPLOIDAL PARASITIC PROTISTS

Aaron R Jex¹, Staffan Svard², Robin B Gasser¹

¹University of Melbourne, Veterinary and Agricultural Sciences, Parkville, Australia, ²Uppsala University, Cell and Molecular Biology, Uppsala, Sweden

A high level of genetic heterozygosity is a significant challenge in the de novo assembly and analysis of the genomes of any organismal group or tissue. For research of parasitic protists, high heterozygosity is a major obstacle to generating highly complete assemblies and exploring complex gene families and repetitive sequence regions (e.g., telomeres) where important virulence genes are often found. An inability to culture many parasitic protists in a cell free system coupled with their microscopic size is a major contributor to this challenge, as it prevents the generation of clonal lines for sequencing and precludes sequencing from individual parasites, at least in any practical sense. A further major challenge is that many parasitic protists are polyploid and some, for example the diarrhoeal pathogen *Giardia duodenalis*, are binucleate parasites that undergo an unknown rate of internuclear recombination and genetic homogenisation. Using existing bioinformatic approaches, we present a strategy for assembling non-redundant and highly complete genomes from heterozygotic and polyploid templates. We utilize this approach to explore genetic and haplotypic diversity in *Giardia duodenalis* isolates sequenced using PacBIO or Illumina methodologies and explore important aspects of its biology, including the evolution of complex gene families and relative rates of sexual and asexual recombination.

A NOVEL SOFTWARE TOOL AND PIPELINE FOR THE CLASSIFICATION OF META-GENOMICS SEQUENCING DATA, AND THEIR APPLICATION TO THE DIAGNOSIS OF NEUROPATHOLOGICAL INFECTIONS OF THE NERVOUS SYSTEM

Florian P. Breitwieser¹, Daehwan Kim¹, Li Song², Carlos A Pardo², Steven L Salzberg^{1,2,3}

¹Johns Hopkins University, Center for Computational Biology, Baltimore, MD, ²Johns Hopkins Hospital, Department of Neurology, Baltimore, MD, ³Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, ⁴Johns Hopkins University, DepartmentsDepartment of Computer Science, Baltimore, MD

Whole-metagenome sequencing is revolutionizing clinical diagnosis and tracing of pathogens in infectious diseases. Powerful software tools such as Kraken, MetaPhlAn and MEGAN/DIAMOND enable microbe identifications from the sequencing data. However, the interpretation of the results for clinical diagnosis requires careful studying of the background of the samples, and validation of potential hits. We present a pipeline that aims at enabling better interpretation by capturing clinical data, running metagenomics analysis through a diverse range of tools, and visualizing and comparing the results across samples.

Furthermore, we present a novel software tool for species-level metagenomics identification, Centrifuge. While the applicability of many metagenomics tools is limited in terms of memory requirements, run-time, or sensitivity, Centrifuge can be run on desktop machines with less than 8GB of RAM (includes the human genome, and all completed bacterial and viral genomes), has comparable or better sensitivity and precision than state-of-the art tools, and a run-time less than twice of Kraken. The system uses a novel indexing scheme based on the Burrows-Wheeler transform and the FM index, optimized specifically for the metagenomic classification problem.

We utilized and validated Centrifuge and the pipeline on a series of patients with neurological symptoms indicating possible infections. Direct sequencing of brain biopsies generated 8.3 million to 29.1 million sequence reads per sample, which allowed us to successfully identify the infectious agent in 3 out of 10 patients, demonstrating the power of large-scale unbiased sequencing as a novel diagnostic tool. Separate clinical assays confirmed each of the three positive cases.

COMPREHENSIVE PROFILING OF SOMATIC MOSAICISM IN HUMAN BRAIN

Taejeong Bae^{1,2}, Flora M Vaccarino^{3,4}, Alexej Abyzov^{1,2}

¹Mayo Clinic, Center for Individualized Medicine, Rochester, MN, ²Mayo Clinic, Department of Health Sciences Research, Rochester, MN, ³Yale University, Child Study Center, New Haven, CT, ⁴Yale University, Department of Neurology, New Haven, CT

With emerging evidences, it is becoming apparent that each cell in the human body has its own genome, a phenomenon called as somatic mosaicism. Such somatic variations include single nucleotide variants (SNVs), small insertions and deletions (indels), transposable element insertions (TEI) and larger copy number variations (CNVs) and structural variations (SVs). Even though somatic mosaicism may have functional and pathological implication, there is no comprehensive estimate of the number and allelic frequency of genomic variations in normal somatic cells in various tissues of human body. Difficulty in detecting somatic mosaic variants is due to them present in a small proportion (could be as low as fraction of a percent) of cells within a tissue. To circumvent this problem, we adopted an approach to sequence the genome of clonal cell populations derived from single brain progenitor cells and, thereby, identify genomic variations present in the founder cell and manifested in each clone at 50% allele frequency. This approach is also advantageous over single cell sequencing as it is free of amplification artifacts. For data analysis we developed a workflow to synergize calls from popular somatic variant calling programs: MuTect, SomaticSniper, Strelka and VarScan for SNVs, Scalpel, Strelka, and VarScan for indels, and CNVnator for CNVs. By applying the workflow to compare germline genomes of different individuals we performed data driven estimation of its (workflow's) sensitivity. By applying it to real data for 6 clones from an individual healthy brain, we detected 200-500 SNVs per clone with over 75% sensitivity, 10-30 indels per clone at over 40% sensitivity, and 1-5 CNVs per clone. Independent validation revealed specificity of nearly 100% of the generated calls. This analysis uncovers extensive somatic mosaicism existing in human brain.

DECOUPLING ARRAY CGH SAMPLE-REFERENCE HYBRIDIZATION PAIRS FOR NORMALIZATION OF LOG RATIO ARTEFACTS.

Aled R Jones¹, Kevin J Ryan², Adele Corrigan³, Joo Wook Ahn¹

¹Applied Bioinformatics, Guy's & St Thomas' NHS Foundation Trust, London, United Kingdom, ²Applied Bioinformatics, Viapath, London, United Kingdom, ³Genetics Laboratories, Viapath, London, United Kingdom

Array CGH is a well established technique that is used in clinical genetics laboratories for detection of copy number variation. A DNA sample and a reference sample are labeled fluorescently with cyanine 3 and 5 respectively, before being hybridized together to an array of oligonucleotide probes. Following hybridization, fluorescence intensities are measured and normalized, before $\log(2)$ ratios are calculated.

Copy number variants (CNVs) are then detected using an algorithm that calls genomic regions that are significantly distinct from those with a $\log(2)$ signal intensity ratio of 0; regions of normal diploid copy number in the sample have equal fluorescence to the reference resulting in a $\log(2)$ ratio of 0. However, fluorescent labeling of the sample and reference can vary in efficiency when the two are extracted from differing starting materials (e.g. biopsied material from an embryo as the sample and cultured cells as the reference) and/or extraction methods. This causes $\log(2)$ ratios to deviate from a value of 0 and can reduce the performance of CNV calling. Therefore, matching appropriate samples and references is crucial for optimizing array CGH performance; however, it is not always possible to predict this behaviour and furthermore, an appropriate reference sample may not always be available in a clinical diagnostic setting where samples are received from multiple sources.

We present the Embryo vs Embryo algorithm (EvE), which enables post-hoc pairing of hybridization partners, allowing the opportunity to optimize sample-reference pairs and normalize skewed $\log(2)$ ratios. We have validated and successfully applied EvE to array CGH data generated from embryos at the Assisted Conception Unit at Guy's Hospital, demonstrating its potential to improve results when testing embryos from couples carrying balanced translocations.

5-HYDROXYMETHYLCYTOSINE IN *DAPHNIA PULEX*

Gediminas Alzbutas^{1,2}, Dovile Strepetkaite¹, Eimantas Astromskas¹, Rasa Sabaliauskaite¹, Kestutis Arbaciauskas⁴, Arunas Lagunavicius¹, Juozas Lazutka³

¹Thermo Fisher Scientific, Research and Development, Vilnius, Lithuania,

²Vilnius University, VU Institute of Biotechnology, Vilnius, Lithuania,

³Vilnius University, Faculty of Natural Sciences, Vilnius, Lithuania,

⁴Nature Research Center, Laboratory of Evolutionary Ecology of Hydrobionts, Vilnius, Lithuania

We report discovery of 5-hydroxymethylcytosine in DNA of *Daphnia pulex* clone of European origin (established from the pond in Lithuania) along with its draft genome. The presence of 5-hydroxymethylcytosine was revealed by two methods: i) dot blot analysis using the anti-5-hmC antibody, ii) sequencing of DNA enriched for 5-hmC (EpiJET 5-hmC Enrichment Kit, Thermo Fisher Scientific). It was discovered that 5-hmC distribution across genome is not random: in exons it is twice as frequent as in introns. In the draft genome genes were identified using GlimmerHMM that was trained on related, already published genome of the "TCO" clone of American origin (identity ~ 93 %). Detected protein coding genes were annotated with PANNZER. Then we searched for relations between the predicted functions of genes and the density of hmC in their exons. The findings indicate that high density of 5-hmC is associated with genes that are involved in G-protein signalling pathways and are related to endocrine system, molting cycle. On the other hand, exceptionally low 5-hmC density is found in genes of ribonucleoproteins and genes of proteins that are predicted to be related with TGF-beta signalling pathway.

INTEGRATED ANALYSIS OF NUMEROUS HETEROGENEOUS GENE EXPRESSION PROFILES FOR DETECTING ROBUST DISEASE-SPECIFIC BIOMARKERS AND PROPOSING DRUG TARGETS

David Amar¹, Tom Hait¹, Shai Izraeli^{2,3}, Ron Shamir¹

¹Tel Aviv University, The Blavatnik School of Computer Science, Tel Aviv, Israel, ²Sheba Medical Center, Hematology-Oncology, Safra Children's Hospital, Ramat Gan, Israel, ³Tel Aviv University, Sackler School of Medicine, Tel Aviv, Israel

Genome-wide expression profiling has revolutionized biomedical research; vast amounts of expression data from numerous studies of many diseases are now available. Making the best use of this resource in order to better understand disease processes and treatment remains an open challenge. In particular, disease biomarkers detected in case-control studies suffer from low reliability and are only weakly reproducible. Here, we present a systematic integrative analysis methodology to overcome these shortcomings. We assembled and manually curated more than 14,000 expression profiles spanning 48 diseases and 18 expression platforms. We show that when studying a particular disease, judicious utilization of profiles from other diseases and information on disease hierarchy improves classification quality, avoids overoptimistic evaluation of that quality, and enhances disease-specific biomarker discovery. This approach yielded specific biomarkers for 24 of the analyzed diseases. We demonstrate how to combine these biomarkers with large-scale interaction, mutation and drug target data, forming a highly valuable disease summary that suggests novel directions in disease understanding and drug repurposing. Our analysis also estimates the number of samples required to reach a desired level of biomarker stability. This methodology can greatly improve the exploitation of the mountain of expression profiles for better disease analysis.

OPTIMIZATION OF *DE NOVO* TRANSCRIPTOME ASSEMBLY AND DIFFERENTIAL EXPRESSION ANALYSIS OF SALT TOLERANCE GENES IN THE HALOPHYTE *SUAEDA FRUTICOSA*

Joann Diray-Arce¹, Mark J Clement², Bilquees Gul³, M Ajmal Khan⁴, Brent L Nielsen¹

¹Brigham Young University, Department of Microbiology and Molecular Biology, Provo, UT, ²Brigham Young University, Department of Computer Science, Provo, UT, ³University of Karachi, Institute of Sustainable Halophyte Utilization, Karachi, Pakistan, ⁴Qatar University, College of Arts and Sciences, Doha, Qatar

Several studies of naturally occurring salt-tolerant plants called halophytes have suggested their economic potential as crops. They can endure high amounts of salt in their systems because of their unique adaptation mechanisms at the physiological and molecular levels. While the physiological and biochemical properties of some halophyte species have been characterized, little is known about the molecular basis for salinity tolerance in other species. Our research focuses on an obligate halophyte species, *Suaeda fruticosa* that grows optimally at 300 mM NaCl and has the ability to sequester salts into vacuoles to reduce buildup of ions for long-term survival. This study reports the optimization of de novo transcriptome assemblies of *Suaeda fruticosa* using various algorithms necessary to convert RNA-seq data into a usable de novo transcriptome. We trimmed and normalized the reads for pre-assembly quality check and used Trinity and Velvet-Oases to generate multiple assemblies and compared their mapping efficiencies using GSNAP. Optimization of the assemblies was performed and compared using clustering methods (CAP3, CDHIT-EST and Isorefuse) that reduced the number of transcripts while retaining the mapping coverage. We provided a superior clustering algorithm of splice variants that can be used to improve the usability of a transcriptome. This maximizes the coverage of reads while reducing the number of transcripts without losing important information needed for de novo transcriptome assembly. We performed differential expression analysis on our most preferred assembly and found 519 differentially expressed genes are either up- or down-regulated when comparing plants grown with optimal salt and in the absence of salt. We have annotated these sequences and supported our data with qRT-PCR validation of selected genes involved in salt tolerance. Phylogenetic analysis for *Suaeda fruticosa* was performed and compared with other plant species using InParanoid and MAFFT. These data provide a resource for the discovery of potential genes important for salt tolerance in this species and may serve as a reference for other succulent halophyte plants.

MOLECULAR DELINEATION OF TWO MAJOR ONCOGENIC PATHWAYS GOVERNING INVASIVE DUCTAL BREAST CANCER DEVELOPMENT

Luay Aswad^{1,2}, Surya P Yenamandra¹, Ow Ghim Siong¹, Anna V Ivshina¹, Vladimir A Kuznetsov^{1,2}

¹Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore 138671, Singapore, ²Nanyang Technological University, School of Computer Engineering, Singapore 637553, Singapore

Abstract

Invasive ductal carcinoma (IDC) is the major histo-morphologic type of breast cancer. IDC is diagnosed in 800,000 women per year in USA. Histological grading (HG) of IDC is widely adopted by oncologists as a prognostic factor. However, HG evaluation is highly subjective, with only 50%-85% inter-observer agreements. Specifically, the subjectivity in the assignment of the intermediate grade (grade 2, HG2) breast cancers (comprising ~50% of IDC cases) results in uncertain disease outcome predictions and sub-optimal systemic therapy. Despite several attempts have been made to identify the mechanisms underlying the HG classification systems, their molecular bases are poorly understood. Herein, after analyzing TCGA and other integrative genomic and clinical datasets, we found the 22-genes tumor aggressiveness grading classifier (22g-TAG) reflecting a global bifurcation in the IDC transcriptomes resolving the patients with HG2 tumors into two genetically and clinically distinct subclasses: histological grade 1-like (HG1-like) and histological grade 3-like (HG3-like). The gene expression profiles and clinical outcomes of these two subclasses are similar to the HG1 and HG3 tumors, respectively. We further reclassified IDC into the low genetic grade (LGG=HG1+HG1-like) and the high genetic grade (HGG=HG3-like+HG3) tumor classes. We demonstrated the validity of this dichotomization on the genome scale by extensive integrative data analyses and suggested independent oncogenic pathways for each tumor class. For the low- and high- aggressive HG2 subclasses we found subclass-specific DNA alterations, mutations, oncogenic pathways, cell cycle/mitosis and stem cell-like expression signatures discriminating HG1-like and HG3-like tumors. We found similar molecular patterns in the LGG and HGG tumor classes. Our results suggest the existence of two genetically-predefined IDC classes (LGG and HGG), driven by distinct oncogenic pathways. They provide novel prognostic and therapeutic biomarkers and could open unique opportunities for personalized neo-adjuvant and adjuvant systemic therapy of IDC patients.

DETECTION OF PATHOGEN INTEGRATION SITES IN CANCER

Gnanaprakash Balasubramanian^{1,2,3}, Barbara Hutter^{1,2,3}, Hamza Khan^{1,4},
Benedikt Brors^{1,2,3}

¹G200 Applied Bioinformatics, German Cancer Research Institute (DKFZ), Heidelberg, Germany, ²Research Program Translational Cancer Research, National Center for Tumor Diseases (NCT), Heidelberg, Germany, ³German Consortium for Translational Cancer Research (DKTK), Heidelberg, Germany, ⁴Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

Viruses are causal agents in various debilitating diseases including certain cancers. Globally 12% of the diagnosed malignancies have been associated with an oncovirus. Characterizing the viral infectious agent associated with cancer is of great importance in treating this dreadful disease. We have developed a method named D-ViSioN (Detection of Integration of Virus(s) by SingletoN(s)) that predicts viral integration sites based on Next Generation Sequencing. Our method also gives an overview of the associated viral infectome. D-ViSioN can be applied as a cancer pathogen discovery tool to decipher novel viral associations. We have applied this method to the whole genome sequencing data of Epstein Barr virus (EBV) transformed immortalized lymphoblastic cells. We have successfully obtained the EBV integrome and infectome of the cell line. Our method correctly detects the experimentally known integration sites of EBV. Genes affected by viral integration are involved in various basic cell processes like cell cycle control, cytoskeleton remodelling, vesicular traffic and cell communication. Thus, D-ViSioN gives a glimpse of the process of cell immortalization and gives valuable clues to decipher the phenomenon of oncogenic transformation. D-ViSioN is routinely used to characterize Viromes of two major Precision (Personalized) Oncology projects of the German Cancer Research Institute (DKFZ) namely INFORM (INdividualized Therapy FOR Relapsed Malignancies in Childhood) and HIPO-021 (Prospective genetic characterization of tumors from individual patients of special interest).

AUPAIRWISE: BIOLOGICALLY FOCUSED RNA-SEQ QUALITY CONTROL USING CO-EXPRESSION

Sara Ballouz, Jesse Gillis

CSHL, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY

A principal claim for RNA-sequencing has been greater replicability, typically measured in sample-sample correlations of gene expression levels. Replicability of transcript abundances in this way will provide misleading estimates of the replicability of conditional variation, which is what is of interest in expression analyses. Heuristics which implicitly address this problem have emerged in quality control measures to obtain ‘good’ differential expression results. However, these methods involve strict filters such as discarding low expressing genes or using technical replicates to remove discordant transcripts, and are costly or simply ad hoc.

As an alternative, we’ve performed a series of sample and replicate-based analyses of RNA-seq data and show that gene-gene correlations of expression levels across conditions and between replicates are a more useful measure of replicability. Through the re-analysis of reference RNA-seq expression sets, we are able to recapitulate SEQC guidelines for replicability, but using substantially less data. Thus, our gene-level method of replicability allows for flexibility and specific tailoring of the analyses in even small data sets that cannot be achieved with the general recommended filters. We then show that this gene-level replicability of differential activity can be modeled in a co-expression framework, using known co-expressing gene pairs as pseudo-replicates instead of true replicates. We use this as a quality control metric: by modelling the effects of noise that perturbs a gene’s expression, we can then measure the aggregate effect of this perturbation on these co-expressing gene-pairs (ie. ‘housekeeping interactions’). Perturbing expression by only 5% is readily detectable (AUROC~0.73), which makes this a straightforward method to help customize experiments. We have named this method AuPairWise, and is available as a set of R scripts (github.com/sarbal/AuPairWise).

Our ongoing assessment is to apply AuPairWise as a method of validating methodological pipelines on real data. In particular, we are testing the parameter search space of the RNA-seq alignment STAR, to find which choices gives the output most closely aligned to known biology (as determined through expected housekeeping co-expression), an approach differing substantially from ones using simulated datasets.

INTEGRATED GENOMIC ANALYSIS SUGGESTS INHERITED PREDISPOSITION TO CANCER IN THERAPY-RELATED ACUTE LYMPHOBLASTIC LEUKEMIA (TR-ALL)

Riyue Bao¹, Fuhong He², Mark M Sasaki³, Jane E Churpek⁴, Lei Huang¹, Qianfei Wang², Jorge Andrade¹, Kenan Oneil³

¹The University of Chicago, Center for Research Informatics, Chicago, IL, ²Beijing Institute of Genomics Chinese Academy of Sciences, Laboratory of Genomic and Precision, Beijing, China, ³The University of Chicago, Department of Pediatrics, Chicago, IL, ⁴The University of Chicago, Department of Medicine, Chicago, IL

Therapy-related acute lymphoblastic leukemia (tr-ALL) is a rare secondary leukemia following treatment for primary malignancies. However, at least 15% cases occur in the absence of antecedent genotoxic therapy, suggesting that a large proportion may not be therapy-induced, but associated with inherited predisposition to cancer. To test this hypothesis, we obtained the family history for nine tr-ALL patients, and found that all met at least loose criteria for Li-Fraumeni Syndrome (LFS), a hereditary cancer predisposition syndrome caused by a germline mutation in *TP53*. To determine whether any had high penetrance cancer-predisposing mutations, we performed whole exome sequencing (50x or greater) on five patients with available DNA. We discovered that one patient harbors a *bona fide TP53* germline cancer-predisposing mutation (c.1015G>T, p.E339*, exon 10), where the premature stop-codon leads to truncation of the protein oligomerization domain at the C-terminal. Of note, loss of heterozygosity (LOH) occurs at the same position in the leukemia tissue of the same patient. We identified 2,146 rare exonic germline variants (MAF<0.001) in 804 genes, including 14 in known cancer predisposition genes. Pathway analysis suggested that those mutated genes are significantly enriched in MAPK signaling, DNA double-strand break repair by homologous recombination, and BRCA1-regulated DNA damage response (p<0.01). Interestingly, we discovered a highly connected gene network consisting of 20 core members centered at *TP53* and *BRCA1/2* carrying rare deleterious germline mutations, suggesting compromised p53 and BRCA function at network level. These data suggest that a subset of patients previously thought to have tr-ALL may actually have LFS, and that tr-ALL may be a previously unsuspected component of LFS. Additionally, our studies suggest that there may be a network of variants that together deregulate BRCA1/p53 signaling, thereby predisposing individuals to ALL either by themselves or in cooperation with exposure to genotoxic chemotherapy.

FROM THE GROUND TO THE CLOUD IN JUST MINUTES: BUILDING A CUSTOMIZED GALAXY ANALYSIS SERVER USING ONLY A WEB BROWSER

Daniel Blankenberg^{1,2}, The Galaxy Team²

¹Penn State University, Department of Biochemistry and Molecular Biology, University Park, PA, ²The Galaxy Project, <https://galaxyproject.org>, Everywhere, PA

Galaxy (<https://galaxyproject.org>) is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. Galaxy makes bioinformatics analyses accessible to users lacking programming experience by enabling them to easily specify parameters for running tools and workflows. Analyses are made transparent by allowing users simple access to share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis. Extending Galaxy with new tools, visualizations, datasources, and external resources has been designed to be a plug-n-play process.

The Galaxy Project provides a free to use Galaxy service at <https://usegalaxy.org> that allows any researcher to upload their own data and perform complex analyses. Additionally, over 70 other public servers, not supported by the Galaxy Project, are also available. Although these public Galaxy instances have been proven to enable real research, with hundreds of papers citing their use, there are varying limits imposed on storage size, compute resources, tool availability, and data access restrictions that can prevent successful completion of a project on a public site.

Galaxy has been designed to be deployable on local compute resources and virtualized infrastructure, including cloud-style resources, with minimal effort. However, one will still need to install the tools and reference data that are required to perform meaningful analyses. To address these needs, we have developed the Galaxy ToolShed and Data Manager framework. The Galaxy ToolShed functions as an AppStore for Galaxy, allowing Galaxy administrators to install new Galaxy tools and utilities, along with their binary dependencies, through a point-and-click interface in real-time. In addition to tools, a special type of utility, known as Galaxy Data Managers, allows administrators to manage a particular Galaxy instance's built-in cache of reference datasets using a web-based interface. Data Managers are installable from the Galaxy ToolShed. Data Managers can build reference data, such as indexes for mapper tools, or, when available, can make use of precomputed indexes by downloading.

Here, we demonstrate just how easy it is to run your own Galaxy server with a custom set of tools and reference datasets. We will begin from a standard laptop computer, launch a Galaxy instance on the Cloud, download and install required tools and dependencies, acquire and build reference datasets and indexes, and perform a computationally intensive analysis in less than 20 minutes.

A MODIFIED MAKER STRUCTURAL GENOME ANNOTATION METHOD REVEALS NOVEL GENE PREDICTIONS OF HIGH AND LOW GC CONTENT IN RICE

Megan J Bowman, Jane A Pulman, Kevin L Childs

Michigan State University, Plant Biology, East Lansing, MI

Accurate structural genome annotation depends on precise gene model prediction. Hidden markov models (HMMs) are used in the structural annotation process to mathematically predict specific genic regions and are dependent on training sets of genes to identify gene structure boundaries. In general, HMMs are trained on a random selection of genes that may represent the variance in gene structure found in a given genome. One example of this variance is GC content, which differs across plant species and is bimodal in many grass genomes. This bimodality presents a challenge in accurately predicting gene structure using HMMs. We hypothesize that grass genes with extreme GC content would not be well predicted with an HMM trained on a random selection of genes. To address this, we developed training sets of genes with both high or low GC content from the genome of rice (*Oryza sativa* ssp. *Nipponbare*) for the ab initio prediction programs SNAP and AUGUSTUS. The HMMs developed through this work were employed by the MAKER annotation pipeline to predict 1,598 new evidence-supported gene models as compared to the default MAKER protocol. The gene predictions identified by high and low GC HMMs were compared to MAKER genome annotations with non-GC specific HMMs and then combined through MAKER to create a new structural annotation. Additionally, tissue specific gene expression and codon usage bias analyses have identified important biological functions for these novel evidence supported gene predictions in rice. This work has enhanced the structural annotation of the rice genome and has created a novel approach to GC-specific HMM development for use with the MAKER annotation pipeline.

A SOFTWARE TOOL FOR DATA INTEGRATION IN A DIAGNOSTIC LABORATORY

Riccardo Brumm¹, Sebastian H Eck¹, Betina Ebert², Ina Vogl³, Sandra Kuecuek¹, Sabine Rath¹, Verena Hasselbacher¹, Christina Sofeso¹, Birgit Busse¹, Soheyla Chahrokh-Zadeh¹, Christoph Marshall¹, Karin Mayer¹, Imma Rost¹, Hanns-Georg Klein¹

¹Center for Human Genetics and Laboratory Diagnostics Dr.Klein, Dr.Rost and Colleagues, Munich, Germany, ²Steag & Partner AG, St. Gallen, Switzerland, ³GENOLYTIC GmbH, Leipzig, Germany

The implementation of Next-Generation Sequencing in a clinical diagnostic setting opens vast opportunities through the ability to simultaneously sequence all genes contributing to a certain indication at a cost and speed that is superior to traditional sequencing approaches. Especially in the case of rare, heterogeneous disorders this may lead to a significant improvement in diagnostic yield. On the other hand, the practical implementation of NGS in a clinical diagnostic setting involves a variety of new challenges which need to be overcome. Among these are the generation, analysis and storage of unprecedented amounts of data, strict control of sequencing performance, validation of results, interpretation of detected variants and reporting. Exonic regions of more than 500 custom selected genes are enriched in parallel by oligonucleotide hybridization and capture (Agilent QXT), followed by massively-parallel sequencing on the Illumina NextSeq instrument. During data analysis, only genes from the requested indication (grouped in subpanels) are selected to limit interpretation to relevant genes, while simultaneously minimizing the possibility of unsolicited findings. Data analysis is performed using the CLC Genomics Workbench (v.8.0.3 CLCbio) and custom developed Perl scripts. Target regions which fail to reach the designated coverage threshold of 20X are re-analyzed by Sanger sequencing. Additionally, identified candidate mutations are independently confirmed. All detected variants are imported into an in-house relational database scheme which may be queried via a web interface for dynamic data analysis and filtering. Information from all 500 genes is used in an anonymized way for internal variant frequency calculation, quality control and the detection of potential sequencing artifacts. In order to integrate NGS data within our diagnostic laboratory we developed a software tool, MIDAS. MIDAS integrates patient data from our LIMS system, data from the routine Sanger sequencing workflow as well as phenotype data, based on the Human Phenotype Ontology (HPO) with the NGS results. In particular, Genotype-Phenotyp relations identified in one patient are made available for all other cases to aid the interpretation and build a comprehensive knowledge base.

We have applied this approach to more than 350 samples from a variety of different genetic disorders. We use the outlined approach for the diagnostics of arrhythmogenic cardiac disorders (LQTS, HCM, DCM), connective tissue disorders (EDS, TAAD), rare kidney disorders (Nephrotic Syndrome, CAKUT), neurological disorders (Noonan syndrome, Microcephalies), metabolic disorders (MODY diabetes) and coagulopathies.

GMOD IN THE CLOUD 2.0

Scott Cain¹, Stephen Ficklin², Colin Diesh³, Lacey Sanderson⁴, Lincoln Stein¹

¹Ontario Institute for Cancer Research, Stein Lab, Toronto, Canada, ²Washington State University, Main Lab, Pullman, WA, ³University of Missouri, Elsik Lab, Columbia, MO, ⁴University of Saskatchewan, Plant Sciences, Saskatoon, Canada

GMOD in the Cloud is a virtual server, available through the Amazon compute cloud, equipped with a suite of preconfigured GMOD components and GMOD in a Box is an identical implementation in a VirtualBox virtual machine. The 2.0 versions of both include updated releases of several GMOD components including a Chado 1.3 database, JBrowse, Tripal 2.0, and WebApollo 2.0. Users can clone the GMOD Amazon Machine Image (AMI) or download the VirtualBox implementation from the GMOD ftp site, and either way create their own server for storing data and making it available to the public or a select group of users. Making use of virtualization--hosting data and/or applications on existing networked computer systems--to give users access to preconfigured, extensible servers is an alternative to building and maintaining large computing infrastructure in-house. Potential applications of the GMOD images range from short term usage, such as during annotation jamborees, to the long term provision of access to genome data and applications to the community.

GALAXY TOOL WORLD PROGRESSION:HAPPIER DEVELOPERS, HAPPIER USERS

John Chilton¹, Martin Čech¹, Björn Grüning², Eric Rasche³, Galaxy Team^{1,4}

¹Penn State University, Department of Biochemistry and Molecular Biology, University Park, PA, ²University of Freiburg, Department of Computer Science, Freiburg, Germany, ³Texas A&M University, Center for Phage Technology, College Station, TX, ⁴Johns Hopkins University, Department of Biology, Baltimore, MD

The Galaxy Project provides a free platform for domain agnostic data intensive research. The number of Galaxy instances maintained around the globe is continuously rising (the yearly number of public servers starting 7/2011: 15-20-35-60-73) often being in very diverse settings (compare e.g. GalaxyP, deepTools, CoSSci, Oqtans, VirAmp, Osiris) ultimately bringing more users with wider demands to the project. This presents substantial challenges in tool development, testing, discovery, and distribution.

Galaxy uses the Tool Shed (TS) as an App Store-like platform for tool exploration and deployment with reproducible workflow sharing support. Today the TS contains more than 3,400 tools from different areas of computational research and a vigorous community is maintaining and revising these tools.

Driven by community feedback solicited via questionnaire¹ we identified and focused on the areas that would benefit the most from improvements:

Tool Shed - by providing integration with Jenkins and Github we have simplified and automated the tool publishing process.

Tool testing - we built the CLI toolkit Planemo that makes tool creation and testing dramatically easier.

Tool discovery - we have rewritten search from the ground up to allow deployers to identify high quality tools more easily.

A community based commission (IUC) is maintaining a best practice guide to address the first issue and define high quality standards for Galaxy tools².

We also present Planemo - command-line utility that targets the tool complexity problem and can help the developer boost productivity and improve quality of products. It streamlines the tool development process in general because it enables the developers to create high-class Galaxy tools and lowers the barrier to integrate their tools resulting in more unified interfaces and improved software quality.

Jenkins build scripts leveraging Planemo have been developed to automate testing and deployment of tool repositories providing vital feedback by posting the results back to GitHub. The scripts can also automatically deploy tool repositories to the TS reducing the overhead of managing large number of tools. We firmly believe the presented work will enable the Galaxy community to scale up its already incredible contributions and boost collaboration.

¹ Data available at

<https://wiki.galaxyproject.org/Community/GalaxyAdmins/Surveys/2014>

² <http://galaxy-iuc-standards.readthedocs.org/>

SLNCKY: A SOFTWARE AND NOVEL APPROACH FOR ANNOTATION AND EVOLUTIONARY ANALYSIS OF LNCRNAs

Jenny Chen^{1,2}, XiaoPeng Zhu³, Alexander A Shishkin⁴, Sabah Kadri¹, Itay Maza⁵, Jacob H Hanna⁵, Aviv Regev^{1,6,7}, Manuel Garber^{1,3}

¹Broad Institute of MIT and Harvard, Cambridge, MA, ²Massachusetts Institute of Technology, Division of Health Sciences and Technology, Cambridge, MA, ³University of Massachusetts Medical School, Bioinformatics and Integrative Biology, Worcester, MA, ⁴California Institute of Technology, Department of Biology and Biological Engineering, Pasadena, CA, ⁵Weizmann Institute of Science, Department of Molecular Genetics, Rehovot, Israel, ⁶Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, ⁷Howard Hughes Medical Institute, Chevy Chase, MD

Recent advances in transcriptome sequencing have led to the discovery of thousands of long non-coding RNAs (lncRNAs) across prokaryotes and eukaryotes. Though several lncRNAs have been shown to play important roles in diverse biological processes, the function and mechanism of most lncRNAs remains unknown. Two nontrivial obstacles lie between transcriptome sequencing and functional characterization of lncRNAs: 1) difficulty in identifying truly noncoding genes from *de novo* reconstructed transcriptomes, and 2) prioritizing hundreds of resulting putative lncRNAs from each sample for downstream analysis.

Here, we present a software package, *slncky*, for computational lncRNA discovery: *slncky* filters a *de novo* transcriptome down to transcripts likely to be noncoding and implements a novel approach for evolutionary analysis of lncRNAs. In particular we propose several metrics of transcript and sequence evolution (transcript identity, splice site conservation, indel rate) that reveal distinct classes of lncRNA evolution.

We apply *slncky* to pluripotent RNA-seq data across four mammalian species and use additional functional data to validate *slncky*'s lncRNA predictions. We further apply *slncky* to existing catalogues consisting of >700 RNA-Seq experiments in human and mouse and narrow over 100,000 lncRNAs into ~47,000 high confidence transcripts and ~300 highly constrained lncRNAs which we believe are under purifying selective pressure.

PICKARMSITE, INVESTIGATE THE PREFERENCE OF USING ONE ARM FOR MIRNAS

Tingwen Chen^{1,2}, Cheng-Yang Lee^{1,2}, Petrus Tang^{1,2,3}

¹Chang Gung University, Molecular Medicine Research Center, Taoyuan, Taiwan, ²Chang Gung University, Bioinformatics Center, Taoyuan, Taiwan, ³Chang Gung University, Graduate Institute of Biomedical Sciences, Taoyuan, Taiwan

Up/down regulation of miRNA have been shown to correlate with many diseases including cancer. Mature, functional miRNAs are derived from their pre-miRNAs which can produce two mature miRNA, called 3' arm and 5' arm miRNAs. For many miRNAs, one arm can be detected in a much higher expression level than the other. The differences have been shown to be species specific, tissue specific or developmental stage specific. Many studies have found the arm preferences for miRNAs differ in various cancers and name this phenomenon as "arm switch". Given the differences of the sequences of the two arms, the change in arm preferences can result in changes in the regulation of a large number of downstream target genes. In this study, we wanted to thoroughly investigate these "arm switch" events in 20 different cancer types by utilizing publicly available miRNA expression data from the Cancer Genome Atlas (TCGA). We applied statistical test for miRNA expression values in normal samples and tumor samples to provide a reliable standard to filter for potential "arm switch" events. To further understand the outcome of these events, we also predicted targets for both arms and investigated the functions of these predicted target genes. We finally summarized all these data into our database PickArmSite which not only identifies cases of arm switch events but also suggests consequences of these events. We believe PickArmSite can shed lights on the roles of miRNA in cancer development.

ABUNDANT INVERTED DUPLICATES IN THE HUMAN AND MOUSE GENOMES AS FUNCTIONAL REGULATORY ELEMENTS EVOLVING UNDER SEX-RELATED SELECTION

Zhen-Xia Chen¹, Yong Zhang², Ge Gao³, Brian Oliver¹, Manyuan Long⁴

¹National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, ²Chinese Academy of Sciences, Institute of Zoology, Beijing, China, ³Peking University, College of Life Sciences, Beijing, China, ⁴University of Chicago, Department of Ecology and Evolution, Chicago, IL

Background

Inverted duplicates or repeats can lead to genome instability and many may have no function, but some functional small RNAs are processed from hairpins transcribed from these elements. It is not clear whether the pervasive numbers of such elements in genomes, especially those of mammals, is the result of high generation rates of neutral or deleterious duplication events or positive selection for functionality.

Results

We identified large numbers of inverted duplicates within intergenic regions of two mammalian genomes: 1,932,227 in human and 1,919,926 in mouse. Inverted duplicates were also abundant in 19 other vertebrate genomes. 11% of human and 16% of mouse inverted duplicates are expressed in gonads. Structural characterization of these inverted duplicates revealed significantly longer arms and shorter spacers compared to simulated inverted duplicates in randomly shuffled chromosomal sequences, and the proportion of inverted duplicates forming hairpin structures is much higher than that for mirror duplicates or simulated inverted duplicates, suggesting that inverted duplicates may produce hairpin RNAs. To infer their potential functional roles and detect the involved evolutionary forces, we tested if the well-known signatures of a few forms of sex-related forces of evolution on the functional genetic elements on the X-chromosomes applied to these duplicates. We found that inverted duplicates distributed differently with their ages regarding their X- and autosomal linkage and the strata structures in the X chromosomes. Moreover, we detected a burst of inverted duplicates origination in the vertebrate phylogeny after the emergence of the X chromosome.

Conclusion

These observations, reminiscent of similar distribution and evolutionary dynamics of functional duplicates and known non-coding RNA regulators, are consistent with the hypothesis that intergenic inverted duplicates dispersed across genome may play functional roles as genome wide expression regulators, as manifested in the different dynamics between X- and autosomal linkage.

BUILDING MORE EXPRESSIVE GALAXY WORKFLOWS

John M Chilton¹, Galaxy Team^{1,2}

¹Pennsylvania State University, Biochemistry and Molecular Biology, University Park, PA, ²Johns Hopkins University, Department of Biology, Baltimore, MD

Galaxy is a data analysis platform used to integrate diverse command-line utilities into a unified and user friendly web-based interface. A salient feature of Galaxy is that it allows researchers to extract analysis histories as examples out into reusable workflows. Despite the popularity of this feature, the kinds of workflows that could be expressed by Galaxy in the past have had critical limitations. Perhaps chief among these are that the workflow engine planned out every job right at submission time (leaving little room for dynamic computation) and Galaxy workflows have required a static number of inputs. Many common and relatively simple analyses in bioinformatics require running a variable number of inputs across identical processing steps (“mapping”) and then combining or collecting these results into a merged output (“reducing”). Likewise - many requested advanced features such as pausing workflows, splitting inputs up into multiple datasets, and conditionals all require the re-evaluation of workflows overtime. This work will discuss how we have started addressing these limitations. In particular we will present dataset collections and a real, pluggable Galaxy workflow subsystem - together these features deal with these limitations and vastly increase the expressiveness of Galaxy workflows.

Galaxy dataset collections allow gathering datasets into potentially nested hierarchies of lists and pairs and operating on them as units. Pre-existing Galaxy tools can be used without modification to “map” operations across dataset collections to produce new collections. Likewise tools that consume many datasets can be readily used to “reduce” these collections. For newly developed tools - a wide range of extensions to Galaxy tooling format exist to consume and produce dataset collections. In addition to presenting these additions to Galaxy, extensions to the workflow system to tie together these analyses and innovative UI elements such as the paired list dataset collection builder will be presented.

Specific biologically relevant examples to highlight the power of dataset collections and the new workflow engine will be presented. These will include community developed workflows from various groups for analyses including RNA-seq, phylogenomics, and proteomics.

While this work greatly enhances the expressivity of Galaxy workflows, much remains to do be done. An outline for the future of Galaxy workflow development will be laid out - including conditionals, iteration, nested workflows, and more flexible connections between steps (e.g. mapping output metadata to input parameters for instance).

ASSESSING NEW TOOLS AND BEST PRACTICES FOR RNA SEQ DATA ANALYSIS AND VISUALIZATION WITH IPLANT CYBERINFRASTRUCTURE

Kapeel M Chougule^{1,4}, Andrew Olson¹, Jurandir Vieira De Magalhaes², Peter Van Buren^{1,4}, Liya Wang^{1,4}, Doreen Ware^{1,3,4}

¹Cold Spring Harbor Laboratory, Ware Laboratory, Cold Spring Harbor, NY, ²Embrapa, Parque Estação Biológica, Brasília, Brazil, ³USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY, ⁴iPlant Collaborative, Bio5 Institute, University of Arizona, Tucson, AZ

Common analysis steps in all RNA-seq experiments include quality control, read alignment, assigning reads to genes or transcripts, and estimating gene or transcript abundance. Choice of analysis tool and experimental strategy are critical in ensuring high quality data for analysis and visualization of the results. In this study we compare and assess new tools available for spliced alignment(STAR & HISAT), assembly(StringTie) and quantification(Balloon, eXpress, Sailfish, Kallisto and Salmon) for RNA seq data using cyberinfrastructure(CI) services developed by iPlant collaborative project.

The CI unites high-end computing, large scale data storage and networking with user-interfaces that simplify creation, integration and sharing of software tools and workflows. Here we compare the established Tuxedo protocol for RNA seq analysis against the new tools and evaluate the performance of the workflows on the CI. Most of these new tools were benchmarked on human RNA seq datasets, we want to test the efficacy and robustness of these tools on the CI using RNA seq data from plants. The alignment and assembly tools were integrated in Discovery Environment platform that executes jobs in a Docker container while the quantification tools were integrated as a virtual image in Atmosphere cloud computing service within the CI. The results were visualized on a Biodalliance genome web browser for assessing quality of the annotation.

The iPlant Collaborative is funded by a grant from the National Science Foundation (#DBI-0735191)

DECONVOLVING GENE EXPRESSION PROFILES FOR TUMOR POPULATIONS WITH PRIOR FREQUENCY INFORMATION

Christopher Cremer, Quaid Morris

University of Toronto, Department of Computer Science, Toronto, Canada

High-throughput, genome-wide sequencing can profile the abundance of thousands of RNAs in a tumor, providing a comprehensive snapshot of tumor gene expression. Previous studies have used mRNA expression profiles to identify cancer sub-types, and to predict patient prognosis and treatment response. Predictive potential is stifled, however, by tumor heterogeneity within single samples. A tumor sample is not a uniform collection of genetically identical cells, but a mixture of different cell populations. This heterogeneity arises from contamination by non-cancerous tissue, as well as evolution that occurs throughout a tumor's lifetime. These distinct cellular populations exhibit heterogeneity through not only genomic mutations, but also differences in gene expression. Beyond informing our understanding of cancer evolution, these gene expression differences constrain the discovery of biomarkers necessary for personalized cancer therapeutics.

At present, subclonal populations in a tumor sample can be modeled by analyzing DNA sequencing data for simple somatic mutations (SSMs) and copy number variations (CNVs) unique to each population. To build on these efforts, we have developed a computational model using RNA expression data that addresses how contamination by normal tissue and tumor subpopulation heterogeneity affect gene expression. Our model represents tumor expression profiles as linear combinations of their constituent cancerous populations and non-cancerous contamination. Expression profiles for each population in a tumor are learned by iterative optimization. Prior knowledge of population frequencies within a tumor is drawn from DNA-sequencing-based methods, serving as a regularization factor that penalizes the model when it deviates too far from estimated population frequencies in generating tumor expression profiles. Our method extends existing methods for deconvolving tumor expression data, by incorporating frequency estimates as prior information, thereby, pushing the deconvolution towards meaningful latent factors.

In evaluating our model using simulated data, we found it accurately recovers correct tumor expression profiles. We are now investigating how variance in estimates for numbers of subpopulations and their associated frequencies affects model performance. Application of our model to clinical data will grant insight into how mutations affect gene expression in cancer, improving understanding of tumor evolution and informing development of improved treatments.

COMPARISON OF CHROMOSOME STRUCTURE ACROSS CONDITIONS USING A THREE DIMENSIONAL CHROMOSOME BROWSER

Steven W Criscione¹, Yue Zhang¹, Benjamin Siranosian¹, Alan Hwang¹, Marco De Cecco¹, John M Sedivy¹, Nicola Neretti^{1,2}

¹Brown University, Molecular Biology, Cell Biology, and Biochemistry, Providence, RI, ²Brown University, Center for Computational Molecular Biology, Providence, RI

High-throughput conformation capture, including Hi-C and 5C methods, uncovered an unprecedented view of the 3D conformation of chromosomes. These methods revealed that the chromosome is structured into compartments, referred to as A and B, which correlate with markers of active euchromatin and repressive heterochromatin. At a second-level the genome is organized into topological associated domains containing loci that strongly interact in cis. Despite these insights there are challenges to visualizing this information in a 3D model of the chromosome. First, multiple structural solutions are compatible with Hi-C data which can complicate comparisons across experimental conditions. Second, other information provided by 3D DNA fluorescence in situ hybridization (FISH) experiments imposes relevant spatial constraints on chromosome structure. Third, visualization of 3D chromosome structure together with other genomic data, such as ChIP-seq data, remains unresolved. Here we introduce 3DC-Browser, a 3D Chromosome browser package for visualizing chromosome structure together with improvements to a consensus model of chromosome structure. Our consensus model can leverage spatial constraints from 3D DNA FISH and chromosome painting experiments to yield a more realistic chromosome model. The consensus method includes an option to solve the structure for an experimental condition from an initialized state provided by a control structure. We demonstrate that this strategy can remove artifacts that complicate comparing structures from two conditions such as arbitrary reflections. Finally, we provide a standalone web interface for visualizing the resultant 3D model and for integrating the 3D model with data from other genomic experiments.

WRANGLING DATA INTO TRACK HUB VISUALIZATIONS WITH *HUBWARD*

Ryan K Dale

National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Laboratory of Cellular and Developmental Biology, Bethesda, MD

Comparing data among publications should be an integral part of doing science. In theory, processed data from published genomics studies would be provided in standard file formats, ready for visualization in a genome browser. In practice, supplementary data in papers and in repositories like GEO and ArrayExpress have different formats depending on the preferences of the submitter. This lack of standardization hinders data comparison.

hubward is a tool to manage the download of original data files (e.g., "supp_table_S1.xls"), convert them into standardized file formats (bigBed, bigWig, VCF, BAM) that can be uploaded to the UCSC genome browser, create a UCSC track hub (a set of tracks along with metadata, documentation, data provenance, and visualization configuration), and upload the track hub to be shared with collaborators -- all in a reproducible, easily-updated manner.

Due to the proliferation of bespoke data formats, some degree of custom work cannot be avoided when visualizing a data set. *hubward's* guiding principle is to restrict the scope of such custom work while retaining the flexibility to support arbitrary data formats. The design is purposefully lightweight, driven by config files processed by a command-line interface that supports the management and maintenance of many created track hubs. Convenience functions for commonly-used tasks are provided; for example, one function applies a colormap to the score column of a BED file to create a colored bigBed file. Finally, *hubward* generates and uploads UCSC track hub files based on the config files, allowing the data to be viewed alongside any other data in the UCSC genome browser.

Visualizing existing data remains critical for informing experimental design and strategy and for interpreting new results. *hubward* aims to reduce the overhead of this visualization while providing mechanisms for reproducibility, maintainability, data provenance, and data reusability. Source code and documentation are available at <https://github.com/daler/hubward>.

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES OF DEVELOPING MOUSE TOOTH WITHIN THEIR GENOMIC LOCATIONS

Rishi Das Roy, Outi Hallikas, Elodie Renvoise, Jukka Jernvall

Institute of Biotechnology, University of Helsinki, Helsinki, Finland

The complex tissue types in multi-cellular organisms develop from stem cells through a process of cell differentiation. In this process, tissue specific expressions of genes are regulated at various developmental stages with the combination of transcription factors and regulatory elements. It is also hypothesized that changes in gene expression could be a primary factor causing variation in morphology. Our previous studies have revealed the diverse shape and size of mammalian tooth as a good model to study the evolution and developmental aspects.

Therefore we measured the gene expression levels of mouse tooth and jaw (as control) at embryonic day 13 and 14 through Affymatrix Mouse Exon 1.0 array. The gene expression were measured using robust multi-array average algorithm (R package: aroma.affymetrix) and the genes were annotated using latest custom chip definition files (CDF) from Microarray lab (<http://brainarray.mbni.med.umich.edu/Brainarray/default.asp>). Thereafter the differentially expressed genes (DE_genes) were identified using linear models (R package: limma) and compared with “Gene expression in tooth database” (<http://bite-it.helsinki.fi/>). However many significant tooth developmental genes were not successfully identified in DE_genes list due to lower than 1 log fold change in their expression level. Nevertheless a careful observation of these genes revealed that they are indeed differentially expressed in comparison to their neighbouring genes from the same chromosomal location. We hypothesize that localized differential expression of single gene in its chromosomal surroundings could be a signature of its critical role in the development of studied organ.

Available methods evaluate each gene independently to identify their differential expression. To overcome this limitation, we are now developing computational methods to identify crucial genes from their localized expression patterns and to validate them. Overall our study and method will help us to identify more potential tooth developmental genes.

COMPARING ALGORITHMS TO GENOTYPE SHORT TANDEM REPEATS IN NEXT-GENERATION SEQUENCING DATA

Harriet Dashnow^{1,2}, Alicia Oshlack¹

¹Murdoch Childrens Research Institute, Bioinformatics, Parkville, Victoria, Australia, ²The University of Melbourne, Biosciences, Parkville, Victoria, Australia

Short tandem repeats (STRs) are short (2-6bp) DNA sequences repeated in tandem, which make up approximately 3% of the human genome. These loci are prone to frequent mutations and high polymorphism. Dozens of neurological and developmental disorders have been attributed to STR expansions. STRs have also been implicated in a range of functions such as DNA replication and repair, chromatin organisation and regulation of gene expression.

Traditionally, STR variation has been measured using capillary gel electrophoresis. This process is time-consuming and expensive, and so has tended to limit STR analysis to a handful of loci.

Next-generation sequencing has the potential to address these problems. However, determining STR lengths using next-generation sequencing data is difficult. For example, many callers are limited by sequencing read lengths and polymerase slippage during PCR amplification introduces stutter noise.

Recently, a small number of software tools have been developed genotype STRs in next-generation sequencing data. We have performed a general comparison of the tools published to date, identifying their application domains, assumptions and limitations.

We have assessed the performance of some of the most popular STR genotyping tools on human next-generation sequencing data. When comparing STR callers we have observed drastic differences in which STR loci are identified as variant. Surprisingly, even for variant loci reported in common between tools, there is markedly low concordance between the specific genotype calls.

Finally, we draw together our findings to comment on the considerations when choosing and running an STR genotyping tool, with an emphasis on applications to human disease.

THE IMPACT OF RNA DEGRADATION ON THE ABILITY TO DETECT FUSIONS USING TRUSEQ RNA LIBRARY PREPARATION

Jaime I Davila¹, Wang Xiaoke², Numrah Fadra¹, Nair Asha¹, McDonald Amber², Crusan Barbara², Kandelaria Rumilla², Jen Jin², Klee Eric^{1,2}, Kipp Benjamin², Halling Kevin²

¹Mayo Clinic, Health Sciences Research, Rochester, MN, ²Mayo Clinic, Laboratory Medicine and Pathology, Rochester, MN

The identification of gene fusions using next generation sequencing has become widely available and is increasingly used in a clinical setting. The Illumina TruSeq RNA library is a popular method for library preparation that uses a poly-A pulldown strategy which is optimal for high quality RNA samples. When sequencing partially degraded samples, the poly-A pulldown chemistry can cause coverage profiles with less read coverage at the 5' end of the gene resulting in the reduced ability to detect fusions the farther the fusion breakpoints are from the 3' end of the gene. In particular, BCR/ABL fusions whose breakpoints are farther than 5 kb from the poly-A tail are difficult to detect in degraded specimens whereas FGFR3/TACC3 fusions whose breakpoint is less than 1kb from the poly-A tail are easy to detect even in degraded samples.

We sought to more precisely characterize the effect that RNA degradation has on the ability to detect fusions using RNA-seq with the TruSeq RNA library preparation method. Universal Reference RNA (UHR) samples were fragmented at different levels and their RNA integrity number (RIN) was measured. We then performed RNA-seq analysis on these samples and plotted the gene coverage level as a function of the distance from the poly-A tail and an exponential decrease in coverage was found which became more pronounced as the RIN decreased. We developed a tool that plots the sensitivity of detection of the fusion as a function of the distance of the fusion breakpoint to the 3'UTR. This tool provides a quantitative way to measure the effect of degradation and its impact on the ability to detect fusions and aids in assessing the quality of fragmented RNA samples for RNA-seq.

ADDITIONAL VARIANTS AMONG THE MH-GRID COHORT DISCOVERED AFTER ALIGNMENT TO AN ANCESTRY SPECIFIC REFERENCE GENOME

Adam R Davis¹, Ryan A Neff¹, Shurjo Sen¹, Cihan Oguz¹, Rakale Quarells², Gary H Gibbons¹

¹National Human Genome Institute, Cardiovascular Disease Section, GMCID, Bethesda, MD, ²Morehouse School of Medicine, Cardiovascular Research Institute, Atlanta, GA

Abstract:

Whole genome sequencing studies across certain populations, such as those with African ancestry, are often underpowered due to a larger divergence between the common reference genome and the true genetic sequence of the population. However, a common reference genome is not designed to account for this divergence in population-specific studies. Strong signals from common (MAF>50%) single nucleotide polymorphisms (SNPs), insertion-deletions (indels), and structural variants (SVs) can make alignment and variant calling difficult by masking nearby variants with weaker genetic signals. We present the results generated from alignment to an African descent population-specific reference genome by applying variants present in a majority of individuals with African descent from all phases of the 1000 Genomes Project and the International HapMap Consortium. We compared alignment of MH-GRID samples between the population-specific and the hg19 reference. We identified 680,285 single nucleotide polymorphisms at MAF>50% in the MH-GRID population. We demonstrate that utilization of a population-specific reference improves variant call quality, coverage level, and imputation accuracy. We discovered additional SNPs by alignment to the population specific reference in union across all samples, including exonic variants that are clinically relevant to resistant hypertension in African Americans.

(*) Corresponding Author

The project described was supported by Dr. Gary H. Gibbons, PI, Division of Intramural Research of the NIH, National Human Genome Research Institute. We are grateful to the original MH-GRID study at Morehouse School of Medicine funded by the National Institute on Minority Health and Health Disparities (NIMHD) Grant Number 8 U54 MD007588-04/Formerly Grant Number U54 RR026137 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).

ESAT – A NEW TOOL FOR ANALYZING END-SEQUENCING RNA-SEQ DATA

Alan Derr¹, Alexey A Sergushichev^{2,5}, Sabah Kadri³, Sebastian Kadener⁴, Maxim N Artyomov², Manuel Garber¹

¹University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA, ²Washington University in St. Louis, Pathology & Immunology Department, St. Louis, MO, ³Broad Institute, Cambridge, MA, ⁴Hebrew University of Jerusalem, Dept. of Biological Chemistry, Jerusalem, Israel, ⁵ITMO University, Computer Technologies Department, Saint Petersburg, Russia

In whole-genome RNA-Seq analysis, we generally seek to measure the relative levels at which each gene or isoform is present in a set of samples. To date, most quantification tools (e.g. RSEM, Cufflinks) assume that reads are distributed across the full length of each transcript and use that assumption to provide a normalized measurement of expression such as Transcripts per Million (TPMs).

However, many newer RNA-Seq protocols, especially those adapted to low-input or single-cell RNA-Seq, target the end of transcript mRNAs, resulting in libraries with strong 3'- or 5'-biases. Standard quantification methods are not appropriate for these “end-sequenced” libraries.

We present ESAT (End Sequencing Analysis Toolkit), a tool specifically designed for analyzing end-sequencing-based libraries. ESAT focuses on maximizing the information provided by: (i) searching for possible starts/ends that may not be annotated to ensure that all genes are optimally scored (ii) identifying and quantifying all potential transcript start/end sites within each gene locus (iii) supporting protocols that incorporate unique molecular identifiers (UMIs) and using UMIs to detect and remove duplicate reads from PCR duplication.

ESAT is designed to process many (thousands) of samples together with the goal of comparing differences not only in gene expression but also leveraging the specific nature of end-sequence libraries to enable analysis of differential usage of polyadenylation or transcription start sites. ESAT is flexible, with command-line parameters to allow the user to select experiment-specific features such as the width of the transcript scan window, the amount of overlap between successive windows, the method for handling multimapped reads, and a p-value threshold to determine which windows are considered candidate polyadenylation or transcription start sites. We will describe the features of ESAT in detail, and present several applications using both published and unpublished end-sequencing data sets of both single cell and large cell populations.

RECONSTRUCTING THE EVOLUTIONARY HISTORY OF TUMOURS

Amit G Deshwar¹, Quaid Morris^{1,2}

¹University of Toronto, Electrical and Computer Engineering, Toronto, Canada, ²University of Toronto, Donnelly Center for Cellular and Biomolecular Research, Toronto, Canada

Tumours often contain multiple, genetically diverse subclonal populations. Identifying these sub-populations and their evolutionary relationships can aid in the understanding of cancer development and progression, as well as, ultimately impacting treatment decisions.

Subclonal reconstruction algorithms attempt to infer the prevalence and genotype of multiple, genetically-related subclonal populations using the variant allele frequency (VAF) of simple somatic mutations (SSMs) and large copy number variants. The low read-depth typical of these experiments (~50x) makes this problem extremely challenging.

A number of machine learning algorithms have been used to address this difficult problem; however, methods that only consider somatic mutations individually are unable to recover tumour phylogenies in many cases, and are prone to other estimation errors. We will describe a new approach that overcomes these limitations; thereby greatly expanding the collection of tumours whose evolutionary histories can be reconstructed.

It is often possible to determine the mutation status of >1 mutation in a single cell when a single read covers >1 SSM. We have incorporated these reads into our new PhyloSpan method which adds considerable power to the phylogenetic reconstruction of the tumor subclonal populations. These multi-SSM reads aren't available for many pairs of mutations, but a small number of them can provide substantial phylogenetic certainty. For example, if two SSMs are found in the same evolutionary branch, then we expect to see reads containing both mutations.

An often-overlooked factor in accurate subclonal reconstruction is the "winner's curse". Only SSMs with enough reads supporting the mutation can be called, leaving many more mutations that are present in the sample but not called. Not taking this truncation into effect results in significantly skewed estimates in the number of subclonal mutations and the cellular prevalence of subclonal populations. We have recently extended our existing algorithms to model the data as truncated distributions.

A current controversy in the field of subclonal reconstruction is how important neutral evolution is in explaining the subclonal mutations observed in tumours. Our new method includes neutral and clonal evolution as competing explanations for VAFs. This permits us to ascertain which tumours and which cancer types have subclones under selection and which do not.

Together these three innovations permit much more precise reconstruction of the evolution history of tumours, even for those sequenced with relatively low read depth.

PRECISION-STAR: UNBIASED ALLELE AWARE MAPPING OF RNA-SEQ READS TO PERSONAL GENOMES.

Alexander Dobin, Thomas R Gingeras

Cold Spring Harbor Laboratory, Functional Genomics, Cold Spring Harbor, NY

Being a part of the drive towards precision medicine, personal genomics holds the promise for identifying genetic predispositions for common diseases, diagnosis and management of hereditary disorders, characterization and individual treatment of cancer, and genotype guided drug research and administration. Precision-STAR, an extension of the RNA-seq aligner STAR, utilizes genotype information to produce more accurate allele aware mapping of RNA-seq data. Unlike many other approaches, Precision-STAR incorporates the variants in the form of Variant Call Format (VCF) files directly at the mapping step, and produces the allele-specific alignment in one passage. Each short variant (SNP or short indel) is appended to the reference genome together with its surrounding sequence, while keeping track of its genomic locus and allele. In the seed search phase, the reads are mapped to both the reference genome sequence and the short variant sequences, which eliminates reference mapping bias. In the seed stitching phase, the seeds from the reference genome and short variant sequences are used together to construct the final highest scoring alignment. Because of the favorable logarithmic scaling of the STAR's mapping speed with the reference length, the addition of the short variance sequences does not significantly reduce the mapping speed.

The end product of the Precision-STAR mapping run includes haplotype information (maternal, paternal, undetermined) for each read. It also produces allele specific quantifications at the level of individual short variants, as well as for entire genes and transcripts. Precision-STAR is also capable of haplotype aware mapping and allele-specific expression (ASE) computation even without existing genotype information using the following algorithm. In the 1st pass, the reads are mapped to the reference genome, and the short variants are detected. In the 2nd pass, the short variants are added to the reference genome as described above, and the reads are re-mapped to the on-the-fly generated personal genome. This approach is accurate for medium and highly expressed genes.

Using simulated and real data we show that our approach eliminates most of the reference mapping biases and achieves high fidelity in detecting ASE. To demonstrate superior efficiency of Precision-STAR, we reprocessed the ENCODE and GEUVADIS data to produce bias free personal genome alignments and ASE.

APOLLO: A PLATFORM FOR COLLABORATIVE GENOME CURATION AND ANALYSIS

Nathan A Dunn¹, Monica C Munoz-Torres¹, Colin Diesh², Deepak Unni², Eric Yao³, Ian Holmes³, Christine Elsik², Suzanna E Lewis¹

¹Lawrence Berkeley National Labs, Genomics Division, Berkeley, CA, ²University of Missouri, Animal Sciences, Columbia, MO, ³University of California Berkeley, Bioengineering, Berkeley, CA

Modern genome sequencing projects face several challenges not experienced by their pioneering predecessors. They address highly diverse scientific aims: from comparing the genomes of closely related species, to better understanding population dynamics, to exploring the molecular mechanisms behind processes such as adaptive radiation. In many cases, the genomes themselves have lower coverage, contain more frequent assembly errors, and lack comparable sequences from closely related species. Lastly, the communities for these burgeoning new genome projects are often relatively small and geographically dispersed. In addition to providing a genome browser that allows visualization of automated gene models, Apollo serves the diverse needs of these new genome projects by enabling collaborative, real-time curation (akin to Google Docs) of genomic elements in terms of structure and functional information. Researchers from nearly one hundred institutions worldwide, some also involved in model organism databases and specific research consortia (such as i5K and VectorBase), are currently using Apollo for distributed curation efforts in over sixty genome projects across the tree of life: from plants to arthropods, to fungi, to species of fish and other vertebrates including human, cattle, and dog.

Our most recent release, Apollo 2 (<http://genomearchitect.org>), improves usability, scalability, and customization to better support newer genome projects. The web-based client, which uses the JBrowse genome viewer, has added a sidebar that provides a detailed view of annotations, sequences, and organisms as well as a new reporting structure, and websocket support to improve real-time communication. The server was rewritten using the Grails framework (Spring / Hibernate / Groovy) to more robustly scale a single server over multiple organisms while better supporting additional curators. The architecture is simplified to use a single database (e.g., PostgreSQL, MySQL, H2) to store organization and annotation data. Furthermore, the entire secure REST API used to build Apollo is exposed. This allows, for example, genomic features to be injected into Apollo from an automated curation process or organization-specific metadata to be extracted directly from Apollo using a SQL query or REST. Apollo 2 has an improved standard set of tools but also increases the ability to customize Apollo and integrate it into a modern genome sequencing project's curation pipeline.

ASSEMBLY-FREE COMPARATIVE GENOMICS OF *TRICHOMONAS VAGINALIS* AND THREE OTHER TRICHOMONADS

Daniel Ence^{1,3}, Claudia P Marquez², Mark Yandell^{1,3}, Ellen J Pritham¹

¹University of Utah, Department of Human Genetics, Salt Lake City, UT,

²University of Texas at Arlington, Department of Biology, Arlington, TX,

³Eccles Institute of Human Genetics, Utah Center for Genetic Discovery, Salt Lake City, UT

Genome projects involving phylogenetically distant organisms can be hampered by a scarcity of annotations from closely related organisms. Transposon rich genomes can further complicate assembly and annotation. The genome of the G3 strain of *Trichomonas vaginalis* suffered from both of these issues. The original genome annotation in 2007 reported ~60,000 protein-coding genes, although this number is almost certainly inflated by the inclusion of gene fragments and open-reading frames from transposable elements. A key form of evidence for gene annotation are comparisons with closely related organisms, which are limited for *T. vaginalis* due to its distant relationship with animals, fungi and plants. Originally comparisons were made with *Giardia lamblia*, *Dictyostelium discoideum*, and *Entamoeba histolytica*. Here we repurpose the superfast metagenomics tool, Taxonomer to explore the inter- and intraspecies conservation of protein-coding genes among ten strains of *T. vaginalis* and 3 other trichomonad species. We identified genomic reads with homology at the nucleotide and protein level to genes annotated in *T. vaginalis* G3 reference assembly. The gene annotations in *T. vaginalis* are based on our recent re-annotation of the G3 reference assembly, which makes use of a thorough annotation of transposable elements in the G3 reference assembly. After classifying reads to the *T. vaginalis* genes, we performed *de novo* assembly of the reads classified to each gene and constructed multiple sequence alignments of each gene. This approach recovers conserved genes better than standard alignment tools, which placed only a small numbers of reads from the non-*T. vaginalis* samples on the G3 reference assembly. In closing, we note that the approach described here bypasses two major requirements of comparative genomics (i.e., the need for genome assemblies and annotations for every species of interest). These results are the first ever genome wide study of conservation of protein-coding genes among species in this phylogenetically distant and understudied group of organisms. In the future, we plan on developing additional comparative genomics applications for Taxonomer, including the exploration of TE landscapes and identification of conserved noncoding regions between species.

BIOLOGICALLY BASED DISEASE CLASSIFICATION FOR CHILDHOOD ARTHRITIS.

Simon W Eng^{1,2}, Rae S Yeung^{1,2}, Quaid Morris³

¹University of Toronto, Department of Immunology, Toronto, Canada, ²The Hospital for Sick Children, Division of Rheumatology & Department of Cell Biology, Toronto, Canada, ³University of Toronto, Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada

Juvenile idiopathic arthritis (JIA) encompasses a heterogeneous set of autoimmune disorders characterized by joint inflammation. We sought to determine whether unique clinical and biological patterns underlie the complex and varied phenotypes. In support of this, several Canada-wide studies have collected vast, heterogeneous clinical and biological data from patients with JIA. Here, we describe some initial dimensionality reduction analyses to decode the heterogeneity of JIA and identify clinical and biological composite descriptors (or factors) that distinguish patients from each other.

Using principal component analysis (PCA), we established an initial set of meaningful clinical and biological factors with treatment-naïve patients from the REACCH OUT (Research in Arthritis in Canadian Children, Emphasizing Outcomes) cohort. In clinical and cytokine data, PCA identified 4 factors representing pro-inflammatory cytokine levels, disease activity, age of diagnosis, time to diagnosis, and modes of immune activation linked with autoimmunity—clearly suggesting links between aberrant immune activation and disease presentation. In joint involvement data, PCA identified 3 factors describing contrasts between groups of co-activated joints, specifically total number of joints inflamed, fingers versus toes, and large weight-bearing joints versus small joints.

Having established the initial factors, we found interesting biological ones by extending our analysis to the BBOP (Biologically Based Outcome Predictors of JIA) cohort, which contains a superset of patient features in the REACCH OUT cohort. First, we projected data from treatment-naïve patients from BBOP onto the initial factors. Then, with gene expression data from peripheral blood, PCA identified 6 factors, which we characterized using gene set enrichment analysis, enrichment maps, and word clouds. One factor linked adaptive immunity with increased cellular growth, and another described a pronounced innate immune response, pointing to the underappreciated role of the innate immune system in JIA. Our preliminary work demonstrates that we can identify clinically and biologically meaningful patterns that distinguish patients from each other and provide novel insight into JIA. Our pattern recognition results provide a foundation for cluster analysis to identify patient subgroups from these indicators, providing the basis for a biologically based disease classification for JIA.

TUNESIM : TUNABLE VARIANT SET SIMULATOR FOR NGS READS

Bertrand Escalière^{1,4}, Sonia Van Dooren^{1,2}, Raphael Helaers³, Gianluca Bontempi^{1,4}, Guillaume Smits^{1,5,6}

¹IB2, Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium, ²Vrije Universiteit Brussel, UZ Brussel, Center for Medical Genetics, Reproduction and Genetics, Reproduction Genetics and Regenerative Medicine, Brussels, Belgium, ³Université Catholique de Louvain, De Duve Institute, Brussels, Belgium, ⁴Université Libre de Bruxelles, Machine Learning Group, Brussels, Belgium, ⁵Université Libre de Bruxelles, Hôpital Erasme, Center for Medical Genetics, Brussels, Belgium, ⁶Hôpital Universitaire des Enfants Reine Fabiola, Genetics, Brussels, Belgium

NGS analysis software and pipelines optimization is crucial in order to improve discovery of disease causing variants. Developing a simulator able to insert known human variants at a realistic minor frequency and artificial variants in a tunable controlled way would allow to overcome three optimization limits: determination of exact software/pipeline calling sensitivity and accuracy; optimization on the appropriate population; and the capacity to dynamically stress one variable at the time on a realistic variant background.

We implemented TuneSIM, a wrapper around NGS dwgsim reads simulator (<http://sourceforge.net/projects/dnaa/>), to simulate reads with realistic mutations. TuneSim can be applied to any NGS dataset type. In order to generate data as realistic as possible, TuneSim inserts variants following an haplotype block structure. We computed TruSeq exome +/- 200 bp blocks using Plink and vcf files from IKG Western European individuals, and fed the results into TuneSim. TuneSim generates a random number following an uniform distribution and compares it to the haplotype frequencies of the block, such as to retain one of the possible haplotypes. TuneSim performs this operation twice to create two independent DNA strands. It then inserts individual dnbSNP variants not part of the blocks using their MAF, their exonic status and random numbers generated following an inverse transform sampling procedure. If the generated probability or its square is lower than the GMAF, the variant is inserted as an heterozygous or an homozygous variant, respectively. In addition, TuneSim offers the possibility to insert random variants with desired Ti/Tv ratio and changes dwgsim base quality scores to realistic scores. Each step is optional. Without dwgsim, TuneSim could simply be used to generate realistic variant datasets and/or be adapted to existing simulators.

We already used TuneSim for exome pipelines comparisons, and in-house development of NGS tools. Results will be presented. TuneSim will be hosted on a web interface allowing users to download the code and generated datasets, compare accuracy results and suggest new simulations.

BENCHMARKING ULTRAFAST WORKFLOWS FOR HUMAN GENOME VARIANT CALLING

Gloria Redon^{1,2}, Volodymyr Kindratenko¹, Victor Jongeneel^{1,2}, Liudmila S Mainzer^{1,2}, Christopher Fields^{1,2}

¹U. of Illinois, Natl Ctr for Supercomputing Applications, Urbana, IL, ²U. of Illinois, Inst for Genomic Biology, Urbana, IL

As the pace of implementing personalized medicine concepts increases, high-throughput variant calling on hundreds of individual genomes per day is a reality that will likely be faced by sequencing facilities across the country in the near future. While the scientific best practices for human variant calling workflows have been well defined, they also pose serious computational challenges at this high scale. Therefore, efforts in both academia and the private sector have focused on developing alternative “monolithic” workflows that may eliminate these bottlenecks and substantially reduce the computational cost per individual genome. We have tested and benchmarked three of these workflows: Genalice (<http://www.genalice.com/>), iSAAC (http://www.illumina.com/documents/products/whitepapers/whitepaper_isaac_workflow.pdf) and BALSAMIC (<https://peerj.com/articles/421/>) can deliver variants on a deeply sequenced whole human genome in minutes to hours. The short turn-around time allows users to easily re-run the workflow if anything goes wrong with the analysis, and removes the need for check-pointing, usually implemented by breaking up the workflow into modular components that are quick to rerun. The monolithic (as opposed to modular) nature of the ultra-fast workflows eliminates the intermediary files, lightening the load on the file system and reducing the data footprint on disk. A single computational module in place of a workflow also obviates the need for complex workflow management, with its required error-checking, quality controls and job dependencies. While these ultrafast variant callers require advanced servers with large amounts of RAM and high core count, they do eliminate the need for supercomputers to process large amounts of data, and significantly reduce the processing cost per genome. Their downside is that they allow less experimentation with the procedures and parameters for variant calls, and are thus only applicable to the production analysis of genomes (primarily human) for which standard operating procedures have already been well established. We show that there are very significant differences in the performance of the three workflows we tested, in terms of sensitivity, specificity, speed, and robustness in handling noisy datasets. While none give results of a quality that is comparable to a well-tuned “best practice” workflow, some come close enough to be of clinical utility.

GENOMIC REGION AND SAMPLE SELECTION STRATEGY FOR VARIANT DISCOVERY AND ASSOCIATION ANALYSIS

Steven M Foltz^{1,2}, Kai Ye^{1,2}, Li Ding^{1,2}

¹Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, ²Washington University School of Medicine, Department of Medicine, St. Louis, MO

Computational discovery of genomic variants relies on having appropriate coverage and modeling systematic error in the data. Depth of coverage varies throughout the genome and exome. Variant discovery software may produce false positive results in genomic regions with low coverage or high sequence error rates. We propose an algorithm that automatically performs sample and genomic region selection by evaluating coverage depth, coverage variability, and sequence error patterns. One sequence error pattern we have observed is a fixed gap in the same position of many reads, regardless of location in the genome. By effectively minimizing batch effects, our approach helps control the quality of data produced by various sequencing instruments and facilities for effective variant discovery and meta-association analysis.

TRANSCRIPTOME STATES IDENTIFIED BY PROBABILISTIC MODELING OF CLIP-SEQ DATASETS IN YEAST

Mallory A Freeberg, James Taylor

Johns Hopkins University, Department of Biology, Baltimore, MD

UV-induced crosslinking and immunopurification of an RNA-binding protein (RBP) followed by deep sequencing of its bound RNAs (CLIP-seq and derivative protocols) is an increasingly popular method for identifying *in vivo* transcriptome-wide sites of RBP interactions at nucleotide resolution. Consequently, a large collection of published deep-sequencing datasets is available representing RNA targets of hundreds of RBPs for many organisms. Initial analyses of RNA target sites for individual RBPs have revealed important mechanistic insights into RBP-mediated post-transcriptional gene regulation. However, integration of all available RBP:RNA interaction site information what is missing is a transcript-centric, is lacking.

Inspired by work from the ENCODE group to identify chromatin states (*e.g.* promoter, enhancer) from histone modification marks, DNA polymerase occupancy, and transcription factor binding sites, we have identified biologically distinct transcript states across the entire yeast protein-coding transcriptome based on empirical evidence of physical interactions between yeast transcripts and over 50 RBPs, including ribosomes. We developed a probabilistic model to learn transcript states by modeling read count data produced by CLIP-seq experiments from multiple published studies. After learning and characterizing these states, we incorporated transcript characteristics – primary RNA sequence motifs, predicted RNA secondary structures, sequence conservation, and expression levels – and uncovered that specific post-transcriptional regulatory features correlate with different combinations of transcript states. Importantly, our model allows yeast transcript states to be easily updated as more CLIP-seq datasets become available, and transcript states can be learned from CLIP-seq data for any organism with only minor modifications.

ORTHOGONAL SEQUENCING OF THE HUMAN EXOME

Alexander Frieden², Niru Chennagiri², Eric White¹, Daniel Lieber², John Thompson¹

¹Claritas Genomics, Research and Development, Cambridge, MA, ²Claritas Genomics, Software and Informatics, Cambridge, MA

Whole exome sequencing studies have revolutionized the diagnosis of genetic disorders by identifying and providing clinically valuable information about pathogenic variants throughout the exome. However, the sheer volume of variants that are identified in each individual patient makes confirmation and interpretation of the data very challenging. ACMG practice guidelines for clinical laboratory standards for next-generation sequencing (NGS) recommend that “all disease-focused and/or diagnostic testing include confirmation of the final result using a companion technology”, a daunting task given the number of variants involved. Many of these variants are rare, often unique to the patient and with uncertain significance with respect to disease. Typically, Sanger sequencing is used for confirmation. However, Sanger sequencing is expensive and time consuming.

To minimize the time necessary for confirmation and interpretation, Claritas Genomics has developed an approach using complementary sequencing technologies to analyze variants from whole exome sequences that have been restricted to phenotypically-defined, clinically relevant regions of interest (ROI). Each individual’s exome is sequenced using two separate DNA target capture and sequencing methodologies. DNA is captured using a hybridization-based approach and sequenced using Illumina NGS. In parallel, DNA is captured using an amplification-based approach and sequenced using Ion Torrent NGS. Because independent and complementary technologies are used, the need for Sanger sequencing for confirmation is reduced. For 19119 genes, across a full exome, approximately 90% of the variants are detected on both platforms and thus are orthogonally confirmed. When this analysis is carried out on DNA with an extensive truth set (HapMap sample NA12878), 100% of the orthogonally confirmed variants are true positives, supporting the use of independent capture and sequencing approaches. The use of orthogonal sequencing platforms provides immediate confirmation of most variants and also yields better sensitivity than either platform alone can generate, providing more sensitive and timely results for patients.

K-MER SPECTRA FILTERS TO ASSEMBLE HIGH QUALITY, CONTIGUOUS, COLLAPSED MOSAICS FROM NON-MODEL HETEROZYGOUS GENOMES

Gonzalo A Garcia Accinelli¹, Darren Heavens¹, Jens Maintz², Diane Saunders^{1,2}, Matt Clark¹, Federica Di Palma¹, Bernardo J Clavijo¹

¹The Genome Analysis Centre, Computational Genomics, Norwich, United Kingdom, ²The Sainsbury Laboratory, Crop Genetics, Norwich, United Kingdom

As whole genome sequencing becomes cheaper and easily accessible, we are moving from a single-reference, model organism, genome scenario towards a myriad of interesting but extremely challenging individual genomes. Highly heterozygous genomes, previously avoided by sample choice or specific breeding, need to be sequenced and assembled but current technology still has struggles with them. Typical pipelines produce assemblies with poor contiguity, duplicated sequence and most worryingly even loss of homozygous content. In this work, we propose a post-contigging filter approach that takes advantage of a new generation of assemblers that conserve some representation for all the possible variants of every locus.

We use the output from DISCOVAR de novo, and apply expectation maximization heuristics based on the k-mer spectra of the raw reads to create a mosaic genome representation collapsing the haplotypes into one choice per locus instead of trying to reconstruct all combinations. The filter keeps all the homozygous content and discards roughly half of the heterozygous content in the final set of contigs. This collapsing simplifies the scaffolding problem downstream reducing the number of alternative paths for the use of long range information, which in turn produces a higher quality and more contiguous assembly with all the content for complete collapsed mosaic. We tested our approach on ash tree (*Fraxinus excelsior*) and yellow rust (*Puccinia striiformis*), two highly heterozygous organisms. We are able to reduce the duplicated content by 60% and discard as much as 40% of the heterozygous content with negligible (<1%) loss of homozygous content. The assembly contiguity is also much greater with an N50 increased by 25% compared to the standard pipelines. Both the increased contiguity and the absence of spurious duplication enable more powerful downstream analysis. Our current implementation of the filter consists of a set of scripts that can be included in any existing pipeline between the contigging and the scaffolding steps.

DIDA: A FIRST DIGENIC DISEASES DATABASE

Andrea Gazzo^{1,2}, Dorien Daneels*^{1,3}, Elisa Cilia*^{1,2}, Maryse Bonduelle³, Marc Abramowicz^{1,5}, Sonia Van Dooren^{1,3}, Guillaume Smits^{1,4,5}, Tom Lenaerts^{1,2,6}

¹Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels, Belgium, ²Machine Learning Group, ULB, Brussels, Belgium, ³Center for Medical Genetics, VUB, Brussels, Belgium, ⁴HUDEURF, VUB, Brussels, Belgium, ⁵Center for Medical Genetics, Hopital Erasme, Brussels, Belgium, ⁶AI lab, VUB, Brussels, Belgium

DIDA (Digenic diseases DATabase) is a novel database that provides for the first time a manually curated collection of genes and associated variants involved in digenic diseases (DD). The database is accessible at <http://dida.ibsquare.be>.

Digenic inheritance is the simplest form of oligogenic inheritance for genetically complex diseases and has been defined as follows: "inheritance is digenic when the variant genotypes at two loci explain the phenotypes of some patients and their unaffected relatives more clearly than the genotypes at one locus alone" (Schäffer, 2013).

It has been shown that human disease may in certain cases be better described by more complex inheritance mechanisms, overcoming the limitation of the monogenic model and introducing the oligogenic one. All currently available resources that might be mined for information on DD do not provide a detailed variant pair information as is done in DIDA. As such, DIDA and its future developments, have the potential to promote new avenues in medical, biological and bioinformatics research. Making such information publically available is the essential first step towards intensifying research into the combinatorial nature of many diseases, even when those diseases were classically considered to be monogenic.

We have collected and curated 216 digenic variant pairs (comprising of 134 distinct genes and 367 distinct variants), which are linked to 42 DD. Manual curation is crucial to ensure the quality of the database. All digenic pairs are directly associated with the related publications, and many additional annotations and cross-links to other related databases are provided.

DIDA is a relational database structured in 4 tables representing the main domain concepts (entities): genes, single variants, digenic pairs and diseases. Each table contains information representing properties or attributes of the corresponding entity. The web interface provides browsing and search functionalities, as well as documentation, help pages and general database statistics.

DIDA is an innovative research resource, which provides information and insight into how digenic pairs jointly lead to disease. It provides an important tool for clinical and molecular geneticists to find the most comprehensive information on the digenic nature of their diseases of interest and is a first step towards oligogenic diseases elucidation.

PIPES - A TOOL FOR CLASSIFYING LONG RNA READS.

Sam Kovaka¹, Alex Dobin², Thomas R Gingeras²

¹Clark University, Worcester, MA, ²Cold Spring Harbor Laboratory, Functional Genomics, Cold Spring Harbor, NY

Alternative splicing is a common phenomenon in eukaryotes, with more than 75% of multi-exon human genes having multiple isoforms. Despite the significant effort to annotate the human genome, there are likely many unknown isoforms to be discovered. While standard short-read RNA sequencing is a commonly used method for identifying individual splice sites, determining the full length transcript structure from short-read data requires complex analyses which are not always accurate. High throughput long-read sequencing technologies from companies like Pacific Biosciences and Oxford Nanopore have the ability to sequence whole RNA molecules, providing entire intron structures and making it possible to accurately identify new isoforms.

It has been reported that a significant (>25%) proportion of the long reads do not match annotated transcripts, which makes it crucially important to classify the novel isoforms differences with respect to the annotated transcripts. To accomplish this task we developed PIPES, Pairwise Isoform Polymorphism Encoding and Scoring tool. In our method, the long RNA reads are mapped to the genome using a spliced sequence aligner such as BLAT, GMAP or STAR. Next, we find the annotated transcripts that overlap each of the alignments. The differences between the reads and annotated transcripts are represented with an encoding scheme termed PIPE, similar to the SAM CIGAR. PIPE strings contain characters to represent matches and mismatches in splice donors/acceptors and transcript terminals. Next, the differences between the read alignments and the annotated transcripts are scored (similarly to the local alignment schemes) according to their types: skipped exon, new exon, retained intron, split exon, shifted donor/acceptors, terminal new exon, and terminal skipped exon. The highest scoring PIPE designates the closest matching annotation for a long read. Finally, the reads are classified into broader categories such as annotated, intergenic, fragment, complete, incomplete etc. Importantly, PIPES can collapse novel reads into putative novel transcripts, which are then classified with the above scheme. We applied our pipeline to the PacBio long read datasets consisting of 1,383,979 circular consensus reads in three human tissues, 59% of which were mapped to the human genome in a circular consensus fashion. 331,898 (40%) of all mapped reads were classified as full length or fragments of annotated transcripts, while 269,231 (33%) of the reads were found to be novel isoforms. Collapsing the reads resulted in 170,075 distinctive transcripts. Interestingly, 86% of these transcripts were found to be novel. Even among the transcripts supported by 5 or more reads, 48% are found to be novel isoforms. Potential evolutionary, mechanistic and regulatory implications derived from the use of PIPES will be discussed.

RED: AN INTELLIGENT, RAPID, ACCURATE TOOL FOR DETECTING REPEATS DE-NOVO ON THE GENOMIC SCALE

Hani Z Girgis

University of Tulsa, Tandy School of Computer Science, Tulsa, OK

Background: With rapid advancements in technology, the sequences of thousands of species' genomes are becoming available. Within the sequences are repeats that comprise significant portions of genomes. Successful annotations thus require accurate discovery of repeats. As species-specific elements, repeats in newly sequenced genomes are likely to be unknown. Therefore, annotating newly sequenced genomes requires tools to discover repeats de-novo. However, the currently available de-novo tools have limitations concerning the size of the input sequence, ease of use, sensitivities to major types of repeats, consistency of performance, speed, and false positive rate.

Results: To address these limitations, I designed and developed Red, applying Machine Learning. Red is the first repeat-detection tool capable of labeling its training data and training itself automatically on an entire genome. Red is easy to install and use. It is sensitive to both transposons and simple repeats; in contrast, available tools such as RepeatScout and ReCon are sensitive to transposons, and WindowMasker to simple repeats. Red performed consistently well on seven genomes; the other tools performed well only on some genomes. Red is much faster than RepeatScout and ReCon and has a much lower false positive rate than WindowMasker. On human genes with five or more copies, Red was more specific than RepeatScout by a wide margin. When tested on genomes of unusual nucleotide compositions, Red located repeats with high sensitivities and maintained moderate false positive rates. Red outperformed the related tools on a bacterial genome. Red identified 46,405 novel repetitive segments in the human genome. Finally, Red is capable of processing assembled and unassembled genomes.

Conclusions: Red's innovative methodology and its excellent performance on seven different genomes represent a valuable advancement in the field of repeats discovery.

REFSEQ ANNOTATION OF FUNCTIONAL ELEMENTS ON THE HUMAN AND MOUSE REFERENCE GENOMES

Tamara Goldfarb*, Catherine M Farrell*, Sanjida H Rangwala, Terence D Murphy, Donna R Maglott, Kim D Pruitt

National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD

The National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) (www.ncbi.nlm.nih.gov/refseq/) database provides a non-redundant dataset of genomic, transcript, and protein sequence standards including annotated reference genomes. Annotation from RefSeq sequences are represented in NCBI's Gene database (www.ncbi.nlm.nih.gov/gene/), providing a centralized resource to view information such as genomic context, phenotypic information, variation, and interaction data. The scope of the RefSeq and Gene projects thus far has emphasized representation of genic regions, but such regions represent only a small fraction of the genome. Given that there are many conserved non-genic sequences with important biological functions, it is important to represent these by reference standard sequences. Annotation of these elements on the reference genome is not easily accessible to public users who lack specific research knowledge and expertise. The RefSeq group is thus initiating a project to curate the location and attributes of functional genomic elements on the human and mouse reference genomes. This new effort will be integrated with our current annotation and associated with RefSeq and Gene records. In our initial pilot project, curation will focus on known strong recombination hotspots and known regulatory elements, such as enhancers, locus control regions, and boundary elements. Only those regions with functional and experimental validation in the literature will be included. Records in the NCBI Gene database for these regulatory elements will include short summaries, comprehensive literature references, and annotation of sub-features. These records will constitute a unique resource for information about these non-genic features and provide great value to the biomedical community.

This work was supported by the Intramural Research Program of the National Library of Medicine, NIH.

*These authors contributed equally

STRUCTURAL ALTERATION OF TRANSCRIPT ISOFORMS IN HUMAN CANCERS

Leonard D Goldstein^{1,2}, Eric Stawiski^{1,2}, Thong Nguyen¹, James Lee³, David Stokoe³, Robert Gentleman⁴, Somasekar Seshagiri¹

¹Genentech, Molecular Biology, South San Francisco, CA, ²Genentech, Bioinformatics & Computational Biology, South San Francisco, CA, ³Genentech, Discovery Oncology, South San Francisco, CA, ⁴23andMe, Computational Biology, Mountain View, CA

Structural alteration of transcript isoforms plays an important role in human cancers. Cancer-specific splice variants can result in inactivation of tumor suppressors and activation of oncogenes. Examples of known activating variants include EGFRvIII and MET exon 14 skipping. Moreover, mutations in regulatory splicing factors can result in altered transcript isoforms. High-throughput sequencing of RNA (RNA-seq) has been used successfully for identification of transcripts from fusion genes, but has been underutilized for the study of intra-genic transcript variants.

We have uniformly processed RNA-seq data from The Cancer Genome Atlas (TCGA) project (6,748 cancer and 636 normal samples) and our own in-house cancer sequencing projects. RNA-seq data for 2,958 normal samples from the Genotype-Tissue Expression project (GTEx) were processed with the same tool chain and serve as normal controls. Reads were aligned with the splice-aware aligner GSNAP, which can accurately align reads across both known and novel splice junctions. We use our recently developed R/Bioconductor software package SGSeq for prediction and quantification of splice variants.

Our analysis detected known oncogenic splice variants (EGFRvIII in ~11% of glioblastoma and ~2% of lower grade glioma and MET exon 14 skipping in ~3% of lung adenocarcinoma) as well as several previously uncharacterized recurrent candidate activating variants in known oncogenes and loss-of-function variants in known tumor suppressors. Moreover, we identified transcript alterations associated with recurrent splicing factor mutations. We are investigating potential mechanisms and functional consequences of these alterations.

METAUVEN: A TAVERNA-BASED PIPELINE FOR THE ANALYSIS OF SHOTGUN METAGENOMIC DATA

Giorgio Gonnella, Laura Glau, Stefan Kurtz

University of Hamburg, Center for Bioinformatics ZBH, Hamburg, Germany

Due to the decreasing costs of DNA sequencing, metagenomics became a flourishing area of research in the recent years. The amount of metagenomics datasets, their sizes and the diversity of tasks to analyse them are constantly growing. Several solutions for a semi-automatic analysis of metagenomics data are available. However, most of them can hardly be installed locally (e.g. MG-Rast, [MPD+08]; WebCarma, [GJT+09]) or are difficult to integrate with other tools (e.g. MEGAN, [HMW+11]).

Here we present MeTavGen, a flexible analysis pipeline for shotgun metagenomics datasets based on Taverna [W+13]. MeTavGen can be run locally on a single computer or an SGE computer cluster. The pipeline glues together several external software tools using custom Ruby and Bash scripts. MeTavGen can easily be extended by the user to include more analysis steps.

As input the pipeline expects either preprocessed reads or metagenomics contigs. Genes in the input sequences are annotated and their protein products are aligned to the NCBI NR database. The results are analyzed with MEGAN to determine the distribution of taxa in the metagenome. Furthermore, the genes are functionally annotated and a statistical analysis of the functional profiles of the most common taxonomic groups is performed by STAMP [PTHB14], using a similar workflow to the one adopted in [PGKL14].

The results of the analysis are presented in an automatically generated dynamic HTML report, which allows to easily access a large number of tables, metabolic pathway maps, bar and hierarchical plots, generated using MEGAN, STAMP, R and KronaTools [OBP11].

References

- [GJT+09] Gerlach *et al.*, BMC bioinformatics, 10:430, 2009.
- [HMW+11] Huson *et al.*, Genome Research, 21:1552-1560, 2011.
- [MPD+08] Meyer *et al.*, BMC bioinformatics, 9:386, 2008.
- [OBP11] Ondov *et al.*, BMC bioinformatics, 12(1):385, 2011.
- [PGKL14] Perner *et al.*, Applied and environmental microbiology, 2014.
- [PTHB14] Parks *et al.*, Bioinformatics, 30(21):3123-3124, 2014.
- [W+13] Wolstencroft *et al.*, Nucleic acids research, 41:W557-61, 2013.

CHARACTERIZATION OF DNA DAMAGE RESPONSE AND R-LOOPS IN EWING SARCOMA

Aparna Gorthi¹, Yidong Chen², Alexander Bishop¹

¹Univ of Texas Health Science Center at San Antonio, Cellular & Structural Biology, San Antonio, TX, ²Univ of Texas Health Science Center at San Antonio, Epidemiology & Statistics, San Antonio, TX

Ewing sarcoma is a rare, yet highly aggressive family of bone and soft tissue tumors that afflicts children and young adults. It is the prototypical example of mesenchymal tumors driven by a fusion oncogene involving the EWSR1 gene, most frequently EWS-FLI1. Ewing sarcoma is exquisitely sensitive to genotoxic agents, and more recently PARP1 inhibitors, predominantly when detected at an early stage. However, the molecular basis for this sensitivity is relatively underexplored. The effectiveness of these drugs suggests that Ewing sarcoma tumors are unable to respond to and repair exogenously induced DNA double strand breaks.

We took a systems biology approach to delineate the pathways that are dysregulated in Ewing sarcoma, both basally as a function of the oncogenic program, as well as in response to genotoxic stress. Evaluation of the dynamics of transcriptional response to DNA damage (time course RNA-seq) revealed elevated replication stress and an inability to control transcription stress. We discovered that Ewing sarcoma cell lines have widespread accumulation of R-loops, which are DNA structures consisting of an RNA:DNA hybrid and a displaced complementary DNA loop. To further characterize the R-loops, we performed DRIP (DNA:RNA Immunoprecipitation)-Seq with Ewing sarcoma and controls with and without exposure to DNA damage. We also established a bioinformatics pipeline for analyzing DRIP-seq data as well as comparing it with orthogonal data sets such as RNA-seq and ChIP-seq. We identified a subset of genes that are altered transcriptionally upon damage in control cell lines but fail to do so in Ewing sarcoma. Further comparison with DRIP-seq indicates that these genes are associated with R-loops in Ewing sarcoma, providing a first clue into their extreme chemosensitivity. Investigation into the unique molecular profile of Ewing tumors also revealed an entire network of repair and cellular homeostasis pathways that are upregulated to compensate for the widespread ‘genomic stress’ incurred as a result of R-loops and impaired recombination in Ewing sarcoma. Further, we provided initial proof-of-principle for targeting some of these pathways as alternative therapy. We have also discovered that these pathways (FEN1, RNASEH2 and Fanconi Anemia) are similarly upregulated in neural crest cancer stem cells and to a smaller extent, in bone marrow mesenchymal stem cells. Up-regulation of these pathways may provide a tolerant environment for the establishment of EWS-FLI1 and lends strength in favor of their role as cells-of-origin.

IMPROVEMENT OF THE ASSEMBLY OF HETEROZYGOUS GENOMES OF NON-MODEL ORGANISMS

Anaïs Gouin¹, Anthony Bretaudeau^{2,3}, Emmanuelle d'Alençon⁴, Claire Lemaitre¹, Fabrice Legeai^{1,2}

¹Inria/IRISA équipe GenScale, Campus de Beaulieu, 35042 Rennes Cedex, France, ²INRA IGEPP, Domaine de la Motte, 35653 Le Rheu Cedex, France, ³Inria/IRISA, GenOuest Core Facility, Campus de Beaulieu, 35042 Rennes Cedex, France, ⁴INRA DGIMI, Université de Montpellier 1, 34000 Montpellier, France

Whereas the number of non-model organisms being sequenced has drastically increased, the extraction of biological information from such data is hampered by the low quality of the draft assemblies. In particular, the combination of a high level of heterozygosity and short reads sequencing leads to fragmented assembly and the overestimation of the gene content and of the genome size. Recently, new assemblers have been developed to better handle heterozygous data. But, the complete re-assembly of a genome involves automatic and manual re-annotations tasks that are very cost-effective. Thus, we present here a novel method to detect and correct false duplications due to heterozygosity (two alleles instead of one consensus sequence) in diploid draft assemblies. In addition, the method is able to relocate and merge supernumerary gene annotations.

The method is based on a whole genome self-alignment (*Lastz + AxtChain*) allowing the detection of highly similar regions. These can have two origins: either allelic regions or duplicated regions. To distinguish between them, three criteria are used: 1/ their location inside scaffolds: contrary to duplications, unmerged haplotypes come from the same locus and must share the same genomic contexts, 2/ their cumulative read depth (close to the expected one) and 3/ their level of redundancy in the whole assembly. Next, Detected pairs of allelic regions needs to be merged into one unique sequence in the assembly: either by the complete deletion of the redundant scaffolds or by the construction of meta-scaffolds (scaffolds joined together) keeping only the allele present in the longest scaffold of the pair. Genes located on the merged alleles need to be correctly re-annotated. This is performed using *Exonerate* and *Augustus*. The former allows to identify the location of the deleted genes onto the remaining allele. The latter is used to predict new genes or consensus ones.

We applied this method to an heterozygous wild type insect genome assembly. This leads to a drastic reduction of the genome assembly size (coherent with the expected size estimated by flow cytometry) and to the increase of the N50. Most of the new meta-scaffolds were confirmed by several additional resources : mate pairs, BAC ends sequence mapping and synteny analysis. Moreover, about 80% of gene predictions located in removed fragments have been either relocated or merged with their complementary allele.

CONCORDANCE AND CONTAMINATION CHECKER FOR WGS AND WES MATCHED SAMPLE STUDIES

Ewa A Grabowska, Phaedra Agius, Kanika Arora, Nora C Toussaint, Dayna M Oswald, Vladimir Vacić

New York Genome Center, Computational Biology Group, New York, NY

With the decreasing cost of sequencing, the number of samples that can be analyzed at the same time is rising rapidly. With more samples being processed within a lab, the probability of sample mix-ups and cross-contamination increases. Studies that utilize matched samples (tumor-normal, multiple treatment conditions, experiment replicates, etc.) are an integral part of all genomic research projects. Sample mismatch or contamination may result in an incorrect analysis outcome, decreased sensitivity, and/or specificity.

Here we present a fast and robust method to perform concordance verification, as well as contamination checking for matched human samples in whole-genome and whole-exome sequencing experiments. Based on a list of over 7000 preselected exonic markers, characterized by being highly variable in the 1000 Genomes dataset, we can reliably detect sample swaps and estimate contamination levels. Evaluation of sample contamination is based on a likelihood model. In contrast to `verifyBamIid`, our algorithm uses a small set of highly sensitive markers and allows for only two possible alleles for each marker. This approach greatly decreases running time, while still accurately estimating contamination levels.

We also offer a solution to the problem of sample contamination versus sample rearrangement, which is a common issue in many cancer studies. For this, we detect all homozygous markers in a normal sample, and use them to check for contamination in the matched tumor samples, utilizing the fact that homozygous markers are not sensitive to copy number changes and genomic rearrangements.

Our tool has been successfully applied to multiple synthetic datasets, as well as real production projects at the New York Genome Center, where it provides a robust measure to ensure the data quality control. It can be easily extended to non-human organisms.

ADAGE: A METHOD FOR THE UNSUPERVISED INTEGRATION OF GENE EXPRESSION EXPERIMENTS APPLIED TO *PSEUDOMONAS AERUGINOSA*

Jie Tan¹, John H Hammond², Deborah A Hogan², Casey S Greene^{1,3}

¹Geisel School of Medicine at Dartmouth, Genetics, Hanover, NH, ²Geisel School of Medicine at Dartmouth, Microbiology and Immunology, Hanover, NH, ³University of Pennsylvania Perelman School of Medicine, Systems Pharmacology and Translational Therapeutics, Philadelphia, PA

The growth in genome-scale data for different species in publicly available databases provides the opportunity for hypothesis generation by the application of new analytical methods for biological interpretation of these data. Here, we present an unsupervised machine-learning approach, ADAGE (Analysis using Denoising Autoencoders of Gene Expression) and apply it to the interpretation of all of the publicly available gene expression data for *Pseudomonas aeruginosa*, an important opportunistic bacterial pathogen. Without any prior knowledge of any genome structure or gene function, the *P. aeruginosa* ADAGE model found that co-operonic genes often participated in similar processes and accurately predicted which genes had similar functions. Using newly generated data, we were able to use the model to identify gene expression differences between strains, to identify the cellular response to low oxygen, to predict the involvement of biological processes in previously published data despite low level expression differences is directly involved genes, and to identify processes that are most highly responsive to different environmental perturbations. Our results include the ADAGE model interpretation of all publicly available *P. aeruginosa* GeneChip experiments an open source code for ADAGE that can be readily applied to other species and systems.

THE GENOME AND TRANSCRIPTOME OF THE REGENERATION-COMPETENT FLATWORM, MACROSTOMUM LIGNANO

Kaja A Wasik¹, James Gurtowski¹, Xin Zhou⁴, Olivia M Ramos¹, M. Joaquina Delas¹, Giorgia Battistoni¹, Osama El Demerdash¹, Ilaria Faciadori³, Dita B Vizoso³, Peter Ladurner³, Lukas Scharer³, W. Richard McCombie¹, Gregory J Hannon², Michael C Schatz¹

¹Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY, ²CRUK Cambridge Institute, Li Ka Shing Centre, Cambridge, United Kingdom, ³University of Basel, Evolutionary Biology, Basel, Switzerland, ⁴Stony Brook University, Molecular and Cellular Biology Graduate Program, Stony Brook, NY

The free-living flatworm, *Macrostomum lignano*, much like its better known planarian relative, *Schmidtea mediterranea*, has an impressive regenerative capacity. Following injury, this species has the ability to regenerate almost an entirely new organism. This is attributable to the presence of an abundant somatic stem cell population, the neoblasts. These cells are also essential for the ongoing maintenance of most tissues, as their loss leads to irreversible degeneration of the animal. This set of unique properties makes a subset of flatworms attractive organisms for studying the evolution of pathways involved in tissue self-renewal, cell fate specification, and regeneration. The use of these organisms as models, however, is hampered by the lack of a well-assembled and annotated genome sequences, fundamental to modern genetic and molecular studies. Here we report the genomic sequence of *Macrostomum lignano* and an accompanying characterization of its transcriptome. The genome structure of *M. lignano* is remarkably complex, with ~75% of its sequence being comprised of simple repeats and transposon sequences. This has made high quality assembly from Illumina reads alone impossible (N50=222 bp). We therefore generated 130X coverage by long sequencing reads from the PacBio platform to create a substantially improved assembly with an N50 of 64 Kbp. We complemented the reference genome with an assembled and annotated transcriptome, and used both of these datasets in combination to probe gene expression patterns during regeneration, examining pathways important to stem cell function. As a whole, our data will provide a crucial resource for the community for the study not only of invertebrate evolution and phylogeny but also of regeneration and somatic pluripotency.

SYSTEMATIC ANALYSIS OF ALTERNATIVE POLYADENYLATION DURING NEUROGENESIS OF MURINE EMBRYONIC STEM CELLS

Kevin Ha^{1,3}, Benjamin Blencowe^{1,3}, Quaid Morris^{1,2,3}

¹University of Toronto, Molecular Genetics, Toronto, Canada, ²University of Toronto, Computer Science, Toronto, Canada, ³University of Toronto, Donnelly Centre for Cellular and Biomedical Research, Toronto, Canada

Alternative polyadenylation (APA) is a post-transcriptional process by which multiple RNA transcript isoforms with distinct 3' ends derived from the same gene can be produced. It has been estimated that 30-50% of mammalian genes contain more than one cleavage and polyadenylation site^{1,2}. Changes in APA patterns have been observed during development of mouse embryonic stem cells (ESCs)³⁻⁵. Recent studies have also shown that RNA binding proteins (RBPs) can regulate APA through binding of cis-acting motifs in the 3'UTR, and its mis-regulation is associated with disease⁶. However, the regulatory mechanisms involved in APA are not completely understood. We developed a framework for studying APA from RNA sequencing (RNA-Seq) data by examining the expression of 3'UTR isoforms to estimate differential usage of polyadenylation sites. This method was applied to a published RNA-Seq time series capturing various stages of neurogenesis⁷. Principal component analysis revealed a subset of transcripts that expressed short 3'UTRs during the early stages of differentiation and subsequently lengthen, confirming previous findings³. Gene ontology analysis identified several functions related to neurogenesis and stem cell development. Compared to constitutively expressed 3'UTRs, many of these lengthening 3'UTRs do not contain the canonical polyadenylation signal, AAUAAA, which is most strongly associated with polyadenylation site selection. Hence, we hypothesize that other mechanisms regulating 3'UTR lengthening during neurogenesis may be at play, such as the role of RBPs. Our analysis demonstrates the feasibility of using RNA-Seq to study APA and its underlying regulatory mechanisms.

1. Tian, B. et al. *Nucleic Acids Res* 33, 201–12 (2005).
2. Shepard, P. J. et al. *RNA* 17, 761–72 (2011).
3. Ji, Z. et al. *Proc. Natl. Acad. Sci* 106, 7028–33 (2009).
4. Lackford, B. et al. *EMBO J* 33, 878–89 (2014).
5. Ji, Z. & Tian, B. *PLoS One* 4, e8419 (2009).
6. Batra, R. et al. *Mol. Cell* 56, 311–322 (2014).
7. Hubbard, K. S. et al. *F1000Research* 2, 35 (2013).

CLAMMS: A SCALABLE PIPELINE FOR CNV CALLING AND QUALITY CONTROL, APPLIED TO OVER 40,000 EXOMES

Lukas Habegger¹, Jonathan S Packer¹, Evan K Maxwell¹, Colm O'Dushlaine¹, Alexander Lopez¹, Samantha N Fetterolf², Joseph B Leader², David J Carey², David H Ledbetter³, Frederick E Dewey¹, Rostislav Chernomorsky¹, Aris Baras¹, John D Overton¹, Jeffrey G Reid¹

¹Regeneron Pharmaceuticals, Regeneron Genetics Center, Tarrytown, NY,

²Geisinger Health System, Weis Center for Research, Danville, PA,

³Geisinger Health System, Genomic Medicine Institute, Danville, PA

Many algorithms have been developed for calling copy number variants (CNVs) from whole-exome sequencing (WES) read depths, but existing tools are difficult to automate, lack scalability, and are only suitable for calling rare variants. To address these limitations we developed CLAMMS, a new algorithm that is easy to parallelize and calls CNVs of any allele frequency. Common variants are handled naturally by exon-level copy number estimates from mixture models, which are then smoothed by a Hidden Markov Model. CLAMMS' scalability is rooted in its use of sequencing quality control metrics, computed for individual samples, as an effective low-dimensional approximation to a sample's overall coverage profile. Each sample's read depths are normalized against those of the 100 samples nearest to it in this low-dimensional metric space. A k-d tree allows a sample's nearest neighbors to be found in $O(\log n)$ time.

We applied CLAMMS to a population-scale sequencing project of over 40,000 exomes. We initially validated selected common and rare variants with TaqMan qPCR. We then trained quality control procedures using transmission rates of called CNVs in ~3,000 parent-child duos identified using PRIMUS, an algorithm for building pedigrees from population sequence data via identity-by-descent. These procedures, which supplement coverage-based quality metrics with auxiliary information such as allele balance and zygosity of SNPs within a called CNV region, achieve high transmission rates and sensitivity. For example, we detect an average of 0.58 small (< 10 kb), rare ($AF < 1\%$) variants per sample at a transmission rate of $>47\%$. Even rare single-exon calls achieve a transmission rate of 41.5%. We also demonstrate a significantly improved false positive / false negative trade-off compared to CNVs called from microarray data using PennCNV.

An initial survey of CNVs in our sequenced population identified over 12,000 distinct copy number variant loci, the vast majority of which (~92%) are extremely rare ($AF < 0.01\%$). Conversely, $>70\%$ of variants identified in the average individual are common ($AF > 5\%$). Our results highlight the benefits of detecting CNVs across the allele frequency spectrum as well as the need for larger sample sizes to improve the statistical power of association tests.

PROOVREAD-3.0: PACBIO HYBRID ERROR CORRECTION FOR DI-/POLYPLOID GENOMES, METAGENOMES AND TRANSCRIPTOMES

Thomas Hack^{1,2}, Frank Förster², Matthias G Fischer¹

¹Max-Planck-Institute for Medical Research, Biomolecular Mechanisms, Heidelberg, Germany, ²University of Wuerzburg, Department of Bioinformatics, Würzburg, Germany

Long sequencing reads (PacBio SMRT, Nanopore MinION) are becoming a more and more valuable resource in the quest towards high quality assemblies of genomes, metagenomes and transcriptomes. Their main disadvantage, error rates of more than 15%, can be overcome by different computational correction methods and programs. For typical data sets, accuracies of >99% are obtained in many cases. Most of the current algorithms, however, are based on alignment and consensus approaches, either using complementary short read data or long read self-alignments. While these methods produce reliable results for the majority of sites, they often fail to properly distinguish signals from different, yet highly similar regions: in particular heterozygous sites in di-/polyploid samples or high similarity regions in metagenomes and transcriptomes are affected. With the latest release of our hybrid correction software *proovread-3.0*, we are addressing this challenge. We have extended the basic Illumina-to-PacBio alignment and consensus approach to include identification and stabilization of variable sites, coverage independent inference of the most likely state from the raw read and a final refinement step performing read-backed phasing. As a result *proovread-3.0* is capable of consistently discriminating paralogous transcripts, homologous regions in metagenomes and haplotypes in genomic samples.

UTILIZATION OF VERY LARGE READ DEPTH SEQUENCING DATA FOR THE DETECTION OF VARIATION IN A FATHER-MOTHER-SON TRIO

Nancy F Hansen¹, James C Mullikin¹, Genome in a Bottle Consortium²

¹National Human Genome Research Institute, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, ²National Institute of Standards and Technology, Genome Scale Measurements Group, Gaithersburg, MD

The Genome in a Bottle Consortium (GIABC) is a public-private-academic consortium led by the National Institutes of Standards and Technology (NIST) with the goal of facilitating the translation of whole genome sequencing data to clinical practice. As part of this project, the GIABC has gathered data from a variety of sequencing platforms to fully characterize several reference materials, all of which will then be made available, along with high confidence variant calls, to investigators for use in benchmarking their own sequencing methods.

One of these whole genome sequence datasets consists of very high read depth (300x) Illumina paired-end (150x150 bp reads) data from each of three Ashkenazi individuals (a father-mother-son trio) from the Personal Genomes Project. We use this large dataset to examine the value of very high read depth in different types of variant calling, and specifically to assess the value of high read depth for detecting de novo variation. Additionally, we assess the accuracy and replicability of variant calling as a function of read depth across multiple technical replicates. In this way, we can answer questions with regard to optimal read depth for the detection of different types of variants.

TEPEAKS: A TOOL FOR INCLUDING REPETITIVE SEQUENCES IN CHIP-SEQ AND CLIP-SEQ ANALYSES

Ying Jin, Yuan Hao, David Molik, Oliver Tam, Molly Hammell

Cold Spring Harbor Laboratory, Bioinformatics Shared Resources, Cold Spring Harbor, NY

Next generation sequencing technologies are widely used in characterizing genome-wide binding patterns of transcription factors, histone modifiers, and other chromatin associated proteins through the use of ChIP-seq assays. Many of these chromatin-associated factors bind to repetitive regions of the genome where unique alignment of short reads can be problematic. Most existing methods either discard non-uniquely mapped reads or randomly choose one from the multiple alignments. Both strategies reduce the accuracy in determining enrichment in repetitive regions such as regions with transposable element insertions. Since TEs cover between 20-80% of most eukaryotic genomes, this can result in lowered accuracy of binding site analyses for a large fraction of the genome. We have developed TEpeaks, a method for identifying ChIP-seq peaks genome-wide, that includes the repetitive fraction of the genome as well as uniquely mappable sites. TEpeaks carefully distributes multiply mapped reads using the uniquely mapped reads as a guide and optimizes the assignment by the expectation maximization (EM) algorithm. Moreover, TEpeaks provides multiple normalization options and also includes a module for differential binding analysis. TEpeaks is part of the larger TEtoolkit package that also includes an RNA-seq analysis suite, Tetranscripts (Jin et al., 2015). Together, this package allows the analysis and integration of a wide variety of sequencing datasets such as ChIPSeq, CLIPSeq, RNA-seq, and GRO-seq.

CNVTHRESHER: COMBINING MULTIPLE LINES OF EVIDENCE TO CONSTRUCT HIGH-QUALITY CNV CALL SETS

Jason Harris, Gábor Bartha, Stephen Chervitz, Deanna M Church, Richard Chen

Personalis, Inc., R&D, Menlo Park, CA

Robust detection of Structural and Copy-Number Variants (SVs and CNVs) using short-read sequencing data is a notoriously difficult problem. Many of the available SV/CNV callers require significant tradeoff between sensitivity and specificity, and are often inherently sensitive to limited categories of variants. Exacerbating the problem is the fact that experimental factors and systematic alignment errors can mimic the signals on which CNV detectors rely. As a result, there are very few reliable and complete “Gold” CNV call sets, from which we could measure the performance of various detection methods.

We present **CNVThresher**, a tool that annotates CNV detections in WGS data with metrics designed to reflect the level of evidence present in the aligned reads in support of each putative call. The tool examines the pileup of aligned reads in the vicinity of the call, and looks for several features that are expected for real CNVs: (1) Morphological features of the read-depth profile: the read depth should transition sharply at the breakpoint to a level reflecting the CNV’s copy number, and this level should be sustained throughout the feature’s width. (2) Read mapping anomalies: including read pairs with anomalous insert size, and aligned soft-clip edges at the breakpoints. (3) Allele distribution features: we expect a lack of biallelic positions inside heterozygous deletions, and we expect to see positions with unbalanced allele ratios inside duplications. (4) Quality control metrics: to flag calls that may be more likely false positives, we also measure the fraction of reads that had a mapping quality of zero (both inside the feature, and in the flanking regions), and the density of non-Reference bases in the flanking regions. Together, the CNVThresher annotations provide a robust method for evaluating the level of evidence present in the aligned reads for any set of CNV calls.

THE LANDSCAPE OF MICROSATELLITE INSTABILITY IN CANCER EXOMES

Ronald J Hause¹, Emily H Turner², Mallory Beightol², Colin C Pritchard², Jay Shendure¹, Stephen J Salipante²

¹University of Washington, Genome Sciences, Seattle, WA, ²University of Washington, Laboratory Medicine, Seattle, WA

Microsatellites, 2-5 base pair repetitive sequences present throughout the human genome, can abnormally shorten or lengthen because of defects in the DNA mismatch repair (MMR) system, resulting in a “microsatellite instability” (MSI) phenotype. MSI is a key prognostic and diagnostic tumor phenotype that has been well studied by conventional methods. However, both the genomic landscape of MSI events and differences in MSI among cancer types remain poorly illuminated. We here present a comprehensive analysis of the landscape of MSI in cancer exomes. We catalogued MSI events at over 500,000 incidentally sequenced microsatellite loci across 4,224 cancer exomes spanning 18 different cancer types from The Cancer Genome Atlas. We constructed a global classifier for MSI that achieved 93.75% sensitivity and 98.5% specificity compared to gold-standard MSI calls based on the revised Bethesda guidelines. We observed that MSI-low (MSI-L) samples did not display significant differences from MS-stable (MSS) samples in the number of MSI events and support discontinuation of the use of MSI-L as a distinct classification. Comparative examination of MSI revealed both cancer-specific and core loci associated with global MSI, such as a frameshift mutation in an 8-bp polyadenine tract in exon 10 of the tumor suppressor ACVR2A observed in 56.6% of MSI-high (MSI-H) cancers. Lastly, we investigated the relationships between MSI and mutations in MMR genes, gene regulatory features, and clinical covariates. Our results provide a comprehensive view of MSI in cancer exomes, highlighting both conserved and cancer-specific MSI properties and identifying candidate genes underlying predisposition to global MSI. Future work will attempt to functionally validate these candidates as causally influencing the cancer phenotype.

ON THE RELATIVITY OF TIME AND SPACE OF TUMORS: CLINICAL SEGMENTATION IS THE KEY TO ERADICATE BREAST CANCER.

Fritz E Hauser

Aerzte-Gesundheitszentrum Laegern AG, www.ghzl.ch, Project
SEGMENTA, Ehrendingen AG, Switzerland

Despite all the changes that evolution brings about segmentation of a body has always been kept by nature, like an axiom! What teaches this in developing new medical diagnostics and therapies (1)?

Clinical segmentation (2) originating from dermatologic pattern formation and viscerocutaneous reflexes (3) is found in most diseases. The relativity in time and space of i.e. pancreatic cancer has been shown (4).

Examples in tumors, i.e. breast cancers, and metastases (5) will be shown and how diseases keep their space in the body on this road map of segmentation over time. This opens new research developments.

Project SEGMENTA, an upcoming research firm of Medical Scientific Knowledge Enterprise, is essential in order to achieve reliable returns of investments in sciences, biotechnology, and pharmaceutical industries basing on THE NEW MEDICAL ENTITIES (6), almost as "easy" as in *Drosophila*:

Patient + Pattern = Discovery + Product (6).

(1) Hauser FE: Clinical Segmentation & New Medical Entities: Innovation in developing future Blockbusters. Lecture at NOVARTIS Headquarters, Basle, CH, October 24, 2008.

(2) Hauser FE: Clinical segmentation as a major promotor of gene therapy applications. Nature Biotechnology symposium "Gene Therapy: Delivering the Medicines of the 21st Century." Washington, DC, USA, Nov 7-9, 1999.

(3) Hauser W: Lokalisationsprobleme bei Hautkrankheiten. In: Korting GW (1980) Dermatologie in Klinik und Praxis. Bd. I. Allgemeine Dermatologie. Stuttgart, New York 1980, 8.60-8.93.

(4) Iacobuzio-Donahue CA et al: Distant metastasis occurs late during the genetic evolution of pancreatic cancer. doi:10.1038/nature09515 & The patterns and dynamics of genomic instability in metastatic pancreatic cancer. doi:10.1038/nature09460.

(5) Hauser FE: Clinical Segmentation and Signal Transduction in Cancer. Miami 2005 nature biotechnology winter symposia, Miami, Florida, USA, February 5-9, 2005. [<http://www.med.miami.edu/mnbws/documents/hauserSR05.pdf>].

(6) Hauser FE: Pharmaunternehmen sollten kreativer forschen. (Pharma Companies should do research more creatively) On NEW MEDICAL ENTITIES in answer to the New Chemical Entities of the study Re-Inventing Drug Discovery: The Quest for Innovation and Productivity. [Anderson Consulting, NY 1997]. Letter to the Editor: Research of pharma firms should become more efficient. Science under the pressure of economic opportunity. Neue Zürcher Zeitung, CH, 1998/02/3, 46.

(7) Hauser FE: Project SEGMENTA: Patient+Pattern=Discovery+Product. BIO 2001 International Convention & Exhibition, San Diego, CA, USA, June 24-27, 2001.

© 2015, Fritz E. Hauser, segmenta (at) gmx.net, Ehrendingen, Switzerland.

USING THE LANDSCAPE OF GENETIC VARIATION IN PROTEIN DOMAINS TO IMPROVE FUNCTIONAL CONSEQUENCE PREDICTIONS

Jim Havrilla¹, Aaron Quinlan^{1,2}

¹University of Utah, Human Genetics, Salt Lake City, UT, ²University of Utah, Biomedical Informatics, Salt Lake City, UT

Numerous methods (e.g., Polyphen and SIFT) exist to predict the impact of a genetic variant on protein function. The recent RVIS study from Petrovski et al. provides a gene-wide score by regressing the number of common missense variants vs. the total number of variants. The CADD approach from Kircher et al., utilizes annotations and the ancestral genome to determine phenotypic impact by contrasting variants that survived natural selection with simulated mutations using a Support Vector Machine. Lastly, the fitCons score (Gulko et al.) integrates DNase-seq, RNA-seq and histone modification data to create an evolution-based measure of phenotypic function.

None of these methods, however, take direct advantage of protein *domain* information, thereby potentially ignoring valuable information within the various functional portions of a protein. By integrating the ExAC database of protein-coding genetic variation taken from more than 60,000 human exomes with the Pfam database, we have comprehensively measured the landscape of genetic variation among all characterized protein domains. Computing variant densities, dN/dS ratios, and the distribution of those ratios for each domain per protein will allow us to develop a model that should more accurately predict the likelihood that a variant in a particular genomic location will actually lead to phenotypic change. The rationale of the model is that variants coinciding with protein domains with a high dN/dS value or tolerance for variation are less likely to have a functional impact, with the corollary being that variants affecting less tolerant domains are more likely to perturb protein function. We expect that comparing measures of each domain's intra-species variant "constraint" with inter-species conservation measures will further inform variant effect predictions.

We also aim to incorporate non-domain regions, the regions of protein in between and on the side of domains, which we call nodoms, so that we have a point of comparison across a protein. Additionally, we aim to utilize 3D positioning for variants – the location on a protein may indicate whether the variant is as deleterious as we might think. In that same vein, whether a variant overlaps an active site will also be taken into account. We will present our efforts to develop and validate a predictive model that integrates this information to reduce false negative and false positive predictions of the functional impacts of genetic variation in both research and clinical settings.

ENCORE: A COMPREHENSIVE FRAMEWORK FOR CANCER SEQUENCING ANALYSIS

Miao He, Jennifer Becq, Stefano Berri, Mark Ross, David Bentley

Illumina, Population & Medical Genomics Dept, Little Chesterford, United Kingdom

The past decade has seen tremendous benefit from cancer genome sequencing (CGS). Whole genome sequencing (WGS) of cancer allows us to identify somatic changes in the forms of single nucleotide variants (SNVs), insertions and deletions (indels), copy number variants (CNV) and structural variants (SVs). Research in this area has expanded our knowledge of the disease as well as of the patterns of mutations, such as mutational signatures [1], localized hypermutation [2] and of tumour heterogeneity [3]. CGS has also brought huge potential in advancing cancer diagnostics and informing treatment options. As sequencing technologies continue to advance, robust bioinformatics tools are needed to meet the demands of increasing data volume.

We present Encore, a computational framework in R for analysing WGS data of cancers and their matched normal tissues. It contains a rich collection of methods for calculating metrics and producing graphs. Using somatic variants (SNVs, indels, SVs, CNAs) as inputs, Encore generates analysis reports and presents many important features of the tumour genome. Our standard report is designed for a tumour-normal pair and provides summary and analysis including reporting of known cancer variants and hotspots, genome-wide visualization of SVs and CNVs, detection of potential gene fusions, mutational context and signature composition analysis, and localized hypermutation (“kataegis”) and double minute detection. Our multi-sample reports support analysis of multiple tumour samples from a single individual or multiple individuals. They are useful for monitoring cancer progression, comparing laboratory techniques and analysis workflows, and studying cancer cohorts. With these features, Encore provides a powerful framework for in-depth analysis of cancer sequencing data to meet different analysis needs.

References:

1. Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013)
2. Nik-Zainal S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012)
3. Gerlinger M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892 (2012)

METATREEMAP: VISUAL REPRESENTATION OF TAXONOMIC ASSIGNMENT.

Maxime Hebrard, Todd D Taylor

RIKEN Center for Integrative Medical Sciences (IMS), Laboratory for Integrated Bioinformatics, Yokohama, Japan

Metagenomic samples contain hundreds or even thousands of different species. Various software tools, such as MetaBin, have been developed for taxonomic classification of metagenomic reads using many different approaches. Regardless of the methodology, the reads are assigned to specific taxonomic entities (taxa). The classification results constitute a phylogenetic tree with a number of reads assigned to each taxon (node).

While phylogenetic trees with only a few dozen nodes are easy to display, large metagenomic trees contain so many taxa that they are difficult to visualize and comprehend. Not only do we need to legibly display the hierarchy of the tree, we also need to know the number of the reads at each node, especially when comparing two or more samples. Because linear representation of a huge tree is cumbersome and nodes weights are almost meaningless, because circular representation difficult to read and grasp spatially, we have adapted a new representation method that addresses these weaknesses.

A treemap is a drawing method that represents a hierarchy as nested rectangles. Each element of the hierarchy (node) is converted to a rectangle. Each sub-node is then a sub-rectangle. In addition, the area of each rectangle is proportional to the associated quantity (assigned read number). The final result is a tile-like figure where the larger tiles represent the more abundant species in the dataset. An interesting property of treemaps is that sub-branches of a tree are represented as intermediate rectangles, and the area of these rectangles is proportional to the sum of the reads assigned to the sub-branch. Accordingly, all the reads are represented in a compact flat view that maintains the hierarchy. Our tool uses treemaps to enhance the display of phylogenetic rank and allows researchers to easily browse through depth levels by rank selection, color switching, zooming and searching functions. We also display additional information, such as tooltips, inside the treemap and in synchronized spreadsheets (same color and zoom functions). Furthermore, multiples samples can be loaded and visualized at the same time allowing visual and numerical comparison.

MetaTreeMap is a web interface, without any data saving to assure privacy. The software allows online analyses with user-friendly options and dynamic features. The visualization is highly customizable to fit with user needs and can be easily exported for downstream use. It renders phylogenetic trees and focuses on taxonomic assignment visualization producing compact and browsable figures. Thousands of taxa can be rendered in one treemap from one or more samples and all the information can be compared visually or numerically. Data can be imported from a JSON file or from a tabular file. Treemaps can be saved as PNG images, and tree structures can be exported in JSON format.

HIGHLANDER: VARIANT FILTERING MADE EASIER

Raphael Helaers, Miikka Vikkula

de Duve Institute, Université catholique de Louvain, Human Molecular Genetics, Brussels, Belgium

The field of human genetics is being revolutionized by exome and genome sequencing. A massive amount of data is being produced at ever-increasing rates. Targeted exome sequencing can be completed in a few days using NGS, allowing for new variant discovery in a matter of weeks. The technology generates considerable numbers of false positives, and the differentiation of sequencing errors from true mutations is not a straightforward task. Moreover, the identification of changes-of-interest from amongst tens of thousands of variants requires annotation drawn from various sources, as well as advanced filtering capabilities. We have developed Highlander, a Java software coupled to a MySQL database, in order to centralize all variant data and annotations from the lab, and to provide powerful filtering tools that are easily accessible to the biologist. Data can be generated by any NGS machine (such as Illumina's HiSeq, or Life Technologies' Solid or Ion Torrent) and most variant callers (such as Broad Institute's GATK or Life Technologies' LifeScope). Variant calls are annotated using DBNSFP (providing predictions from 6 different programs, and MAF from 1000G and ESP), GoNL and SnpEff, subsequently imported into the database. The database is used to compute global statistics, allowing for the discrimination of variants based on their representation in the database. The Highlander GUI easily allows for complex queries to this database, using shortcuts for certain standard criteria, such as "sample-specific variants", "variants common to specific samples" or "combined-heterozygous genes". Users can browse through query results using sorting, masking and highlighting of information. Highlander also gives access to useful additional tools, including direct access to IGV, and an algorithm that checks all available alignments for allele-calls at specific positions.

ON THE WIDESPREAD AND CRITICAL IMPACT OF BATCH EFFECTS AND SYSTEMATIC BIAS IN SINGLE-CELL RNA-SEQ DATA

Stephanie C Hicks^{1,2}, Mingxiang Teng^{1,2}, Rafael A Irizarry^{1,2}

¹Dana-Farber Cancer Institute, Department of Biostatistics and Computational Biology, Boston, MA, ²Harvard T. H. Chan School of Public Health, Department of Biostatistics, Boston, MA

Single-cell RNA-Sequencing (scRNA-Seq) has become the most widely used high-throughput method for transcription profiling of individual cells. Batch effects and systematic errors have been widely reported as a major challenge in high-throughput technologies. Surprisingly, there is little to no mention of these challenges in published studies based on scRNA-Seq technology. We examined data from five published studies and found that batch effects can explain a substantial percentage of observed cell-to-cell expression variability. Specifically, we found that the proportion of genes reported as expressed varies systematically from batch to batch. Furthermore, we found that the implemented experimental designs confounded outcomes of interest with batch effects, a design that can bring into question some of the conclusions of these studies. Finally, we consider the consequences of failing to adjust for this unwanted technical variability, and propose new strategies to minimize its impact on scRNA-Seq data.

THE ENCODE UNIFORM ANALYSIS PIPELINES CASE STUDIES IN CLOUD-BASED DATA ANALYSIS AND DISTRIBUTION

Benjamin C Hitz¹, J Seth Strattan¹, Esther T Chan¹, Tim Dreszer¹, Jean M Davidson¹, Nikhil R Poddurturi¹, Laurence D Rowe¹, Cricket A Sloan¹, Forrest Y Tanaka¹, Carrie Davis², Alex Dobin², Sarah Djebali³, Roderic Guigo³, Tom Gingeras², Colin Dewey⁴, Xintao Wei⁵, Brenton Graveley⁵, J M Cherry¹

¹Stanford University, Genetics, Palo Alto, CA, ²Cold Spring Harbor Laboratory, Genome Research Center, Cold Spring Harbor, NY, ³Centre for Genomic Regulation, Barcelona, Spain, ⁴University of Wisconsin, Biostatistics, Madison, WA, ⁵University of Connecticut, Genetics and Developmental Biology, Farmington, CT

From Ammon's horn to zone of skin, members of the ENCODE Consortium have measured RNA quantity, RNA-protein interactions, DNA-protein interactions, DNA methylation, replication timing, chromatin structure, and histone modifications in over 4,000 experiments on more than 400 cell or tissue types. The ENCODE Data Coordination Center (DCC) have built a new web resource, the ENCODE Portal, to distribute the results of these experiments. Web-based faceted browsing and search are supported, as is programmatic access through the ENCODE REST API. The ENCODE Data Analysis Center (DAC) have specified uniform processing pipelines for four ENCODE datatypes: ChIP-seq, RNA-seq, DNase-seq, and whole-genome bisulfite sequencing. The DCC have implemented these pipelines and deployed them to a cloud-based platform. The results of these analyses and metadata describing them are distributed through the ENCODE Portal, and illustrate general methods of accessing and interpreting ENCODE data. The ENCODE Portal is <https://www.encodeproject.org/>. The DCC codebase is freely available at <https://github.com/ENCODE-DCC/>.

INTEGRATED APPROACH TO IDENTIFY CLINICALLY RELEVANT CNVs IN CLINSEQ® COHORT

Celine Hong¹, David Ng¹, Jennifer Johnston¹, Dan King¹, Jim Mullikin^{2,3}, Leslie G Biesecker^{1,2}

¹National Institutes of Health, Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, Bethesda, MD, ²National Institutes of Health, NIH Intramural Sequencing Center, National Human Genome Research Institute, Bethesda, MD, ³National Institutes of Health, Comparative Genomics and Cancer Genetics Branch, National Human Genome Research Institute, Bethesda, MD

With the advancements in sequencing technology, the ability to sequence genomes or exomes is becoming widely accessible for researchers. Compared to genome sequencing, exome sequencing is an efficient and an attractive method to study human diseases, due to a limited understanding of non-coding regions and the substantial computational resources required to process and analyze genome sequence data. In addition to identifying single nucleotide variants, copy number variations (CNVs) can be predicted from exomes by using CNV calling algorithms. Here, we pilot a study to show how exomes can be systematically screened for clinically relevant CNVs in a healthy cohort for diagnoses. ClinSeq® cohort comprise 978 participants from the metro Washington D.C. and Baltimore areas. Exome sequencing is done for all participants and their personal and family medical history is recorded. We applied XHMM (eXome Hidden Markov Model) on 978 ClinSeq® exomes. A total of 12,654 CNVs were predicted. A 100% confirmation rate was achieved when validating putative CYP2D6 pharmacogenetic CNVs. Using a 0.5% ClinSeq® population frequency cutoff, there were 123 long CNVs (>200 kb) and 932 small CNVs (<200 kb) remaining. The putative CNVs were further filtered by using HGMD, OMIM, and current literature, and categorized into the following categories: benign, clinically relevant with known significance, clinically relevant with unknown significance, and unknown. Clinically relevant CNVs can be validated by RT-PCR, exome or immuno SNP chip, and genomes. A mosaic Trisomy 12 signal was detected and confirmed in one proband. Trisomy 12 is the third most frequent chromosomal aberration in chronic lymphocytic leukemia (CLL), and patients with Trisomy 12 have 9-11 year median overall survival. The PMP22 deletion, which causes hereditary neuropathy, was detected and confirmed in three probands. Our study demonstrates that the exomes can successfully be utilized to identify clinically relevant CNVs in a healthy population for clinical diagnoses. With further improvements in CNV calling algorithms, screening for CNVs from exome for diagnoses can become a routine clinical care in the future.

DE NOVO ASSEMBLY AND NEXT-GENERATION SEQUENCING TO ANALYZE FULL-LENGTH GENE VARIANTS FROM CODON-BARCODED LIBRARIES

Byungjin Hwang, Hoon Jang, Sunghoon Huh, Duhee Bang

Yonsei University, Chemistry, Seoul, South Korea

Interpreting epistatic interactions is crucial for understanding evolutionary dynamics of complex genetic systems and unveiling structure and function of genetic pathways. Although high resolution mapping of *en masse* variant libraries renders molecular biologists to address genotype-phenotype relationships, long-read sequencing technology remains indispensable to assess functional relationship between mutations that lie far apart. Here, we introduce JigsawSeq for multiplexed sequence identification of pooled gene variant libraries by combining a codon-based molecular barcoding strategy and *de novo* assembly of short-read data. We first validate JigsawSeq on small sub pools and observed high precision and recall at various experimental settings. With extensive simulations, we then apply JigsawSeq to large-scale gene variant libraries to show that our method can be reliably scaled using next-generation sequencing. JigsawSeq may serve as a rapid screening tool for functional genomics and offer the opportunity to explore evolutionary trajectories of protein variants.

THE HUMAN GENE DAMAGE INDEX: A NOVEL GENE-LEVEL APPROACH TO PRIORITIZE EXOME VARIATIONS

Yuval Itan¹, Lei Shang¹, Lluís Quintana-Murci², Shen-Ying Zhang^{1,2}, Laurent Abel^{3,1}, Jean-Laurent Casanova^{1,3,4}

¹The Rockefeller University, Human Genetics of Infectious Diseases, New York, NY, ²Institut Pasteur, Human Evolutionary Genetics, Paris, France, ³INSERM, Human Genetics of Infectious Diseases, Paris, France, ⁴Howard Hughes Medical Institute, New York, NY

The exome of a patient with a monogenic disease contains about 20,000 variations, of which only one or two are disease-causing. Variant- and gene-level *in silico* methods have been developed to select candidate variations prior to their experimental study. We aimed to develop a novel gene-level approach to predict whether specific variations in any given protein-coding gene may be disease-causing. We noticed that 58.32% of exome variants in the general population are found in only 2.42% of human genes. We thus developed the gene damage index (GDI): a novel genome-wide, gene-level estimate of accumulated mutational damage in the general population. We found correlations between GDI and selective evolutionary pressure, protein complexity, coding sequence length, and number of paralogs. We also showed that GDI better differentiates between truly disease-causing and falsely positive genes than three leading gene-level methods (genic intolerance, gene indispensability, and *de novo* excess). We further defined a high-GDI cutoff that can successfully filter out up to 60.62% of spurious variants from patients' exome highly mutated genes. Conversely, a low-GDI cutoff points to prime candidate mutations, in genes never or rarely mutated. This novel method should facilitate human genetic studies, especially for monogenic disorders.

THE MUTATION SIGNIFICANCE CUTOFF (MSC): A GENE-SPECIFIC APPROACH TO PREDICTING THE IMPACT OF HUMAN GENE VARIANTS

Yuval Itan¹, Lei Shang¹, Lluís Quintana-Murci², Shen-Ying Zhang^{1,2}, Laurent Abel^{3,1}, Jean-Laurent Casanova^{1,3,4}

¹The Rockefeller University, Human Genetics of Infectious Diseases, New York, NY, ²Institut Pasteur, Human Evolutionary Genetics, Paris, France, ³INSERM, Human Genetics of Infectious Diseases, Paris, France, ⁴Howard Hughes Medical Institute, New York, NY

The proposed thresholds of significance for current predictors of the biological impact of human genetic variants, such as CADD score, are identical for all genes across the genome. However, we found large differences between the predicted scores of the known pathogenic mutations of 4,015 disease-causing genes. As a result, the prediction rate for proven disease-causing alleles was found to be less than 40%. We thus estimated the 95% confidence intervals of the CADD scores for all known mutations in 17,765 human protein-coding genes. By inference from the subgroup of known disease-causing genes, we defined a gene-specific mutation significance cutoff (MSC) for each of these genes. We then validated the prediction power of this approach through both simulation and analyses of real data. In particular, the true positive prediction rate for a new set of proven disease-causing alleles increased from 34.10% for the regular CADD score to 95.15% with MSC. The MSC can be used to select candidate mutations in the exomes of patients with inborn errors of known or unknown disease-causing genes. The MSC greatly improves the predictive power of methods without gene-specific thresholds.

LARGE SCALE CORRELATION OF EPIGENOMICS DATA

Jonathan Laperle¹, Alexei Nordell-Markovits², Marc-Antoine Robert², David Bujold³, David Anderson De Lima Morais^{3,4}, Michel Barrette⁴, Guillaume Bourque^{3,5}, Pierre-Étienne Jacques^{1,2,4}

¹Université de Sherbrooke, Informatique, Sherbrooke, Canada, ²Université de Sherbrooke, Biologie, Sherbrooke, Canada, ³McGill University Genome Québec Innovation Center, Bioinformatics, Montréal, Canada, ⁴Université de Sherbrooke, Centre de calcul scientifique, Sherbrooke, Canada, ⁵McGill University, Human Genetics, Montréal, Canada

As part of the International Human Epigenome Consortium (IHEC) Data Portal functionalities (<http://epigenomesportal.ca/ihec/>), we implemented a tool aimed at efficiently performing pairwise correlation of thousands of epigenomic datasets, with the possibility for the user to submit its own private data to compare with public ones. This could be useful, for instance, to help the characterization of a dataset or as a quality control. We will present the design and implementation of this tool engineered to run through MPI on a compute cluster in order to analyze massive amount of datasets in a reasonable time frame.

The various features of this tool, integrated into the Galaxy framework of the Genetics and genomics Analysis Platform (GenAP) project, include 1) the support of many genomic file formats (bigWig, WIG, bedGraph, BAM), 2) the possibility to compute the correlations at different resolutions (from 100 bp to 10 Mb), 3) using different metrics (e.g. Pearson, Spearman), 4) on the complete genome as well as on different subsets of regions (e.g. genes, TSS, user-defined), 5) and the processing of the generated correlation matrix with different clustering algorithms to display the results as heatmap and/or dendrogram. We will also offer datasets from model organisms generated by international consortia such as modENCODE as well as data downloaded from GEO/SRA and uniformly processed, along with a user-friendly interface facilitating the selection of the desired datasets to use for the correlations. A comparison of our tool with existing ones will also be presented.

The GenAP project (genap.ca) facilitates the installation and the usage of genetics and genomics state-of-the-art analysis pipelines in the Compute Canada High Performance Computing (HPC) facilities. This resource is already enabling Canadian life science researchers and clinicians to explore their data in novel ways without intermediates, and enrich their research and translational programs.

INTERROGATING THE MECHANISMS OF SCHIZOPHRENIA GENETIC RISK IN THE FULLY CHARACTERIZED HUMAN BRAIN TRANSCRIPTOME

Andrew E Jaffe^{1,2}, Jooheon Shin¹, Richard E Straub¹, Ran Tao¹, Yuan Gao¹, Yankai Jia¹, Leonardo Collado-Torres^{1,2}, Jeffrey T Leek², Thomas M Hyde^{1,3}, Joel E Kleinman^{1,3}, Daniel R Weinberger^{1,3}

¹Lieber Institute for Brain Development, Clinical Sciences, Baltimore, MD, ²Johns Hopkins University, Biostatistics, Baltimore, MD, ³Johns Hopkins University, Psychiatry, Baltimore, MD

Genetic risk for schizophrenia has begun to emerge through large genome-wide association studies (GWAS) in hundreds of thousands of individuals. However, the exact gene(s) and/or transcript(s) that are being regulated by these risk SNPs are largely uncharacterized due to the difficulty in obtaining expression and genotype data in large samples of postmortem human tissue. Advances in RNA sequencing (RNA-seq) have further permitted flexible and largely unbiased characterization of high-resolution transcriptomes, but the incomplete annotation of the human brain transcriptome can potentially affect the ability to use existing tools that rely on complete gene structure information. We have therefore sequenced the transcriptomes of the dorsolateral prefrontal cortex (DLPFC) from 320 non-psychiatric controls across the lifespan at deep coverage, including 50 second trimester fetal samples, and 175 samples from patients with schizophrenia, and characterized their expression profiles across five summarizations that capture elements of transcription – genes, exons, junctions, transcripts, and expressed regions. We show that annotation-agnostic approaches like junction and expressed-region analysis may outperform gene-, exon- and transcript-based approaches when the annotation is incomplete. We further conducted global expression quantitative trait loci (eQTL) analyses across the five expression summarizations in the adult control samples (age > 13, N=237), and identify hundreds of thousands of features that associate with local genetic variation, including extensive genetic regulation of previously unannotated sequence. The eQTLs in junction-level data (N= 53,497 unique junctions annotated to 16,481 genes at FDR < 0.01) showed the largest effect sizes (fold change per allele copy) and identified SNPs as eQTLs with the lowest minor allele frequencies (18.1% versus 23.1-24.2%). We lastly identified eQTLs to specific transcript elements in individual genes in over half of the genome-significant genetic variants for schizophrenia identified genome-wide association studies (GWAS), illuminating potential mechanisms of risk for many of these genetic variants. Leveraging human postmortem brain data can therefore fine map the functional effects of genetic risk variation for schizophrenia identified in large GWAS, and can identify novel targets for drug discovery and more focused biological assays.

PLANT REACTOME: A REFERENCE RESOURCE FOR ANALYZING PLANT METABOLIC AND REGULATORY PATHWAYS

Pankaj Jaiswal¹, Justin Preece¹, Vindhya Amarasinghe¹, Palitha Dharmawardhana¹, Peter D'Eustachio², Sushma Naithani¹, Guanming Wu³, Antonio F Mundo⁴, Robin Haw³, Sheldon Mckay³, Joel Weiser³, Lincoln Stein³, Doreen Ware^{5,6}

¹Oregon State University, Botany and Plant Pathology, Corvallis, OR, ²New York University School of Medicine, Biochemistry and Molecular Pharmacology, New York, NY, ³Ontario Institute of Cancer Research, Toronto, Canada, ⁴European Bioinformatics Institute, Hinxton, United Kingdom, ⁵Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ⁶USDA-ARS, Robin W. Holley Center for Agriculture & Health, Ithaca, NY

The Plant Reactome pathway portal (<http://plantreactome.gramene.org>) of the Gramene database, provides a reference resource for analyzing plant metabolic and regulatory pathways. It uses a data model that organizes gene products, small molecules (compounds) and interactions into reactions and pathways, providing a framework for studying an organism's cellular metabolism and regulation in response to developmental and environmental signals. The current version of the Plant Reactome features about 200 curated reference *Oryza sativa* (rice) pathways and gene homology-based projections for 33 plant species with sequenced genomes, including *Arabidopsis thaliana* and *Zea mays*. The reference pathways and reactions were assembled by manual curation of computationally integrated metabolic pathway data developed earlier using Pathway-Tools software, orthologous projections from the Human Reactome for cell cycle, replication, transcription and translation, and manual reconstruction of new pathways reported in the literature. The presentation will discuss tools for pathway enrichment analysis and homologue pathway comparison, development of the Plant Reactome portal, curation of reference rice pathways, and phylogeny based analyses of projected pathway annotations. The Gramene database project is supported by an NSF award (IOS-1127112). Intellectual and infrastructure support for the Plant Reactome is provided by the Human Reactome award (NIH: P41 HG003751, ENFIN LSHG-CT-2005-518254, Ontario Research Fund, and EBI Industry Programme).

BAMQC: A QUALITY CONTROL TOOL FOR MAPPED NEXT GENERATION SEQUENCING DATASETS

Ying Jin, Oliver Tam, David Molik, Molly Hammell

Cold Spring Harbor Laboratory, Cancer Center, Cold Spring Harbor, NY

Summary: With the recent advances in sequencing technologies, next generation sequencing (NGS) has become a standard analysis tool for the biological sciences. However, raw sequencing data must undergo several analysis steps to retrieve useful information, and often the quality of the sequencing libraries is not fully assessed until all data analyses are complete. A generalized data analysis pipeline for NGS data includes: removing adapters and low-quality reads, mapping to the reference genome or performing a de novo alignment of the sequences, followed by more complicated bioinformatics assessments that depend upon the library type (such as differential expression analysis and generic variant calling). Data analyses can be a bottleneck for sequencing assays, therefore, it is important to have an efficient and easy-to-use tool that can help biologists and even bioinformaticians quickly determine whether the data is sufficient for further analysis. This is especially true for assays that push the boundaries of sequencing technologies, such as single-cell sequencing assays that frequently contain individual libraries of poor quality within a larger cohort of samples. Here, we present a software package, BAMQC, to check the quality of mapped sequencing libraries. It conducts comprehensive evaluations of aligned sequencing data including: genomic mapping distributions, transcriptomic mapping distributions, rRNA contamination rates (for RNA-seq), transcript 5' and 3' read bias (for RNA-seq), PCR duplication rates or biases, sample saturation rates, and sample correlations. Summaries for each QC module are represented in both tabular and graphical format. While many of the quality assessments have been specifically designed for problematic issues common to single cell RNA-seq assays, BAMQC can take any mapped library in the BAM file format. BAMQC can be applied to multiple assays such as RNA-seq, CLIP-seq, Gro-seq, ChIP-seq, DNA-seq and so on.

Availability: BAMQC is implemented in python and is freely available at <http://hammelllab.labsites.cshl.edu/software>.

Contact: mhammell@cshl.edu

UTILIZATION OF HIGH-THROUGHPUT SEQUENCING TO DETECT CANDIDATE GENES ASSOCIATED WITH MOUSE REPRODUCTIVE LONGEVITY

Jyoti Joshi*¹, Kacper Żukowski*¹, Nehil Jain*¹, Jeremy E Koenig¹, Robert G Beiko*¹, Hein van der Steen*²

¹Dalhousie University, Faculty of Computer Science, Halifax, Canada,

²Performance Genomics Inc., Bible Hill, Canada

Reproductive longevity (RL) is a complex trait that depends on a number of factors, such as longevity, ovarian function, fertility, stress resistance and health. Reproduction has a substantial impact on longevity; higher reproduction reduces survival, and limiting reproduction increases life span. Identification of RL genes is important for economically important livestock species in order to improve and increase the overall productivity of their herds, especially in the genomic selection era.

Our principal project objective is to determine the genetic architecture of longevity by identifying which candidate genes and functional pathways are responsible for variation in reproduction in mammals. The project is founded on mouse lines which have been selectively mated for 30 years with fertility and RL data recorded, and corresponding randomly mated control groups. As a result of this, selected (S) mice reproduce up to 86% longer than the control (C) mice and live through 100% more pregnancies. We used whole-genome shotgun sequencing (WGS) and transcriptomics (RNAseq) to investigate variation in single nucleotide polymorphisms (SNPs) and gene expression in the mouse dataset. In both the studies, we analysed the contrast between S and C mice lines. The results of WGS showed over 580,000 significant SNPs (Bonferroni corrected p-value < 0.05) in gene regions covering a total of 13,400 genes. RNAseq data showed that 396 significant genes (adjusted p-value, FDR <5%) were down-regulated and 31 genes up-regulated in the S line, relative to the C line. The result indicates a large number of differential expressed genes in the S line that relate to reproduction are down-regulated and are functional for a longer duration as compared to C.

The selection of gene ranking based on TopQ method and different sets of gene expression levels between S and C mice were produced using WGS and RNAseq respectively, and we performed meta-analysis to produce a comprehensive ranking of selected genes which allowed us to identify relevant functional pathways. Using comparative genomics, mouse results are being used to complement studies in other mammalian species, in order to guide the development of genomic applications. The first target of our comparative approach is the development of commercial DNA marker tests of RL (Herd life) for Holstein cattle breeding. The integration of mouse and cattle data leads to the improvement of genomics evaluation, selection of cattle for RL traits, reduction of replacement costs and increased milk production.

CONSTRUCTION OF TRANSCRIPTION FACTOR NETWORKS FOR OBESITY USING RNASEQ TRANSCRIPTOMICS

Ruta Skinkyte-Juskiene, Lisette J Kogelman, Haja N Kadarmideen

University of Copenhagen, Faculty of Health and Medical Sciences, Frederiksberg C, Denmark

Obesity is the result of an interplay between genetic and environmental factors. However genes don't change so quickly as environment. During the last thirty years obesity rates have tripled and have increased particularly fast in children. Transcription factors (TF) are proteins that control genes' actuation in the genome. By binding to DNA and other proteins hundreds of transcription factors define the core of regulatory networks according to cell statement or environmental conditions. Some MicroRNA networks regulating obesity were described previously (Arner et al., 2015), but none of TFs networks for obesity to our knowledge were constructed before. We built Transcription Factor Networks based on RNAseq data from porcine model for obesity published by Kogelman et al., BMC Medical Genomics 2014 Sep 30;7:57 (PMID: 25270054) and made publicly accessible at the Gene Expression Omnibus (GEO) database of NCBI at: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61271>. The downloaded RNAseq database (Series GSE61271) represented three different degrees of obesity (obese, lean and medium). Using this dataset, we identified 1415 known transcription factors. We then built a TF Network using Weighted gene co-expression network analysis (WGCNA). WGCNA detected three modules possessing correlations with obesity ranging from -0.41 to 0.47 ($P < 0.01$). Functional annotation of selected modules was performed using GOSep and GeneNetwork. Using GOSep, significant results were obtained only from the Blue module, containing 79 GO categories ($FDR < 0.05$) e.g. positive regulation of response to stimulus ($P = 3.89E-08$), positive regulation of immune system process ($P = 9.86E-07$) and one KEGG pathway, i.e. phagocytosis ($FDR < 0.1$). GeneNetwork analysis of the Blue module identified KEGG pathways associated with Tumor syndrome, Osteoporosis (Chemokine signaling pathway, $P = 1.92E-07$) and etc. By narrowing thresholds of inter-modular connectivity and module membership in the Blue module were detected five hub (highly interconnected) genes: MSR1, RPS6KA1, HCLS1, LPXN, and SPI1. In a previous study MSR1 and SPI1 have been detected as regulator genes and they are proposed link to obesity and osteoporosis. Obtained results suggest several promising candidates for further TF network study on the differentially connected genes. Differentially wired TF Networks will be further developed for lean vs. obese, lean vs. median and median vs. obese networks, to provide higher-resolution insight into the mechanisms that drive obesity development.

GENETIC NETWORK METHOD BASED ANALYSIS OF ANTIDEPRESSANT TREATMENT

Majbritt B Madsen¹, Lisette J Kogelman², Henrik B Rasmussen¹, Haja N Kadarmideen²

¹Mental Health Services of the Capital Region of Denmark, Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Roskilde, Denmark, ²University of Copenhagen, Faculty of Health and Medical Sciences, Frederiksberg C, Denmark

Selective serotonin reuptake inhibitors (SSRIs) are the most commonly prescribed antidepressants used for treatment of major depressive disorder (MDD). The response to SSRI treatment varies among individuals, which primarily is thought to stem from genetic factors. Genome-wide association studies (GWAS) have attempted to identify predictive genetic markers for SSRI efficacy and tolerability but with limited success, probably reflecting the need for larger sample sizes, which is difficult to achieve in pharmacogenetic studies. Thus, other methods are needed to identify genetic components, which can lead to stratification of patients and personalised treatment of MDD. Network based methods are novel approaches for analysis of GWAS that incorporates biological information. Previously, a GWAS has been performed on 499 SSRI treated MDD patients¹. MDD symptom severity was scored at baseline, after 4 and 8 weeks. Here we present a network method analysis of these GWAS data. Based on their pairwise correlations we selected the most connected SNPs and clustered them in modules. The Weighted Interaction SNP Hub (WISH) method was used to determine the epistatic interaction among SNPs and select SNPs that jointly affect the response to SSRI treatment. The genes tagged by the selected SNPs were analysed to uncover their involvement in pathways and tissue expression. The following tissues were enriched: nervous-, digestive-, cardiovascular systems, genitalia, bones and joints. Enriched pathways included those of motor skills and limb coordination, carbohydrate transportation and metabolism, behaviour, axon morphology, catecholamine transport and metabolism, acetylcholine receptor function, and vascular endothelial regulation. We suggest that the interpersonal variability of the SSRI treatment response is influenced by genes in the above mentioned pathways, and that the key to personalising SSRI treatment should be found within these genes.

¹Ji Y, Biernacka JM, Hebring S, Chai Y, Jenkins GD, Batzler A, et al. Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. *Pharmacogenomics J.* Oktober 2013;13(5):456–63

BUILDING A META-METAGENOME GRAPH

Andre Kahles, Gunnar Ratsch

Memorial Sloan Kettering Cancer Center, Computational Biology, New York, NY

Accurate identification of the composition of a microbial community from sequencing data has become a central task in clinical research and diagnostics and is quickly gaining further importance. Currently established methods assign variants of 16S ribosomal RNA (rRNA) or reads from whole genome shotgun sequencing (WGS) to single entities in a given taxonomy tree. The major limitations of these approaches arise from incomplete taxonomies, the inaccurate representation of the true phylogenetic relationships by a tree structure and assignment ambiguities. Moreover, species assignments often do not allow for a functional evaluation of resistance or toxigenicity – both of which are of major importance in the clinical setting. Lastly, it is very difficult to include evidence of new and rare species collected in earlier studies into the reference set. We therefore have implemented a new, highly sensitive approach to combine, represent and identify the microbial and/or functional composition of a large set of metagenome samples with a major focus on taking previous knowledge into account. Building on techniques from genome assembly and text compression, we use succinct data structures to efficiently represent all sequence information in a k-mer based assembly graph, which not only represents single species and their individual relationships but also captures intra-species variability. The graph is structured as a self-index that can be used for alignment and annotation of reads arising from metagenome sequencing. Although the graph can in principal be built from 16S rRNA sequences, its full utility is tailored for the use of WGS data, where not only unknown species can be represented in the correct relationship to known species but also single functional entities, e.g., single genes can be identified. The constructed reference graph leverages information from known genomes as well as from the many previous studies, giving access to rare observations not yet present in reference databases. It is designed to integrate further knowledge over time, e.g., to accumulate information over many patients and studies. Thus, it will have greater sensitivity to detect unseen or rarely seen species and inherently represents nearest neighbors with less bias towards species overrepresented in existing databases. Clinical applications of our method include comparing microbial populations of individual patients to find functional differences, the detecting emerging drug-resistances or identifying disease relevant risk/protective taxonomic or functional units. Yet, our method is not limited to the clinical field and we envision a wide range of possible applications.

A DYNAMIC PROGRAMMING APPROACH TO THE RECONSTRUCTION OF PROKARYOTE GENE BLOCK EVOLUTION HISTORY

Carly Schaeffer¹, David Ream², Iddo Friedberg^{2,3}, John Karro^{1,2,4}

¹Miami University, Computer Science and Software Engineering, Oxford, OH, ²Miami University, Microbiology, Oxford, OH, ³Iowa State, Veterinary Microbiology, Ames, IA, ⁴Miami University, Statistics, Oxford, OH

In prokaryote genomes there is a strong syntenic phenomenon known as gene blocks: genes located in close proximity on a chromosome, whose product usually participate in the same cellular function. In many cases these blocks are operons: genes that are co-transcribed to a single mRNA under a strict regulatory mechanism. The underlying selection processes for the formation, conservation, and/or dissolution of such complex genomic systems are an open problem in evolutionary biology. In Ream et al. [1] the evolutionary events that can lead to more complex block structures were investigated (e.g. gene loss, gene duplication, block splitting, etc) with an attempt to determine the block structure of ancestral species, using simple localized heuristics for ancestral assignment. Here we present a formal algorithm for finding the most globally-parsimonious assignment of block-modification events given a set of species with a partially conserved block structure and the corresponding phylogenetic tree. By combining a variation of the standard weighted-parsimony approach to ancestral sequence reconstruction with a modified algorithm for edit-distance computation, we are able to assign each interior a block structure in order to achieve a global-minimization of the total event cost. Further, but incorporating estimated weights from [1], our method has the potential to move from parsimony to maximum likelihood estimation.

[1] Ream DC, Bankapur AR, Friedberg I An Event-Driven Approach for Studying Gene Block Evolution in Bacteria Bioinformatics (2015)

CHOP-STITCH: TARGETED ASSEMBLY OF GENES USING TRANSCRIPTOME ASSEMBLY AND BLOOM FILTER-BASED DE BRUIJN GRAPHS

Hamza Khan, Benjamin P Vandervalk, René L Warren, Inanc Birol

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada

Recent advancements in next-generation sequencing (NGS) technologies have enabled sequencing of non-model organisms with reasonable budgets. Such studies often interrogate the genomes and transcriptomes of these species with massive amounts of short sequences. Reconstructing genomes and transcriptomes using these datasets requires de novo assembly methods, when the species or closely-related species lacks a high quality reference resource. Carrying out full de novo genome assemblies is not a trivial task, especially for organisms with large and complex genomes. For certain applications, limiting the scope of assembly to reconstructing the genic space may be desirable. We hereby present a new method, Chop-Stitch, for targeted assembly of genes using transcriptome assembly and whole genome sequencing data as inputs. We implemented a k-mer based method that first identifies exon-exon boundaries in de novo assembled transcripts with the help of a Bloom filter that represents the k-mer spectrum of genomics reads. The k-mer content of assembled transcripts are interrogated in this Bloom filter to detect exon-exon junctions. The same Bloom filter is also used to construct an implicit de Bruijn graph of the underlying genome. Traversals on this graph are recorded to fill in the nucleotides of the predicted introns searching for paths that connect adjacent exons. The output of our tool is a FASTA file containing sequences for genic regions. We have tested our method on *Caenorhabditis elegans*, using its annotated transcriptome and publicly available whole genome shotgun sequencing data (DRR008444). On this test, Chop-Stitch predicted exon-exon junctions with 85.9% sensitivity and 100.0% precision. This method was able to find the intermediate paths for 88.9% of the predicted exon-exon boundaries. Currently, the tool is being tested with human sequencing dataset to demonstrate its scalability for larger targets. This method could be used effectively to recover functional elements such as protein-coding genes with better accuracy and precision compared to whole genome assemblies.

ROLE OF lincRNA IN T HELPER CELL DIFFERENTIATION

Mohd M Khan^{1,2}, Ubaid Ullah¹, Omid Rasool¹, Sini Rautio³, Zhi Chen¹, Riitta Lahesmaa¹

¹Turku Centre for Biotechnology, University of Turku and Åbo Akademi University,, Turku, Finland, ²Turku Doctoral Programme of Molecular Medicine (TuDMM), Medical Faculty, University of Turku, Turku, Finland, ³Department of Information and Computer Science, Aalto University, School of Science and Technology, Espoo, Finland

Long intervening noncoding RNAs (lincRNAs) are more than 200 nucleotide long intergenic transcripts of very low coding potential playing important roles in many biological process. Little is known about the role of lincRNA in T helper cell subset differentiation. Th17 cells are one of the well-characterized subset of T helper cells that secrete cytokines including IL17, and are responsible for the elimination of bacteria and fungi at mucosal surfaces. Recently there are many reports that both in humans and mice Th17 cells play an important role in the pathogenesis of a diverse group of immune-mediated diseases, including psoriasis, rheumatoid arthritis, multiple sclerosis, inflammatory bowel disease, and Type1 Diabetes. Identification and characterization of non-coding RNA and dissecting the molecular mechanisms through which they may control Th17 cell differentiation and effector functions will may provide new insights into pathogenesis of immune mediated diseases. While profiling transcriptome of differentiating Th17 cells at early time points, we identified a non coding RNA, lincITGAE located next to a protein coding gene ITGAE. Both of the genes are differentially expressed during human Th17 cell polarization. We then demonstrated that depletion of lincITGAE by specific LNAs resulted in decreased expression of IL17a in Th17 cells. Fractionation studies to identify localization of lincITGAE reveal that lincITGAE is present in chromatin fraction of Th17 cell suggesting that may be lincITGAE recruit chromatin modifying complexes and regulate transcriptional activation of Th17 cell specific genes. Gene expression when lincITGAE are downregulated by LNAs and Chromatin isolation by RNA purification (ChIRP) to find interacting partners of lincITGAE are in progress. This study will characterize the mechanism of action of the lincRNA as well as identify new regulators for T helper cell subset differentiation. Understanding the pathways and regulatory mechanisms that define the crucial steps in T helper cell subset differentiation will be critical for the eventual development of rational therapeutics to modify the pathological immune responses in inflammatory diseases.

COMPREHENSIVE, FLEXIBLE PIPELINES FOR ALTERNATIVE POLYADENYLATION ANALYSIS

Hyunmin Kim¹, Jihye Kim², Nova Fong¹, David Bentley¹

¹University of Colorado School of Medicine, Biochem. & Mol. Genetics, Aurora, CO, ²University of Colorado School of Medicine, Division of Medical Oncology, Aurora, CO

Coding gene and long non-coding gene transcripts are cleaved and polyadenylated at their 3' ends. The majority of genes have multiple alternative polyadenylation (pA) sites and their usage preferences are specifically altered in different cellular conditions. Alternative polyadenylation (ApA) fine-tunes mRNA productivity by altering 3' untranslated regions (UTRs). Recently, high throughput polyA-seq datasets have permitted identification of precise pA sites throughout the genome. However, incomplete bioinformatics of polyA-seq remains a major source of heterogeneity in meta-analysis.

The significance of ApA is tested with various statistical assumptions.

Differential usage of distal versus proximal ApA in 3' UTRs can be tested with independent tests. Alternative linear-trend tests are known to be more powerful than the independent tests for detecting shifted distributions.

Whereas the preprocessing steps such as filtering of inter-priming positions and clustering of the alignments cause unpredictable effects on the tests.

Moreover, the performance of testing methods is dependent on the definition of target regions at which individual pA sites are considered. To address these problems we developed a flexible a BASH-oriented pipeline with the following features:

- a. Clustering proximal polyA sites using the 1D k-means clustering algorithm
- b. Filtering interpriming reads using the naïve-Bayes classifier
- c. Considering the variation between replicates using the negative Binomial model (edgeR)

We have validated these tools using public datasets and applied them to our own datasets for analysis of the effects of the translation inhibitor cycloheximide on polyA sites in transformed and non-transformed breast epithelial cells.

IDENTIFICATION AND FILTRATION OF FALSE SOMATIC VARIANTS CAUSED BY VECTOR CONTAMINATION

Junho Kim¹, Ju Heon Maeng¹, Jae Seok Lim², Junehawk Lee³, Jeong Ho Lee², Sangwoo Kim¹

¹Yonsei University College of Medicine, Severance Biomedical Science Institute, Seoul, South Korea, ²KAIST, Graduate School of Medical Science and Engineering, Daejeon, South Korea, ³Korea Institute of Science and Technology Information, Department of Convergence Technology Research, Daejeon, South Korea

Advances in next-generation sequencing (NGS) technologies have remarkably improved the detection limit of somatic variants into a very low-frequency range. However, accurate detection at this range is still confounded by many factors including environmental contaminations. Vector contamination is one of the most commonly occurred issues for sequencing experiments from Sanger to NGS technologies, and is especially problematic because vector-inserted sequences are hardly distinguishable from the sample sequences. Such inserts, which may harbor functional mutations targeted to be cloned, can cause false calling of low-frequent somatic variants at the corresponding site. Numerous vector screening methods have been developed, but none could handle the contamination from inserts by only focusing on searching and removing vector sequence itself. To address the problem, we developed a novel method Vecuum that identifies and filters vector-originate reads including inserts that may cause false somatic calls.

Since vector inserts for molecular cloning are generally constructed from processed cDNAs for their translation, we hypothesized that many vector originated reads will be split at exon junctions to represent a unique structure. Vecuum first examines all clipped sample reads for existence of vector sequence, like previous methods but in a much faster way using bwa fastmap (up to 99% reduction in running time). If a contamination is detected and genomic positions of the vector insertion site are secured, then we separate the vector-originated reads based on the clipping patterns at exon junctions which are restorable in transcript mapping. False somatic calls are identified based on the skewness of mutant allele to vector-originated reads, evaluated by one-tailed Fisher's exact test. Testing on an intentionally contaminated sample, prepared by mixing ten mutant plasmid vectors into normal blood, confirmed that Vecuum successfully detected and filtered all the vector-originated mutations to prevent false variant calling.

GENOMIC-BASED 16S RIBOSOMAL RNA DATABASE WEB SERVER AND TOOLS

Seok-Won Kim¹, Masahira Hattori², Todd D Taylor¹

¹RIKEN Center for Integrative Medical Sciences, Laboratory for Integrated Bioinformatics, Yokohama, Japan, ²Waseda University, Graduate School of Advanced Science and Engineering, Cooperative Major in Advanced Health Science, Tokyo, Japan

We constructed a manually edited 16S ribosomal RNA (rRNA) gene database called GRD. In GRD, both the 5' regions and 3' regions, including the anti-SD sites, have been carefully checked and contaminating sequences have been removed. Because of this careful manual checking of the 16S rRNA sequences, our database should be considered the most reliable reference source for downstream analyses. For utilization of GRD, we developed a web resource (<http://metasystems.riken.jp/grd/>) which allows users to search the database in a number of ways. The 16S rRNA gene sequences and accompanying information for both Bacteria and Archaea can be found by searching by taxonomic name or NCBI Taxonomy ID, or by various keywords, from the home page. Advanced search options allow the use of Boolean terms, search by sequence length, search by presence or absence of Ns in the sequences, etc. An up-to-date phylogenetic tree which is based on the NCBI taxonomy database can be used to select single or multiple taxa for download. Users may perform single or multiple FASTA formatted sequence analysis using BLAST against the GRD database. In addition, the whole GRD database or various subsets of it may be downloaded for analysis or import into other databases. More details are provided on the web site. As a next step, we are currently developing a tool for predicting 16S rRNA genes from genome sequences. General 16S rRNA gene prediction tools hitherto used are based on basic local alignment or Hidden Markov Models using known sequences. Our tool applies various concepts based not only on traditional approaches, but also on the adoption of specialized properties of the 16S rRNA gene, thus leading to more accurate prediction.

IMPROVING ACCESSIBILITY AND USABILITY OF GENOME DATA AT NCBI

Paul Kitts, Michael DiCuccio, Avi Kimchi, Terence Murphy, Kim Pruitt, Tatiana Tatusova

National Center for Biotechnology Information (NCBI), NLM, NIH, Bethesda, MD

The National Center for Biotechnology Information (NCBI) databases contain data for over 50,000 genome assemblies. NCBI has recently made several improvements that: make it easier for users to find and quickly access genome data of interest; provide more convenient data formats; and enrich the data presented in web pages and reports.

We have added new panels to the NCBI Genome Resource (www.ncbi.nlm.nih.gov/genome/) for high profile organisms, such as human and *Salmonella enterica*, that provide quick access to links that allow users to easily execute common actions: download sequences in FASTA format for genome, transcript, or protein; download genome annotation in GFF, GenBank or tabular format; BLAST against genome, transcript, or protein sequences.

We have also redesigned the NCBI genomes FTP site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>) to expand content and facilitate data access through an organized predictable directory hierarchy that has consistent file names and formats. The updated FTP site provides greater support for downloading assembled genome sequences and/or corresponding annotation data. We now provide GFF and feature table formats for all genome assemblies that are annotated. We also instituted the use of accession.version as the primary sequence identifier for both GFF and FASTA files. Having the same identifier in both the FASTA and GFF files supports the use of these files in common RNA-Seq analysis packages and in other analysis pipelines that rely on simple string comparison to match sequence identifiers. We have also started making analysis sets available for the Genome Reference Consortium's human and mouse assemblies (GRCh38 and GRCm38) that are suitable for use with sequence read alignment pipelines. These analysis sets are provided both as FASTA and as index files for BWA, Bowtie and Samtools.

Finally, we have enhanced the search functionality of the NCBI Assembly Resource (www.ncbi.nlm.nih.gov/assembly/) so as to make it easier to find genome assemblies of interest. We also added a link from the Assembly page that provides access to the relevant FTP directory for data downloads and a link to a BLAST web page preconfigured to search against the genomic sequences in the assembly. In addition, we have enriched the Assembly details page with more assembly meta-data and statistics that help to differentiate between the multiple genome assemblies for a particular species.

JBAM QUALITY SCORE COMPRESSION

James Knight

Yale University, Genetics, New Haven, CT

One of the current problems in next-generation sequencing is the size of the BAM files generated from whole-genome sequencing, which can be around 100 GB per sample. One main difficulty in compressing the BAM file information is the per-base quality scores generated for each read. Recently, a lossy "binning" method was introduced, which encodes the 40 possible score values as 8 bins, akin to how the GIF encoding bins the full spectrum of colors down to 256 colors. The JBAM encoding is a method akin to JPEG encoding, using the full 40 quality score values but adjusting closely related neighboring scores to make the scores more compressible. This reduces the space required to store the quality scores, but allows more bases to retain a score closer to their original quality score. The method and the effect on quality scores and on variant calling will be described. This method and its source code is freely available as open source for both academic and commercial use.

CANU: A NEW SINGLE-MOLECULE SEQUENCE ASSEMBLER FOR GENOMES LARGE AND SMALL

Sergey Koren, Brian Walenz, Adam M Phillippy

National Institutes of Health, National Human Genome Research Institute, Bethesda, MD

Single-molecule sequencing is now routinely used to assemble complete, high-quality microbial genomes, but these assembly methods have not scaled well to large genomes. To address this problem, we previously introduced the MinHash Alignment Process (MHAP) for overlapping single-molecule reads using probabilistic, locality-sensitive hashing. Integrating MHAP with Celera Assembler (CA) has enabled reference-grade assemblies of model organisms, revealing novel heterochromatic sequences and filling low-complexity gap sequences in the GRCh38 human reference genome. More recently, in collaboration with the GRC and GIAB consortiums, we have *de novo* assembled five human genomes from PacBio single-molecule, real-time sequencing. In some cases, the resulting assemblies fully resolve chromosome arms and improve on the GRCh38 reference in several metrics.

We have built on this work, creating a successor to CA, named Canu, that is specifically optimized for single-molecule sequencing. Canu represents a complete refactorization of CA, shrinking the code base by 60% and rewriting major components to improve usability and reliability. In addition, Canu introduces improvements in speed, lower coverage requirements, support for nanopore reads, and an adaptive TF-IDF k-mer weighting scheme for efficiently assembling repetitive sequences. For microbial genomes, we have demonstrated that Canu can generate complete, gap-free assemblies from a single Oxford Nanopore MinION flowcell, effectively utilizing both 1D and 2D reads. For large eukaryotes, Canu can generate reference-quality, near-complete assemblies of entire chromosomes. A beta version of Canu is available under a GPL license at <https://github.com/marbl/canu>.

PHYLOGENY OF THE SPIDER MITE SUB-FAMILY
TETRANYCHINAE (ACARI: TETRANYCHIDAE) FROM JAPAN
RECONSTRUCTED BY THEIR TRANSCRIPTOMES

Toshinori Kozaki¹, Tomoko Matsuda*², Kazuo Ishii³, Tetsuo Gotoh⁴

¹Tokyo University of Agriculture and Technology, Laboratory of Genome Science, Fuchu Tokyo, Japan, ²Ibaraki University, Laboratory of Applied Entomology & Zoology, Ami Ibaraki, Japan, ³Tokyo University of Agriculture and Technology, Laboratory of Genome Science, Fuchu Tokyo, Japan, ⁴Ibaraki University, Laboratory of Applied Entomology & Zoology, Ami Ibaraki, Japan

Phylogenetic analysis requires the orthologous genes of multiple species but this analysis usually lacks adequate amount of the sequences especially when the species are over the several taxa. Phylogeny of the sub-family Tetranychinae has been evaluated by COI, ITT2, 18S and 28S rRNAs, but they lacked the accuracy because little information was available from these genes.

In this study, we performed RNA-Seq of the 52 spider mite species that belonged to the family Tetranychidae (4 species of the 2 tribes of the sub-family Bryobiinae and 48 species of the 2 tribes of the Tetranychinae), and the 165 to 247 protein-coding sequences were arbitrarily selected from their transcriptomes to increase the information of the phylogenetic analysis. Several phylogenetic trees were constructed with these sequences to evaluate the phylogeny of the 52 species by Maximum likelihood (ML). Two tribes of the sub-family Bryobiinae, Bryobiini and Petrobiini, were monophyletic and made an out-group to the 48 species of the Tetranychinae. The sub-family Tetranychinae comprised the monophyletic tribe of Eurytetranychini and the polyphyletic tribe of Tetranychini, though the Eurytetranychini did not make an out-group to the Tetranychini but was embedded in the clades of the Tetranychini. At the genus level, 4 genera *Oligonychus*, *Tetranychus*, *Schizotetranychus* and *Eotetranychus* appeared to be polyphyletic and estranged from the topology of the current morphology-based taxonomy. The taxonomy of the sub-family Tetranychinae, therefore, needs to be further confirmed according to both the morphological traits and the molecular facts of this study.

BRINGING GENOMIC DATA INTO FOCUS FOR STUDYING COMPLEX DISEASES IN SPECIFIC BIOLOGICAL CONTEXTS

Arjun Krishnan*¹, Ran Zhang*², Victoria Yao^{1,3}, Chandra Theesfeld¹, Aaron Wong^{3,4}, Alicja Tadych¹, Alan Packer⁴, Alex Lash⁴, Olga G Troyanskaya^{1,3,4}

¹Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, ²Department of Molecular Biology, Princeton University, Princeton, NJ, ³Department of Computer Science, Princeton University, Princeton, NJ, ⁴Simons Center for Data Analysis, Simons Foundation, New York, NY

A big challenge in genomics is the characterization of the genetic and functional dysregulation in complex diseases. Addressing this problem requires systematic computational approaches that can harness the explosion of genomic data, while simultaneously bringing ever-finer biological contexts into focus e.g. tissue, cell-type, sex and age. Towards this goal, we recently developed a Bayesian framework that integrates thousands of gene-expression, protein-interaction and regulatory-sequence datasets to predict tissue-specific functional relationships between genes in each of 144 specific human cell-types and tissues.

Here, using autism spectrum disorder (ASD) as an example, we demonstrate how tissue-specific networks provide a valuable apparatus for generating hypotheses about the molecular basis of human diseases. ASD has a strong genetic basis that remains poorly characterized by sequencing and quantitative genetics studies. Using an evidence-weighted machine learning approach that utilizes the human brain-specific functional gene network, we present here the first genome-wide prediction of autism-associated genes. These predictions are validated using an independent case-control sequencing study of about 2,500 families. Leveraging these genome-wide predictions and the brain-specific network, we demonstrate that the large set of ASD genes – including a host of novel candidates – converges on a smaller number of key cellular pathways and specific early developmental stages of the brain.

Manifesting in early development and being five times more common among boys than among girls, ASD is also an exemplar of diseases whose incidence or severity varies dramatically across the human lifespan and between the sexes. Therefore, our next goal lies in expanding our genomics toolkit to address age- and sex-specificity in addition to tissue/cell-type-specificity. Preliminary results that we will present here demonstrate the promise of our new approaches towards this goal.

* These authors contributed equally.

KOLLECTOR: TRANSCRIPT-GUIDED TARGETED ASSEMBLY OF GENES

Erdi Kucuk^{1,2}

¹ British Columbia Cancer Agency, Michael Smith Genome Sciences Center, Vancouver, Canada, ²University of British Columbia, Graduate Program in Bioinformatics, Vancouver, Canada, ³University of British Columbia, Medical Genetics, Vancouver, Canada

In the last decade, the cost of sequencing has decreased exponentially and the associated, prodigious increase in sequence data is creating computational bottlenecks for processing and analysis. Despite our best efforts, the assembly problem is not yet solved. For some applications, a targeted assembly may be a judicious choice to perform local assembly of sequences of interest. Early targeted assemblers (TASR, MAPAssembler, TRAM) focused on re-assembling de novo, relied on subset of reads sharing k-mers with those sequences of interest, but these methods typically do not scale well and are limiting in their use of sequence information. We have developed Kollector, an alignment-free targeted assembly pipeline that uses thousands of transcript sequences to inform the localized assembly of corresponding gene loci.

The main component of Kollector pipeline is BioBloom Tools, a Bloom filter implementation designed for sequence categorization in a time- and memory-efficient manner. It supports Progressive Bloom filters, a novel Bloom filter-based data structure that can expand greedily and is specifically developed for Kollector. In the Kollector pipeline, a Progressive Bloom filter is initially populated with transcript sequences. Then the genomic data is scanned for reads with a user-defined amount of sequence overlap to the filter, which are added to the filter and used for the accumulation of subsequent reads. This goes on until either the Progressive Bloom filter reaches a predetermined number of elements or all the genomic data is processed. At the end, the Progressive Bloom filter contains intronic regions in addition to the initial seed of transcriptomic sequences. This expanded filter is then used by BBT to recruit genomic reads based on their sequence identity, which are subsequently assembled by ABySS, a well-established de novo genome assembly tool.

The memory-efficient nature of Bloom filters, compared to the methods of alignment used in other tools, allows Kollector to handle large datasets. In order to improve assembly performance in complex genomes, Kollector can take a Bloom filter of repeat sequences as an additional input and use it to tag repeats while extending the progressive Bloom filter. These sequences that are marked as repeats are not used for the expansion of the filter, thus preventing the recruitment of off-target regions. These features make Kollector especially useful for researchers working on non-model eukaryotic organisms.

DOLPHIN: LARGE-SCALE SEQUENCING ANALYSIS PLATFORM

Alper Kucukural, Nicholas Merowsky, Alastair Firth, Manuel Garber

UMass Medical School, Bioinformatics Core, Worcester, MA

High throughput sequencing methods have become very accessible. As sequencing becomes cheaper, simpler and faster, experiments increase in complexity and include many conditions and replicates. Laboratories are able to generate dozens of samples per day and hundreds of libraries every month. The bottleneck becomes the processing and analysis of this ever increasing stream of data. Sequence data processing usually involves multiple programs to perform analysis, e.g. read alignment, peak calling, genome or transcript assembly and quantification.

Existing programs are not designed to process data from “end to end” and take raw input to usable results; instead they are designed and optimized for specific steps in the process. Approaches such as Galaxy, GenePattern and GeneProf attempt to solve this problem by allowing users to build “pipelines” that string specialized programs into end-to-end processes that take raw data into a form that is suitable for analysis. Current solutions were designed when sequencing throughput was lower and users had only a handful of samples to process. As a result they handle a single sample at a time and make no effort to keep experimental details (i.e. metadata). Consequently, they are not well suited to handle the large datasets that are now commonplace.

To address these issues we have created Dolphin, a parallel platform designed to process raw sequence data with the specific goal of handling large datasets. Dolphin keeps metadata information about the experimental conditions and provides an integrated processing and analysis platform. It allows users with limited bioinformatics experience to analyze large numbers of samples on High Performance Computing (HPC) systems through a user-friendly web interface. The UI allows searching, viewing metadata and controlling pipeline (re)execution. Visualization modules show quality results and allow sample comparisons with various plots. Dolphin can backup the files to cloud based storage such as Amazon S3 for easy data sharing, and all files can be uploaded to NCBI Geo and the ENCODE project upon publication.

MULTIMEDIA ANNOTATION USING THE ICLIKVAL BROWSER EXTENSION

Naveen Kumar, Todd D Taylor

Laboratory for Integrated Bioinformatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

The emergence of browser-based web and mobile applications has fundamentally changed the way we generate and consume content. Researchers are producing and consuming more online content than ever in various forms, such as documents, still images, animations, videos, software codes, and many more. There is innumerable information hidden in multimedia content, though it is often not discoverable due to the lack of relevant annotations, such as descriptive keywords.

iCLiKVAL (Interactive Crowdsourced Literature Key-Value Annotation Library) is a web-based application (<http://iclikval.riken.jp>) that uses the power of crowdsourcing to collect annotation information for all scientific literature and media found online. To facilitate the collection of the media and annotation information, we have developed the iCLiKVAL browser extension. The extension is an easy-to-use open-source tool, which uses the iCLiKVAL API to save free-form annotation as “key-value” pairs with an optional “relationship” between them. The idea is to map the online media to a unique URI (Uniform Resource Identifier) and then assign semantic value to it to make the information easier to find and to allow for much richer data searches.

The browser extension facilitates users to bookmark the content or to mark it for later review. It can even be used in offline mode and the data will be automatically synchronized once the user is back online. To use this browser extension users need to be registered with the iCLiKVAL web application. Currently, we are developing this browser extension for Google Chrome and later it will be available for other popular cross-platform browsers.

THE DOE SYSTEMS BIOLOGY KNOWLEDGEBASE: A SYSTEM FOR COLLABORATIVE AND REPRODUCIBLE INFERENCE AND MODELING OF BIOLOGICAL FUNCTION

Vivek Kumar¹, Sunita Kumari¹, James Thomason¹, Mike Schatz¹, Doreen Ware^{1,2}, Sergei Maslov³, Robert W Cottingham⁴, Rick Stevens⁵, Adam Arkin⁶

¹Cold Spring Harbor Laboratory (CSHL), Bioinformatics, Cold Spring Harbor, NY, ²USDA ARS NEA, Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY, ³Brookhaven National Laboratory (BNL), Biological, Environmental & Climate Sciences, Upton, NY, ⁴Oak Ridge National Laboratory (ORNL), Computational Biology and Bioinformatics, Oak Ridge, TN, ⁵Argonne National Laboratory (ANL), Computing, Environmental Science and Life Sciences Research, Argonne, IL, ⁶Lawrence Berkeley National Laboratory (LBNL), Physical Biosciences, Berkeley, CA

The U.S. Department of Energy Systems Biology Knowledgebase (KBase, <http://kbase.us>) aims to provide a computational environment to meet the key challenges of systems biology - predicting and ultimately designing biological function. KBase distinguishes itself as a knowledgebase that supports the "sharing and integration of biological data and any related analysis, modeling and simulation" and not simply a database or a workbench that serves data or canned analyses. KBase integrates commonly used core tools, reference and experimental data, and overlays them with new capabilities for visualization, exploration and predictive analysis designed to accelerate our understanding of microbes, plants and their communities.

The KBase hardware infrastructure is a distributed, private cloud with logical groupings of virtual machines that serve production services and workflows as well as active development. KBase also leverages access to DOE high-performance computing (HPC) resources to support certain large-scale workflows. The production infrastructure is currently distributed between two primary sites at LBNL/National Energy Research Scientific Computing Center (NERSC) and ANL/Magellan to provide redundancy and failover.

KBase has an integrated data model that combines private and public data to enable comparative functional modeling of genes, organisms, and their communities and that supports meta-analysis of both reference data and results from the user community. Using a "plug-in" architecture, it aims to enable external developers to add new data types and tools that operate on KBase data types, leading to easy distribution, comparative tool analysis, and access to enterprise-class computing.

KBase's iPython-based, socially-aware user interface supports a persistent and provenanced environment enabling experimental and computational biologists to work together to share and publish data, approaches, workflows, and thoughts leading to transparent and reproducible scientific results that give credit where credit is due. KBase is supported by the U.S. Department of Energy under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725 and DE-AC02-98CH10886.

A COMPARATIVE ANALYSIS OF NETWORK MUTATION BURDENS ACROSS 21 TUMOR TYPES PREDICTS NEW CANDIDATE CANCER GENES IN THE TAIL OF THE MUTATION DISTRIBUTION OF EXISTING CANCER GENOMES

Heiko Horn^{1,2}, Michael Lawrence², Jessica Hu^{1,2,3}, Elizabeth Worstell^{1,2}, Nina Ilic^{2,4}, Yashaswi Shresta², Eejung Kim^{2,4}, Atanas Kamburov^{1,2}, Alireza Kashani^{1,2}, William Hahn^{2,4}, Jesse Boehm², Gad Getz^{1,2}, Kasper Lage^{1,2}

¹Massachusetts General Hospital, Boston, MA, ²Broad Institute, Cambridge, MA, ³University of Copenhagen, Copenhagen, Denmark, ⁴Dana-Farber Cancer Institute, Boston, MA

Heterogeneity across cancer makes it difficult to find driver genes with intermediate (2-20%) and low frequency (<2%) mutations, and we are potentially missing entire classes of networks (or pathways) of biological and therapeutic value. Here, we quantify the extent to which cancer genes across 21 tumor types have an increased burden of mutations in their immediate gene network derived from functional genomics data. We formalize a classifier that accurately calculates the significance level of a gene's network mutation burden (NMB) and show it can accurately predict known cancer genes and recently proposed driver genes in the majority of tested tumours. Our approach predicts 62 putative cancer genes, including 35 with clear connection to cancer and 27 novel genes, which may point to new cancer biology. NMB identifies proportionally more (4x) low-frequency mutated genes as putative cancer genes than gene-based tests (such as the MutSig suite of tools), and provides molecular clues in patients without established driver mutations. Our quantitative and comparative analysis of pan-cancer networks across 21 tumour types gives new insights into the biological and genetic architecture of cancers and illustrate a powerful, scalable, and cost-efficient computational approach that complements gene-based tests to augment discovery from existing cancer genomes. As more tumors are sequenced in the future our approach should become increasingly powerful.

LARGE-SCALE PREDICTION OF PATHWAYS FROM GWAS AND EXOME-SEQUENCING PROJECTS BY A SYSTEMATIC ANALYSIS OF DIFFERENTIAL PATHWAY ARCHITECTURES IN DIVERSE FUNCTIONAL GENOMIC NETWORKS

John Mercer^{1,2}, Joseph Rosenbluh^{1,3}, Arthur Liberzon¹, Dawn Thompson¹, Thomas Eisenhaure^{1,2}, Steve Carr¹, Jake Jaff¹, Jesse Boehm¹, Aviad Tsherniak¹, Aravind Subramanian¹, Sarah Calvo^{1,2}, Taibo Li¹, Ted Liefeld¹, Bang Wong¹, Jill Mesirov^{1,4}, Nir Hacohen^{1,2}, Aviv Regev^{1,2,5}, Kasper Lage^{1,2}

¹Broad Institute, Cambridge, MA, ²Massachusetts General Hospital, Boston, MA, ³Dana-Farber Cancer Institute, Boston, MA, ⁴University of California, San Diego, San Diego, CA, ⁵Howard Hughes Medical Institute, Chevy Chase, MD

High-throughput technologies in genomics, genetics, epigenetics, transcriptomics, and proteomics have led to the generation of heterogeneous biological networks through which genes are connected if they are functionally correlated in any of the aforementioned data types. These networks share global design features by being scale-free, small world, and modular and have the potential to catalyze genomic interpretation by providing unexpected and systematic insight into the functional wiring of genes. However, it remains challenging to decipher how biological pathways are organized within and between these highly complex networks. We describe a statistical method based on machine learning to test 18 topological metrics across 1,592 pathways (including traditionally defined pathways as well as functional molecular signatures) in five gene networks of i) correlated mRNA expression, ii) phylogenetic patterns, iii) cancer genetic dependencies, iv) cell perturbation profiles, and vi) protein-protein interactions. We show that pathway architectures diverge significantly between networks illustrating that despite similar global design pathways are differentially organized in heterogeneous networks. We provide a web platform (GeNets) that can learn and exploit network-specific pathway architectures to predict unexpected functional relationships between genes and implement it to functionally interpret genetic variants from hundreds of GWAS' and exome sequencing projects. Overall, we provide a computational framework that is a scalable, general, and cost-efficient resource to functionally interpret large genomic data sets using biological networks. GeNets allows users to compare, visualize, and share genome-scale networks through a standardized statistical and visual framework and all results are made available to the genetics community through this resource.

IDENTIFICATION OF THE GENETIC BASIS UNDERLYING
ALTERNATIVE REPRODUCTIVE STRATEGIES IN THE RUFF
(*PHILOMACHUS PUGNAX*)

Sangeet Lamichhane*¹, Guangyi Fan*², Fredrik Widemo*³, Ulrika Gunnarsson¹, Doreen S Thalmann^{4,5}, Marc Höppner^{1,6}, Susanne Kerje¹, Ulla Gustafson⁴, BGI sequencing team², Jacob Höglund⁷, Xin Liu², Leif Andersson^{1,4,8}

¹Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, ²BGI-Shenzhen, Beishan Industrial Zone, Shenzhen, China, ³Department of Wildlife, Fish & Environmental Studies, Swedish University of Agricultural Sciences, Umeå, Sweden, ⁴Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden, ⁵INRA/AgroParisTech, Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France, ⁶Bioinformatics Infrastructure for Life Sciences, Uppsala University, Uppsala, Sweden, ⁷Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden, ⁸Department of Veterinary Integrative Biosciences, Texas A&M University, Texas, TX

The ruff is a medium-sized wading bird that breeds in marshes and wet meadows across the Palearctic zone. The males develop distinctive ornamental feathers (ruff) around their neck during breeding season and show one of the most remarkable mating systems in the animal kingdom. It has three strikingly different male morphs (Independents, Satellites and Faeders) that differ in behavior, plumage colour and body size. Independent males show a spectacular diversity in colour of ruff/head tufts and defend territories at leks. Satellites usually show white ruff/head tufts, do not defend territories and display submissive behavior at leks. Faeder is a rare (< 1%) third morph, mimicking females by its smaller size and female-like plumage. The adaptive evolution in these three male morphs allows them to display a unique courtship behavior and thereby mate with females at leks. Independents attract females by their strongly coloured plumage, elaborate performance display and high degree of aggression, Satellites allow Independents to dominate them, in return getting proximity to females visiting the territories occupied by Independents and Faeders being a female-mimic, get uninterrupted access to mating territories in disguise and attempt to mate females. In this study, we first constructed a de-novo genome assembly of an Independent male and then carried out whole genome re-sequencing of 15 additional Independents, 9 Satellites and 1 Faeder. Using this sequence data, we have identified the genetic basis for the three male morphs and also have proposed a scenario of adaptive changes within this locus that led to the evolution of the spectacular mating system in these birds. We believe, the results of this study will be a textbook example for the evolution of alternative mating strategies in animals.

* equal contribution

THE RESURGENCE OF REFERENCE QUALITY GENOME

Hayan Lee^{1,2}, James Gurtowski¹, Shinjae Yoo³, Maria Nattestad^{1,5}, Shoshana Marcus⁴, Sara Goodwin¹, Richard W McCombie¹, Michael C Schatz^{1,2}

¹Cold Spring Harbor Laboratory, The Simons Center for Quantitative Biology, Cold Spring Harbor, NY, ²Stony Brook University, Department of Computer Science, Stony Brook, NY, ³Brookhaven National Laboratory, Computational Science Center, Upton, NY, ⁴City University of New York, Department of Mathematics and Computer Science, Kingsborough Community College, Brooklyn, NY, ⁵ Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY

Several new 3rd generation long-range DNA sequencing and mapping technologies have recently become available that are starting to create a resurgence in genome sequence quality. Unlike their 2nd generation, short-read counterparts that can resolve a few hundred or a few thousand base-pairs, the new technologies can routinely sequence 10,000 bp reads or map across 100,000 bp molecules. The substantially greater lengths are being used to enhance a number of important problems in genomics and medicine, including de novo genome assembly, structural variation detection, and haplotype phasing.

Here we discuss the capabilities of the latest technologies, and show how they will improve the “3Cs of Genome Assembly”: the contiguity, completeness, and correctness. We derive this analysis from (1) a meta-analysis of the currently available 3rd generation genome assemblies, (2) a retrospective analysis of the evolution of the reference human genome, and (3) extensive simulations with dozens of species across the tree of life.

We also propose a model using support vector regression (SVR) that predicts genome assembly performance using four features: read lengths(L) and coverage values(C) that can be used for evaluating potential technologies along with genome size(G) and repeats(R) that present species specific characteristics. The proposed model significantly improves genome assembly performance prediction by adopting data-driven approach and addressing limitations of the previous hypothesis-driven methodology.

Overall, we anticipate these technologies unlock the genomic “dark matter”, and provide many new insights into evolution, agriculture, and human diseases.

DATA-DRIVEN CHARACTERIZATION OF HUMAN COMPLEX DISEASES

Young-suk Lee^{1,2}, Arjun Krishnan², Olga Troyanskaya^{1,2,3}

¹Princeton University, Computer Science, Princeton, NJ, ²Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, ³Simons Foundation, Simons Center for Data Analysis, New York, NY

There's a surge of genome-wide data in public repositories with much potential for large-scale genome informatics. Gene expression profiling is often used to capture genome-wide snapshots of human disease states. Here we developed a unified computational framework for the characterization of distinctive functional signals in human complex diseases. Leveraging thousands of disease-specific gene expression profiles and known relationships between diseases, our approach identifies molecular-level characteristics unique to each complex disease from both the functional and anatomical perspectives. Our approach, termed URSAHD, is data-driven, and thus not susceptible to literature bias toward specific genes/research areas and provides a data-driven perspective for studying genetically uncharacterized complex diseases, including rare diseases. URSAHD can be used to distinguish between closely-related diseases, identify discerning genes and processes, and associate rare-diseases to the nearest well-studied counterparts for effective drug-repositioning. We find that the most distinctive genes identified by our method are significantly under-studied in the biomedical literature, demonstrating that many key biological processes underlying human pathophysiology are in fact in critical need of further investigation.

FORWARD: A BIOINFORMATICS TOOL TO MANAGE, EXECUTE AND EXPLORE PHENOMIC STUDIES

Marc-André Legault^{1,2}, Louis-Philippe Lemieux Perreault¹, Jean-Claude Tardif^{1,2}, Marie-Pierre Dubé^{1,2}

¹Montreal Heart Institute, Research Centre, Montreal, Canada, ²Université de Montréal, Faculty of Medicine, Montreal, Canada

The phenomic design in genetics is gaining popularity in studies aiming to elucidate complex trait pleiotropy and to evaluate the clinical consequences of mutations. From a bioinformatics perspective, this study design requires particular considerations to maximize both computational efficiency and statistical validity of the experiment, prompting for methodological development.

Here, we describe a new tool called *Forward*, to conduct phenomic experiments in observational cohorts. This tool addresses tedious computational tasks as well as data management and reporting. Mechanisms are included to ensure quality and reproducibility such as the use of human readable experiment descriptions and the archival of analyzed variants, outcomes and results in a relational database. An application programming interface allows further development to expand on the supported data formats and statistical tests. Early versions of this tool support common imputation and microarray data formats (binary *PLINK* and *IMPUTE 2*) and statistical testing for common variants (linear and logistic regression). Optionally, users can use the built-in interactive reporting interface to inspect summary statistics (prevalence, number of missing samples, correlation between outcomes etc.), conduct statistical quality control and interpret the results. A sample report describing a phenomic analysis of the *FTO* gene can be found online (URL: <http://www.statgen.org/forward/fto>). This gene is well known for its association with body mass index and type II diabetes as well as other reported pleiotropic associations. For this analysis, we used the Montreal Heart Institute biobank, an ongoing prospective epidemiological cohort with more than 12,000 individuals with available genetic and phenotypic data. Phenotyping was done using an extensive health questionnaire administered by a research nurse and includes detailed information on cardiovascular health and familial history as well as multiple other demographic and clinical features. *Forward* is a well tested, flexible and open source tool to conduct high throughput phenomic studies.

GENIPE - A PYTHON MODULE TO PERFORM GENOME-WIDE IMPUTATION ANALYSIS

Louis-Philippe Lemieux Perreault¹, Marc-André Legault^{1,2}, Marie-Pierre Dubé^{1,2}

¹Beaulieu-Saucier Pharmacogenomics Centre at the Montreal Heart Institute, Research centre, Montreal, Canada, ²Université de Montréal, Faculty of Medicine, Montreal, Canada

Genotype imputation is now commonly performed following genome-wide genotyping experiments. Imputation increases the density of analyzed data points in the dataset, enabling fine-mapping of either a specific region or across the genome. However, the process of imputation using the most recent publicly available reference datasets can require considerable computation power and necessitates the management of hundreds of large intermediate files. Furthermore, the scarcity of statistical tools available to analyze imputed genotypes (dosage data) narrows the study design possibilities (such as survival analysis and repeated measurements for longitudinal study).

To address the problem of data management and to broaden the scope of statistical tools with imputed data, we have developed *genipe* (**GEN**ome-wide **Imputation Pipeline**), a complete genome-wide imputation pipeline written in Python3. *genipe* uses three of the most commonly used genetics software packages (*PLINK*, *SHAPEIT* and *IMPUTE2*) as well as python modules for additional statistical modeling. *genipe* enables automatic reporting, imputed data indexing and management, and performs a suite of statistical models. Statistical models include linear, logistic or survival regressions, linear mixed effects, and SKAT analysis. *genipe* can be executed on a workstation, or for more efficiency, on a computing server using a batch-queuing system (via a distributed resource management application API). Multiple tests were performed to validate the quality of the modules and results. A set of selected imputed loci can be created using the imputation information value, the imputation probability and completion rate according to user defined thresholds. A report is automatically generated to easily assess the quality of analyses performed and their characteristics (*e.g.* execution time). The report is compiled into a PDF document using *LaTeX*.

Full documentation of the module and tutorials are available at: <http://pgxcentre.github.io/genipe/>. Source code and issue management are available on GitHub (<https://github.com/pgxcentre/genipe>).

FASTQDEMULTIPLEX – A FLEXIBLE DEMULTIPLEXING TOOL FOR ILLUMINA READS

Florian Lenz

University Medical Center Tuebingen, Institute of Medical Genetics,
Tuebingen, Germany

Illumina sequencing systems offer the possibility to sequence several samples within one run. To discriminate between reads of different samples on the same lane, these are typically marked with barcodes. Barcodes are short sequence tags that are read in one or two separate reads, i.e. they are not part of the forward/reverse read of the actual base sequence.

While Illumina's bcl2fastq tool includes the option to assign reads with different barcodes to different FASTQ files, it lacks flexibility: (1) It does not support demultiplexing of reads with barcodes of different length in one pass. (2) It does not support demultiplexing reads with one/two barcodes in one pass. (3) While it includes an option to tolerate deviations from the given barcode sequences, execution is aborted when a barcode clash is found (i. e. a barcode sequence matches the barcode of two samples when considering mismatches). In practice, it would be preferable to continue demultiplexing after discarding reads that could not be assigned to one sample unambiguously.

We present FastqDemultiplex, a novel tool which allows to process reads with different number and lengths of barcodes in one pass. Reads with ambiguous barcodes are put into a separate output file. Furthermore, our tool calculates a lot of interesting statistics about the fraction of reads assigned to each sample and about the barcode sequences.

The tool is publicly available on GitHub as part of the 'ngs-bits' project (<https://github.com/marc-sturm/ngs-bits>).

SPEEDING UP LONG-READ ASSEMBLY BY REDUCING ALIGNMENTS OVERLAP DUE TO REPEATS

Shoudan Liang¹, Paul Peluso¹, Yingping Jiao², Doreen Ware², David Rank¹, Chen-Shan J Chin¹

¹Pacific Biosciences, Menlo Park, CA, ²Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

The state-of-art assembly of long reads requires an all-against-all alignment, which is the most computationally expensive of the assembly steps. For complex genomes, the alignments due to repeats dominate the overlap calculation. The computation can be speed up by up to hundred folds if alignments of repeats can be avoided. One effective way to reduce the alignment of short repeats is by lengthening the minimum allowed overlap in the alignment. This saves computations because the length of overlap can be discerned in seeding step, which is only a small fraction of the alignment computation. However, increasing minimum overlap reduces effective coverage. In order to optimize minimum overlap and other alignment parameters, we derived a formula relating total sequence alignment to minimum overlap for an idealized genome of same size that has no repeats. When reads are randomly sampled from the genome, this curve depends only on the distribution of the read length. We show that by performing alignment using about 1% of data, alignment parameters can be optimized. Examining the genome of human, maize and fungal, we show that a large portion of the alignment was on a few very high copy number reads. The source of these high copy number reads could be chloroplasts, mitochondria, centromeric regions and viruses. We devised a method of removing uniformly-high-copy-number reads.

POSEIDON: A HIGHLY SENSITIVE AND EFFICIENT TAXONOMY CLASSIFIER

EunCheon Lim

Max-Planck Institute for Developmental Biology, Department of Molecular Biology, Tuebingen, Germany

The most abundant forms of life on Earth, the microbes, dwell in human body and their roles in disease development have been studied. Infections by SARS, MERS coronavirus and more dreadful Ebola virus have shown high fatality rates and epidemic potentials since the outbreaks [1]. A fast isolation of virus-specific genomic contents from sputum or blood samples and an accurate detection of species within would enable a timely cope with the spreads of infections and further help to design vaccines.

Metagenomics, a study on a mixture of microbiomes, has some primary questions: the quantitative estimation of species in a sample, and the isolation of target species in exclusion of background genomes. I introduce a highly sensitive and efficient taxonomy classifier, Poseidon, and present the results compared with recently developed algorithms, Kraken [2] and CLARK [3]. They associate a set of k-mers, sequences of length k, to the taxonomy identifiers. Poseidon builds a population index on top of the FM-index, a compressed self-index [4], where the genomes are stored in a compact form and classifies reads by backtracking with variable-length k-mers.

The evaluation has been performed on simulated datasets of 8,294 species by Mason [5] and ART [6]. In all evaluations, Poseidon is the only algorithm keeping a high species-level sensitivity, and achieves the highest genus-level accuracy. In a clade exclusion experiment, the accuracy is measured for *Proteus vulgaris* while the database only comprises *Proteus mirabilis*, and *Proteus penneri*. Poseidon obtains *Proteus*-genus sensitivity of 41.1, which is around 16 higher than Kraken and CLARK with precision of 99.52. Poseidon is 9.4 times more memory-efficient than Kraken and 6.4 times than CLARK. The assignment speed is measured on 100-bp 6,929,444 reads. The speed of Poseidon and Kraken is similar while CLARK is 1.5 times faster than both algorithms.

Poseidon will improve the quality of further analysis in microbiome-related researches through its high accuracy.

[1] Baize S., et al., Emergence of Zaire Ebola virus disease in Guinea, *N. Engl. J. Med.*, 2014, 371

[2] Wood D.E., and Salzberg S.L., Kraken: ultrafast metagenomic sequence classification using exact alignments, *Genome Biology*, 2014, 15, 3

[3] Ounit R., et al., CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers, *BMC Genomics*, 2015, 16:236

[4] Ferragina P., and Manzini G., Opportunistic Data Structures with Applications. *FOCS 2000*. 2000

[5] Holtgrewe M., Mason. <http://www.seqan.de/projects/mason/>

[6] Huang, W., et al., ART: a next-generation sequencing read simulator, *Bioinformatics*, 2012, 28, 4

TRANSCRIPTOME AND EPIGENOME PROFILING OF HUMAN OLFACTORY MUCOSA STEM CELLS AS A MODEL SYSTEM FOR AUTISM SPECTRUM DISORDERS

Tharvesh Moideen Liyakat Ali¹, Mélanie Makhlouf¹, Lam Son Nguyen², Karine Siquier-Pernet², François Féron³, Bruno Gepner³, Laurence Colleaux², Céline Vallot¹, Claire Rougeulle¹

¹CNRS UMR7216 Epigenetic and Cell Fate, Univ Paris Diderot, Sorbonne Paris Cité, Paris, France, ²INSERM UMR 1163, Laboratory of Molecular and pathophysiological bases of cognitive disorders, Paris Descartes – Sorbonne Paris Cité University, Imagine Institute, Necker-Enfants Malades Hospital, Paris, France, ³Faculté de Médecine Nord, Aix Marseille Université, NICN, CNRS UMR 7259, Marseille, France

Olfactory mucosal stem cells (OMSCs) are an easily accessible part of the nervous tissue and could serve as a model to study the pathological conditions related to the brain and other components of nervous system. We have selected this model to study Autism Spectrum Disorder (ASD) and are probing chromatin and transcriptional changes associated with the disease using next generation sequencing approaches.

We are using total rather than polyA+ RNA for RNA-seq analysis, allowing to investigate the expression of both coding and non-coding RNAs. By ChIP-seq, we are characterizing the epigenomic landscape focusing on histone marks such as H3K4me3 and H3K27me3, which could be found in chromatin of active promoter and repressed genomic regions respectively. We are using unsupervised techniques such as PCA and hierarchical clustering to understand and identify major transcriptomic and epigenomic differences between autistic and normal OMSCs, followed by a differential analysis, based on count matrix for both data types.

In parallel, we aim to understand how OMSCs are related to other tissues and cell lines, by comparing our data to available datasets generated from multiple in vitro and in vivo cell types. As a first step, we are analyzing the publicly gathered data sets for potential confounders such as library type, RNA extraction techniques and source of the data. We found that, while overall there are no batch effects, the RNA extraction method impacts on clustering when only long non-coding RNAs are considered. We will discuss unsupervised and supervised analysis of transcriptomic and epigenomic data.

GENOMSAWIT: A ONE-STOP GENOME INFORMATION PORTAL FOR OIL PALM

Leslie Eng-Ti Low¹, Kuang-Lim Chan¹, Mohd Amin Ab Halim¹, Corey Wischmeyer², Smith W Steven², Rozana Rosli^{1,3}, Norazah Azizi¹, Nik Shazana Nik Mohd Sanusi¹, Nadzirah Amiruddin¹, Jayanthi Nagappan¹, Leslie Cheng-Li Ooi¹, Pek-Lan Chan¹, Ngoot-Chin Ting¹, Michael Hogan², Rajinder Singh¹, Meilina Ong-Abdullah¹, Robert A Martienssen⁴, Ravigadevi Sambanthamurthi¹

¹Malaysian Palm Oil Board, Advanced Biotechnology and Breeding Centre, Kuala Lumpur, Malaysia, ²Orion Genomics, St Louis, MO, ³University of South Wales, Genomics and Computational Biology, Pontypridd, United Kingdom, ⁴Cold Spring Harbor Laboratory, Howard Hughes Medical Institute-Gordon and Betty Moore Foundation, Cold Spring Harbor, NY

The oil palm belongs to the family Palmaceae and the genus *Elaeis*. There are two important species in the genus *Elaeis*, *E. guineensis* and *E. oleifera*. In South-East Asia, *E. guineensis* is the commercial oil palm species that is planted for the production of palm oil and palm kernel oil. The importance of the crop led to the publication of its genome in 2013. With the availability of the genome information and related publications, the Genomsawit portal was developed. The portal allows researchers to access the latest publications and data on the genome, as well as information on the Oil Palm Genome Programme, including latest news and bioinformatics tools. The website also contains an oil palm genome browser (MYPalmViewer). Researchers are able to visualize the *Elaeis guineensis* (EG) genome build, which is the genome assembly obtained by aligning scaffolds to genetic maps (T128 and P2). Other tracks available include the comparative genomics track for the *E. oleifera* genome, predicted genes, oil palm transcripts, genetic markers from oil palm publications, *Arabidopsis thaliana* and *Oryza sativa* genes, SwissProt sequences and oil palm GeneThresher methylation filtered data. The browser is searchable via keyword query or Blast interface. MYPalmViewer can be reached from a link in the Genomsawit portal (<http://genomsawit.mpob.gov.my>) or directly via <http://gbrowse.mpob.gov.my>.

DETERMINING THE HYPOXIC GENE EXPRESSION RESPONSE OF *S. CEREVISIAE* CELLS USING RNA-SEQ AND STATISTICAL ANALYSIS OF TIME-COURSE DATA.

Samuel Maclean¹, Gurmannat Kalra¹, Nasrine Bendjilali², Mark J Hickman¹

¹Rowan University, Biological Sciences, Glassboro, NJ, ²Rowan University, Mathematics, Glassboro, NJ

Hypoxia in tissues occurs in several human diseases such as cancer, stroke, and vascular diseases. Eukaryotic cells employ signaling pathways to sense, transduce and respond to oxygen levels; one response is a large change in gene expression which helps cells cope with hypoxia. *Saccharomyces cerevisiae* cells are an excellent model to study this response because many signaling pathways are shared with humans, signaling genes can be accurately deleted or modified, and yeast growth can be rigorously controlled. In order to understand how the hypoxic gene expression is regulated in yeast, we first sought to characterize the expression response after oxygen is withdrawn. At 0, 5, 10, 30, 60, 120, 180, and 240 minutes of hypoxia, Illumina RNA-sequencing was used to quantify expression of all ~7000 genes in wild-type yeast cells. Then, several time-series analyses were performed on HTSeq read counts over the four-hour time period, using R statistical analysis software. First, by monitoring how variability correlates with the number of read counts, we determined a “floor” below which the read counts for a gene were unreliable. Second, oxygen-regulated genes were expected to exhibit a “smooth” time-dependent response over time in hypoxia, and thus have a high autocorrelation. We found hundreds of genes with high autocorrelation, and this autocorrelation was shown by bootstrapping methods to be statistically significant. This autocorrelation, as well as other time-series analyses including DESeq2, identified almost all of the expected oxygen-regulated genes, as well as previously undiscovered genes. These oxygen-regulated genes were subjected to clustering analysis to group genes according to their time expression profile. As expected, genes that cluster together share many of the same functions and same regulatory transcription factors. Such time-course analysis is critical in determining genes that respond to a stimulus (e.g., stress, growth factor) and in grouping genes to define signaling networks. Additionally, we found that genes regulated by the hypoxic response are distinct from those regulated by the environmental stress response, suggesting that hypoxia is not simply recognized as a stress. Our future work will use similar analysis and gene deletions to define all of the signaling pathways that contribute to the hypoxic response. Understanding these hypoxic signaling pathways will give us a glimpse of how the human pathways function, and how disruption of these pathways contributes to human diseases that involve hypoxia.

MICROBIAL GENOME ASSEMBLY USING SYNTHETIC ERROR-FREE READS

Mohammed-Amin Madoui¹, Stefan Engelen¹, Corinne Cruaud¹, Arnaud Lemainque¹, Patrick Wincker^{1,2,3}, Jean-Marc Aury¹

¹Genoscope-CNS, CEA, Evry, France, ²UMR 8030, Université d'Evry, Evry, France, ³UMR 8030, CNRS, Evry, France

The technology of long-read sequencing offers different alternatives to solve genome assembly problems which cannot be resolved adequately by short-read sequencing. We present a new hybrid approach developed to take advantage of long (MinION or PacBio) and short (Illumina) reads. Our method was able to generate synthetic long reads up to 90kb with no error and that span large repetitive regions. The method was applied to several bacterial and small eukaryotes read sets to generate the error-free synthetic reads that were then used to produce highly contiguous and accurate genome sequences. For bacterial genomes, our method outperformed the existing methods of reads correction and the whole strategy (including the sequencing) enabled release of near perfect genomes in less than three days, even in small facilities.

TARGETED SEQUENCING OF FFPE OVARIAN CANCER TUMOUR SAMPLES ON THE ION PGM PLATFORM

Alison Meynert¹, Michael Churchman², Robb Hollis², Tzyvia Rye², Angie Fawkes³, Lee Murphy³, Colin Semple¹, Charlie Gourley²

¹University of Edinburgh, MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom, ²University of Edinburgh, Edinburgh Cancer Research Centre, MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom, ³University of Edinburgh, Wellcome Trust Clinical Research Facility, Edinburgh, United Kingdom

Extensive archives of formalin-fixed paraffin-embedded (FFPE) tumour samples can be exploited for high throughput sequencing (HTS). In conjunction with clinical research databases on patient outcomes and response to treatment, effectiveness of treatment can be conditioned on genetic mutation status. In particular, for ovarian cancer, mutations in the genes BRCA1 and BRCA2 drive the development of a significant proportion of tumours. To undertake this type of study, it is necessary to learn which FFPE samples are suitable for HTS. We took 121 archival FFPE samples of high grade serous ovarian carcinomas and measured DNA quality using a variety of metrics. 115 samples were then targeted with an Ampliseq panel designed to capture BRCA1 and BRCA2 and sequenced on the Ion PGM platform in two pools, with mean on target sample depth from 1300-40000X. Five sample libraries were sequenced twice, once in each pool. We found that most DNA quality metrics do not correlate well with either each other or with the resulting sequence quality. A further challenge with FFPE samples is the degradation of DNA caused by the formalin-fixation. This process introduces strand breaks, fragmenting the DNA into pieces that may be too small for sequencing. It also induces cytosine deamination, which appears as an overrepresentation of erroneous C>T (G>A) mutations at low allele frequencies in the resulting variant call sets. When comparing the mutation spectrum of bi-allelic single nucleotide variants (SNVs) in our samples to several hundred Cancer Genome Atlas (TCGA) ovarian cancer samples, we observed a major bias in our samples towards both C>T (G>A) and A>G (T>C) mutations. Applying a minimum allele frequency cut-off partially addresses the problem; however, genuine sub-clonal mutations will be lost under this approach. Given the difficulty in computationally determining the truth of low allele frequency variants, sequencing FFPE samples to high mean depths with the aim of elucidating low frequency sub-clonal mutations is likely to fail, and it may be more cost effective to sequence several samples from the same tumour instead, relying on spatial heterogeneity to assist in determining cellular heterogeneity.

AUTOMATED TRANSFER OF WORKFLOWS FROM GALAXY TO YABI AND COMMAND LINE TOOLS

David C Molik, Ying Jin, Molly Hammell

Cold Spring Harbor Lab, Cold Spring Harbor, NY

The web-based bioinformatics platform Galaxy gained great popularity as a tool for allowing access to powerful compute clusters and sophisticated bioinformatics software with user-friendly point-and-click interfaces (Goecks et al., 2010). In contrast, command line tools will always be more efficient for those who are comfortable working within a Unix-like environment. Likewise, biologists who understand the computing environment are more likely to understand what is and is not computationally efficient or possible (Dudley et al., 2009). The obstacle lies in training users with little programming experience to be comfortable with the command line. We have explored alternate frameworks to Galaxy that would allow users to design their workflows in a simplified web browser environment, and then automatically transfer these workflows into pipelines suitable for running at the command line.

Yabi is an alternate web-based bioinformatics platform designed by the Center for Comparative Genomics at Murdoch University (Hunter et al., 2012). It fulfills many of the same operations as Galaxy, and is interoperable with the same tools. Yabi provides a similar user experience, providing a graphical user interface to bioinformatics software that can be executed either locally or remotely. Moreover, because Galaxy and Yabi both use simple configuration files, the tool interfaces designed for Galaxy have the possibility of being automatically transferred to the Yabi format. Additionally, Yabi offers a command line tool, yabish, where users can design their workflows within the context of the web-based graphical interface, and then automatically transfer that workflow into a pipeline suitable for running at the command line. This can be an intermediate step between offering a graphical user interface to the end user and having users do all of their analysis on the command line, while providing the benefits of logging, saved workflows, and remote data. For the benefit of server administrators, Yabi accesses data and submits jobs to computational clusters as the end user and not as the daemonized user; meaning data management is more secure and job submission has more equality.

We implemented Yabi as an analogous software application to Galaxy and provide contrasting benefits of both platforms. Moreover, we present tools that automatically parse and reformat tool configuration files for use in either the Galaxy or Yabi format.

SEARCHING AND EXPLORING GRAMENE'S COMPARATIVE GENOMICS DATASETS ON THE WEB.

Joseph Mulvaney¹, Andrew Olson¹, James Thomason¹, Doreen Ware^{1,2}

¹Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY,

²USDA-ARS, USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Gramene is a unique resource that comprises genomic, pathway, genetic marker and expression databases for economically important species across the plant kingdom. We perform comparative analyses across these datasets and make the data available through a performant integrated web service (<http://data.gramene.org>). Here we present a web application (<http://search.gramene.org>) that uses this service to enable researchers to find sets of genes from Gramene and to visualize and explore the results. The application proactively suggests appropriate terms as you type; shows an interactive distribution of result sets across genomes and species; provides statistics and enriched terms in the result set; and allows users to use details of a result to expand or narrow down a search. Each gene provides links to relevant pathway, genome and expression profiles within Gramene and links to third-party databases.

Development of the Gramene database is supported by an NSF award (IOS-1127112).

COMPARATIVE ANALYSIS OF CHROMATIN STATES AND GENE EXPRESSION PROFILES FOR VARIOUS ENDOTHELIUM CELLS

Ryuichiro Nakato¹, Yuki Katou¹, Toutai Mituyama², Hiroshi Kimura³,
Youichiro Wada⁴, Hiroyuki Aburatani⁴, Katsuhiko Shirahige¹

¹Inst Mol Cell Biosci, Tokyo Univ, Bunkyo-ku, Japan, ²Biotech Res Inst, AIST, Koto-ku, Japan, ³Grad Sch Biosci Biotech, Tokyo Tech, Yokohama, Japan, ⁴Res Cen Adv Sci Tech, Tokyo Univ, Meguro-ku, Japan

Analyzing tissue-specificity of genome-wide chromatin state maps facilitates understanding of the mechanism of epigenetic gene regulation and the critical role of histone modification patterns. We have been generating the ChIP-seq (Chromatin immunoprecipitation and deep sequencing) data of histone modifications (H3K4me3, H3K27ac, H3K27me3, H3K36me3 and H3K9me3) for various types of blood vessel endothelium cells including artery, vein and endocardial cells. Here we report the comparative analysis of histone modification patterns with gene expression profiles. Chromatin state maps generated by ChromHMM from endothelium cells, including five types of cells derived from same donor, were characterized and clustered. As reported in early works, enhancer sites were more tissue-specific than strong promoter sites, and gene ontology analysis showed that enhancer and promoter sites were associated with different gene functions. Next, we attempted to identify the tissue-specific sites of H3K4me3 and H3K27ac, using both quantitative (peak intensity) and binary (peak overlapping ratio) manners. We used MACS for peak-calling and DROMPA for normalization and visualization. We found that the binary manner could make better clustering results when the peak intensity (signal-to-noise ratio) differs among replicates. We identified the candidates of tissue-specific histone modification patterns and associated genes for several types of endothelium cells. We also discuss the method for the quality check. Several samples have a GC-rich read distribution, which could be normalized by the GC normalization of DROMPA. Cross-correlation profiles (CCP) generated by spp is efficient to check the signal-to-noise ratio without peak-calling, whereas several input samples had high RSC (Relative Strand Cross-correlation coefficient) scores but had few peaks.

PROBING TRANSCRIPTIONAL REGULATION IN TUMOR SPECIMENS YIELDS HALLMARKS OF PROSTATE CANCER OUTCOME

Ekaterina Nevedomskaya*^{1,6}, Suzan Stelloo*¹, Henk G van der Poel², Jeroen de Jong³, Geert J van Leenders⁴, Guido Jenster⁵, Lodewyk F Wessels⁶, Andre M Bergman⁷, Wilbert Zwart¹

¹Netherlands Cancer Institute, Division of Molecular Pathology, Amsterdam, Netherlands, ²Netherlands Cancer Institute, Division of Urology, Amsterdam, Netherlands, ³Netherlands Cancer Institute, Division of Pathology, Amsterdam, Netherlands, ⁴Josephine Nefkens Institute, Erasmus Medical Center, Department of Pathology, Rotterdam, Netherlands, ⁵Josephine Nefkens Institute, Erasmus Medical Center, Department of Urology, Rotterdam, Netherlands, ⁶Netherlands Cancer Institute, Division of Molecular Carcinogenesis, Amsterdam, Netherlands, ⁷Netherlands Cancer Institute, Division of Medical Oncology, Amsterdam, Netherlands

*Authors contributed equally

Prostate cancer (PCa) is the most common malignancy in men and one of the leading causes of cancer-related deaths in the Western world. Androgen Receptor (AR) is a crucial player on all stages of PCa, including progressive metastatic disease and resistance to therapy. To understand disease development and find prognostic biomarkers we turned to the level on which Androgen Receptor acts: protein-DNA binding and transcriptional regulation. We developed a comprehensive pipeline that starts from unbiased analysis of chromatin accessibility in tumors, followed by identification of transcription factors involved and continues with profiling genomic action of these factors that facilitates identification of disturbed pathways and potential prognostic signature. The pipeline combines both new genome-wide data, as well as a plethora of public data from in vitro experiments and clinical cohorts. The new data included genome-wide profiling of chromatin accessibility by Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)-seq and AR/DNA interaction analysis by Chromatin Immunoprecipitation (ChIP)-seq at different stages of prostate cancer progression. We identified a distinct Androgen Receptor/chromatin binding profile between primary prostate cancers and tumors with an acquired resistance to therapy. We further integrated this genomic data with transcriptomic datasets: gene expression in perturbed cell lines, clinical gene expression and survival data. As a result we come up with a concise gene signature with strong prognostic potential. The power of the proposed approach lies in the integration of multiple datastreams on transcriptional regulation and functional gene expression, which has a potential to provide biologically relevant prognostic signatures. This innovative pipeline for biomarker discovery can be easily implemented in other fields of oncology.

COMPUTATIONAL ANALYSIS OF TARGET SPECIFICITY OF DOUBLE-STRANDED RNA BINDING PROTEIN STAUFEN

Kun Nie^{1,2}, Quaid D Morris^{1,2}

¹Terrence Donnelly Center for Cellular and Biomolecular Research, Banting and Best Department of Medical Research, Toronto, Canada,

²University of Toronto, Department of Molecular Genetics, Toronto, Canada

RNA binding proteins are one of the key regulators of gene expression. As an essential class of RNA binding proteins, double-stranded RNA binding proteins are involved in various aspects of RNA metabolism. *Drosophila* protein Staufen (STAU) is required in the localization of the mRNA of the morphogen *Bicoid* (*bcd*) to the anterior pole, which drives the formation of the anterior pattern of the *Drosophila* embryo. We have shown previously that *Drosophila* Staufen recognize three types of specific secondary structures in the 3'-UTR of its target transcripts¹. Using MC-Flashfold, a fast version of MC-Fold², we identified a new type I Staufen recognition site (SRS) in *bcd* mRNA 3'-UTR that was previously undetected. This site also corresponds to one of the mRNA conformation required by STAU interaction described in a previous research in vivo from Ferrandon *et al*³. Our new analysis also incorporates non-canonical RNA binding sites which provides more detailed RNA secondary structural information. Thus, our study will provide insights into the binding specificities of double-stranded RNA binding proteins as well as post-transcriptional regulations in *Drosophila*.

1. Laver, J. D. *et al*. Genome-wide analysis of Staufen-associated mRNAs identifies secondary structures that confer target specificity. *Nucleic Acids Res.* 41, 9438–9460 (2013).

2. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452, 51–55 (2008).

3. Ferrandon, D., Koch, I., Westhof, E. & Nüsslein-Volhard, C. RNA-RNA interaction is required for the formation of specific bicoid mRNA 3' UTR-STAU-FEN ribonucleoprotein particles. *EMBO J.* 16, 1751–1758 (1997).

COMPARATIVE GENOME ANALYSIS OF MEMBERS OF THE MAGNAPORTHACEAE FAMILY OF FUNGI

Laura H Okagaki^{1,3}, Joshua K Sailsbry², Alexander W Eyer¹, Titus John¹, Cristiano C Nunes^{1,3}, Ralph A Dean^{1,3}

¹North Carolina State University, Center for Integrated Fungal Research, Raleigh, NC, ²Bayer Crop Science, ---, Research Triangle Park, NC, ³North Carolina State University, Plant Pathology, Raleigh, NC

Magnaporthaceae is a family of ascomycetes that includes three fungi of great economic importance that cause disease in cereal and turf grasses: *Magnaporthe oryzae* (rice blast), *Gaeumannomyces graminis* var. *tritici* (take-all disease), and *Magnaporthe poae* (summer patch disease). Comparative genomics provides a powerful method to rigorously evaluate the genetic and evolutionary basis of structure-functional relationships as well as give insight into the biology and pathogenicity of these fungi. We conducted a genome-scale comparative study across 74 fungal genomes to identify orthologous clusters unique to the three Magnaporthaceae species. Gene Ontology annotation, Interpro protein domain identification, and signal sequence identification were used to identify potential orthologous cluster functions that may shed light on Magnaporthaceae pathogenesis. In addition, we examined the relationship between gene evolution (PAML) and distance to repetitive elements found in the genome. We found that of almost 68,000 clusters, almost 3% are specific to the Magnaporthaceae. Of the Magnaporthaceae specific clusters, transcriptional regulators and enzymes were highly represented. No relationship between diversifying/purifying selection and distance to repetitive elements was observed in orthologous clusters. Our data also shows that *M. poae* and *G. graminis* var. *tritici* were more closely related than either were to *M. oryzae*, sharing a higher number of orthologs, more conserved gene function, and larger syntenic regions. A high proportion of genes were species specific, ranging from 28-36%. In addition, genes unique to all three Magnaporthaceae were enriched for small secreted proteins (less than 250 amino acids).

INTEGRATED WEB SERVICES SUPPORTING SEARCH AND INTERACTIVE ANALYSIS TOOLS AT GRAMENE

Andrew Olson¹, Kapeel Chougule¹, Joseph Mulvaney¹, Justin Preece², James Thomason¹, Doreen Ware^{1,3}

¹Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY,

²Oregon State University, Department of Botany and Plant Pathology, Corvallis, OR, ³USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, USDA-ARS, Ithaca, NY

Gramene (www.gramene.org) is a curated resource for comparative functional genomics in crops and model plant species. Gramene hosts assembly and annotation data for 39 plant genomes of interest to crop scientists and plant geneticists utilizing the Ensembl platform. Genotypes of thousands of accessions are maintained in variation databases. Comparative genomic analyses identify conserved sequences, syntenic blocks, and characterize the evolution of gene families. For pathway information, Gramene has adopted the Reactome data model to host curated rice pathways and their projections in other species via orthology relationships.

Gramene is building integrated data stores from these primary data sources to complement the REST APIs offered by Ensembl and Reactome. Lightweight node.js web servers provide HTTP access to the integrated data and auxiliary structured annotations stored in MongoDB and Solr document collections. Unified access to these services is provided through <http://data.gramene.org>.

The new Gramene search interface (<http://search.gramene.org>) uses our Solr endpoint for advanced querying features. For example, the surround query parser is used to search for genes with similar domain architecture. Complex queries involving structured annotations, such as the interpro domain hierarchy, gene ontology, and NCBI taxonomy are composed with filter queries. Facet counting, used throughout the search interface, calculates the genomic distribution of gene search results across all hosted genomes.

We have documented the REST endpoints with swagger and are actively developing interactive views of the data.

Development of the Gramene database is supported by an NSF award (IOS-1127112).

THE IDENTIFICATION OF THE GENE EXPRESSION SIGNATURES OF TISSUE-SPECIFIC EFFECTS OF PIOGLITAZONE TREATMENTS IN A MURINE MODEL OF TYPE 2 DIABETES

Meeyoung Park¹, Amy Rumora¹, Lucy Hinder¹, Junguk Hur², Felix Eichinger³, Matthias Kretzler³, Eva L Feldman¹

¹University of Michigan, Neurology, Ann Arbor, MI, ²University of North Dakota, Biomedical Sciences, Grand Forks, ND, ³University of Michigan, Internal Medicine, Ann Arbor, MI

Diabetic nephropathy (DN) and diabetic peripheral neuropathy (DPN) are the most common complications of diabetes. Even though pioglitazone is frequently used to treat Type 2 Diabetes (T2D), its effect on diabetic complications at the transcriptomic level is not well established. In this study, we aim to elucidate the underlying mechanisms of differential gene regulation patterns that are associated with pioglitazone treatment in kidney (glomeruli and cortex) and nerve [dorsal root ganglia (DRG) and sciatic nerve (SCN)] tissues from *db/db* mice, a murine model of T2D.

Differential expression analysis between *db/db* and pioglitazone treated *db/db* using RNA-Seq was performed to identify significantly dysregulated genes for all tissues. In addition, we applied Self Organizing Maps, an unbiased clustering method, in order to identify coherent expression patterns for all genes between SCN and glomeruli. Functional enrichment and pathway analyses were performed to find key dysregulated genes and pathways.

Differential expression analysis demonstrated that pioglitazone regulated the expression of a large number of genes in SCN and glomeruli, but not in DRG and cortex. In addition, the differentially expressed genes related to the regulation of inflammatory response and programmed cell death, were reversed in glomeruli; however, they were exacerbated in SCN. The SOM analysis revealed that genes overrepresented in the calcium signaling and mitochondrial dysfunction pathways were reversed in kidney but not in nerve.

Our results identified the tissue-specific gene expression signatures of pioglitazone treatment in T2D. These data suggest potential pharmaceutical targets to tailor different diabetic complications.

EFFECTS OF HORMONAL CHANGES AND CIGARETTE SMOKING ON ORAL MICROBIOME

Purnima Kumar, Binnaz Leblebicioglu, Akshay Paropkari

The Ohio State University, Division of Periodontology, College of Dentistry, Columbus, OH

The oral cavity hosts an open ecosystem. As in every ecosystem, the environment plays an essential role in fashioning its community. Previously, we have shown that smoking affects the oral microbiome. This study attempts to understand the ramification of the two vital environmental events on microbial community - pregnancy and smoking.

A total of 44 subjects with no periodontal disease were chosen and their sub-gingival and plaque samples collected. The samples were uniformly collected from non-pregnant non-smokers (control), non-pregnant smokers, pregnant non-smokers and pregnant smoker groups. 16S gene for each sample was sequenced using 454 sequencing platform, and compared to GreenGenes database.

Phylogenetic classification was conducted with quantitative insights on microbial ecology (*QIIME*) and phylogenetic tools for analysis of species-level taxa (*PhyloToAST*) bioinformatics pipelines. Additionally, statistical and network analyses were performed to comprehend the shifting nature of the oral microbiome.

Linear Discriminant Analysis (LDA) highlighted statistically significant group separation ($P < 0.0001$, MANOVA/Wilks lambda) among all groups. Using the interactive tree of life (iTol), OTU's belonging to genera *Aeromicrobium*, *Aggregatibacter*, *Herbaspirillum*, *Johnsonella*, *Neisseria*, *Porphyromonas* and *Pseudomonas* were significantly different ($P < 0.0001$, Tukey-HSD) among groups.

Significant Pearson correlation ($P < 0.05$) and graph theory were used to highlight genus and species level microbial co-occurrence network for each group. The node (total number of OTU's) to edge (total number of correlations) ratio for control group was observed to be consistent ($M = 0.94$, $SD = 0.14$), for both genus and species level. Comparatively, pregnant non-smokers ($M = 0.67$, $SD = 0.25$) and non-pregnant smokers ($M = 0.76$, $SD = 0.67$) both deviate away from the control group. Intriguingly, pregnant smokers ($M = 1.00$, $SD = 0.68$) are observed to be closer to control group than others. Analyzing the variances of the node-to-edge ratio show that habitat selection is at work in all groups when compared to the control group. Compared to the control group, the multifold change in variances of other groups point to the fact that new microbial co-occurrence networks are formed as a result of competitive ability and niche differences expressed by oral microbiome.

The results point out the dissimilarity between microbial composition based on pregnancy and smoking. Oral microbial ecosystem adjusts to each environmental change in a different way. The effects of hormones and smoking are seen individually, and when combined together, they do not have an additive effect on the oral microbiota.

ENHANCING THE UTILITY AND USABILITY OF GEMINI FOR RARE AND COMMON DISEASE RESEARCH.

Brent S Pedersen, Aaron R Quinlan

University of Utah, Department of Human Genetics, Salt Lake City, UT

Studies of human genetic variation leveraging modern DNA sequencing follow the core steps of sequencing, alignment, artifact removal, and variant identification. At this point, all studies also share a common challenge in segregating the small minority of variants underlying a trait or mechanism from the benign variants. Central to this are the complexity and analytical limitations intrinsic to the resulting Variant Calling Format (VCF) files, as well as the need to integrate diverse genome annotations (e.g., dbSNP, ENCODE, ExAC, UCSC, ClinVar) in order to place variants in context.

To address this problem, we have previously developed GEMINI (GEName MINing), a flexible toolset for exploring all forms of human genetic variation. GEMINI has seen wide adoption, due to its integration of genetic variation (in the VCF format) with a diverse and adaptable set of genome annotations into a unified database that facilitate interpretation and data exploration.

Here, we describe multiple substantial changes to GEMINI that improve its speed, scalability, and flexibility. These include dramatic improvements to query speed for queries based on sample genotypes and phenotypes, as well as increased speed and flexibility for creating GEMINI databases for any species and genome build via our new VCFANNO software. Lastly, we will present comprehensive methods for Mendelian inheritance model testing in multiple and arbitrarily large pedigrees. These improvements increase GEMINI's speed by over an order of magnitude and make GEMINI applicable to a broad range of studies of genetic variation in both rare and common disease contexts.

ACCURATE AND EFFICIENT TRANSCRIPT IDENTIFICATION AND QUANTIFICATION USING RNA-SEQ DATA

Mihaela Pertea¹, Geo M Pertea¹, Steven L Salzberg^{1,2}

¹Johns Hopkins University, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, ²Johns Hopkins University, Departments of Biomedical Engineering, Computer Science, and Biostatistics, Baltimore, MD

Transcriptome assembly and gene expression profiling are key components in a vast range of biological experiments today, playing a central role in unraveling the complexity of cell type, cell differentiation, responses to stress, and myriad other conditions. Although transcript assemblers have been developed previously, most of them perform poorly on real, large-scale RNA-seq data sets, severely limiting their impact.

Our method – StringTie - is the first transcript assembler that uses an optimization technique known as maximum flow in a specially-constructed flow network to determine gene expression levels, and it does this at the same time as it is assembling each splice variant of a gene. It is also the first transcript assembler to incorporate techniques from whole-genome assembly, which has the potential to dramatically improve our ability to resolve alternative splice variants. Our results on both simulated and real data demonstrate that, as compared with other leading transcript assembly programs, StringTie produces more complete and more accurate reconstructions of genes and better estimates of expression levels. StringTie is also very fast, with run times ranging from three to 50 times faster than other methods.

One problem that is usually inevitable for transcriptome assembly is a big memory requirement due to the vast amount of RNA-Seq data that needs to be processed at the same time. In highly expressed transcripts, the bundle of reads that come from the same gene locus in the genome can easily reach over 2 million reads. It can be a challenge to store and analyze all the information needed to process these reads: alignment start and ends, cigar string, strand information, mate pairs etc. Here we present new developments to our method, which reduce the big memory footprint usually associated to transcriptome assembly.

NCBI'S GENETIC VARIATION RESOURCES

Lon D Phan, NCBI Variation Working Group

National Institutes of Health, NLM/NCBI, Bethesda, MD

The Variation Resources site (<http://www.ncbi.nlm.nih.gov/variation/>) at the National Center for Biotechnology Information (NCBI) is a gateway for users to access databases and tools that can be used in the fields of personal genomics, medical genetics, and management of clinical variation data. There are five NCBI databases that archive, analyze, display, and report information about germline and somatic variants and the relationship of these variants to phenotype and clinical significance. dbSNP houses short variations; dbVar houses large scale genomic variants; the database of Genotypes and Phenotypes (dbGaP) houses genotypes and phenotypes associations; ClinVar houses reported relationships between human variation and phenotypes; and the Genetic Testing Registry (GTR) provides a central location for accessing inherited and somatic genetic variations that are being tested for a specific trait or disorder. Each submission record is assigned an accession number in its respective database. Submitted variations at the same genomic location are aggregated and assigned a reference identifier (rs# number in dbSNP or nsv|esv in dbVar). The Variation Resource site also includes tools to explore and expedite the analysis of submitted variant and NCBI annotated information and to integrate available information in variation discovery and clinical laboratories workflow. The Variation Viewer allows users to search, navigate, and view variations in genomic and gene context. The Variation Reporter accepts uploaded VCF files and generates a comprehensive report that includes molecular consequences (e.g. missense/nonsense/affects splicing), allele novel to NCBI, and information from NCBI's databases. The Clinical Remap tool converts uploaded variant calls on NCBI36, GRCh37, and GRCh38 and on gene-specific RefSeqs such as RefSeqGene, RNAs, and proteins. The 1000 Genomes Browser allows users to explore variant calls, genotype calls and supporting sequence read alignments (SRA) produced by the 1000 Genomes project. The Phenotype-Genotype Integrator facilitates access to public genome-wide association results based on queries by phenotype or genomic location. In addition, dbSNP provides a VCF report that can be used for filtering large sets of variant calls for variations that are not known to be disease-related. All NCBI variation databases integrate to each other and with other NCBI resources (e.g. BioProjects, Gene, PubMed, Nucleotide, and Protein) and disseminate accurate information about variation to the scientific community.

MIXTURE MODELS THAT ESTIMATE GENE-EXPRESSION ACTIVATION ON A SINGLE-SAMPLE BASIS FOR ANY EXPRESSION PLATFORM

Stephen R Piccolo¹, Evan Johnson²

¹Brigham Young University, Department of Biology, Provo, UT, ²Boston University School of Medicine, Division of Computational Biomedicine, Boston, MA

Commonly, researchers desire to know whether a given gene is actively expressed in a biological sample. Using this information, they might exclude inactive genes from an analysis. Additionally, researchers might compare gene-activation status across multiple data sets, including those that have been profiled using different gene-expression platforms. The ability to compare observations across platforms is crucial in part because researchers often want to compare their findings against publicly available, gene-expression data sets. The public domain contains more than a million gene-expression samples, but these have been generated using a panoply of platforms and thus may not be immediately comparable with each other. Previously, we developed the Universal exPRession Codes (UPC) methodology, which uses probabilistic mixture models to distinguish between active and inactive genes (or transcripts). Key features of the UPC method are that 1) it can account for GC bias and feature length, 2) its interpretation is identical for any gene-expression platform, and 3) it is applied to individual samples---unlike many other methods, which must be applied to multiple samples simultaneously. The single-sample nature of this method is particularly valuable for personalized-medicine applications where biological samples may be processed individually rather than in batches.

Previously, we incorporated the UPC method into the SCAN.UPC package, which is part of the Bioconductor framework and has been downloaded thousands of times. That version enables users to estimate gene-activation status for data generated using Affymetrix microarrays, Agilent two-color microarrays, or RNA-Sequencing. We have extended this package so that the UPC method can be applied to any expression-profiling platform, including Illumina Beadchip microarrays, Agilent one-color arrays, and qPCR values. In addition, we have added functionality that enables data to be downloaded directly from Gene Expression Omnibus in a single line of code and for samples to be processed in parallel. We have also added convenience functions that enable data sets to be adjusted for batch effects (using ComBat). Because many publicly available samples do not contain raw data, we also provide an option to apply the UPC method to data that have been preprocessed using alternative approaches.

We will demonstrate the UPC method's ability to account for nucleotide-composition biases on multiple platforms. We will also provide examples that illustrate our method's ability to identify gene-activation status consistently across samples that have been profiled using multiple platforms---and thus demonstrate the ability to integrate data across platforms. With these added features, the SCAN.UPC package promises to accelerate the process by which scientists analyze their own data sets and integrate with heterogeneous data from the public domain.

TRANSCRIPT DIFFERENTIAL ANALYSIS OF RNA-SEQ DATA WITH SLEUTH

Harold Pimentel¹, Nicolas Bray², Pall Melsted³, Lior Pachter^{1,3,4}

¹UC Berkeley, Computer Science, Berkeley, CA, ²UC Berkeley, Innovative Genomics Initiative, Berkeley, CA, ³University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, Reykjavík, Iceland, ⁴UC Berkeley, Mathematics and Molecular & Cell Biology, Berkeley, CA

We introduce a new method and tool called sleuth for the analysis and comparison of multiple related RNA-Seq experiments at the transcript level. Sleuth uses the idea of bootstraps to ascertain and correct for technical variation in experiments, a response error measurement model for inference that allows for a multitude of experimental designs, and interactive plots that enable real-time exploratory data analysis. To use sleuth, RNA-Seq data must first be quantified with kallisto, which is a program for very fast and accurate RNA-Seq quantification based on pseudoalignment.

The sleuth model for differential analysis allows for complex experimental designs, and by teasing apart technical and biological variability improves on standard approaches to differential expression such as Cuffdiff, DESeq2 and EBSeq.

The sleuth program is implemented in R and has been designed to facilitate the exploration of RNA-Seq data by utilizing the Shiny web application framework by RStudio. The code underlying all plots is available via the Shiny interface so that analyses can be fully open source.

GENOME-WIDE CHARACTERIZATION OF CHROMATIN STATE PLASTICITY

Luca Pinello¹, Alexander Gusev², Hilary Finucane², Jialiang Huang¹, Alkes Price², Guo-Cheng Yuan¹

¹Dana-Farber Cancer Institute, Biostatistics and Computational Biology, Boston, MA, ²Harvard TH Chan School of Public Health, Department of Biostatistics, Boston, MA

With the increasing amount of epigenomic data, a pressing challenge is to understand how the chromatin states are regulated; in particular the mechanisms for their cell-type specific establishment and maintenance. Here we propose a computational method based on information theoretic approaches to systematically quantify the variability of the chromatin states and study how this variability is linked to gene expression variation.

Using histone modification data from 9 human cell lines from the ENCODE project we are able to automatically highlight important functional regions based on variability of chromatin states that we called highly plastic regions (HPRs). The HPRs are enriched for many known functional categories such as super-enhancers, promoters and polycomb repressed regions. Moreover we find that the HPRs are highly enriched for GWAS-associated non-coding variants and, in some cases, explain significant heritability after controlling for enrichment of broadly used annotations for this task such as open chromatin or enhancer regions. We are currently using the chromatin state variation as a guide to search for functionally important genetic variants.

We also identify regions of co-variability based purely on histone modification data, and show that such regions correlate well with long range chromatin interactions data obtained by Hi-C or ChIA-PET assays. For example, the co-variability measure shows a clear depletion in correspondence of the boundaries of topological associated domains (TAD); suggesting that this approach could be helpful to highlight and refine functional modules within the TAD.

Our analysis provides new insights into the organization and dynamic change of cell-type specific chromatin structure during development and a valuable tool for investigating the mechanisms of chromatin state establishment and usage.

NEW GENETIC APPROACHES IN PATIENTS WITH TRANSPOSITION OF THE GREAT ARTERIES

Alex Postma^{1,2}, Fleur Tjong³, Julien Barc³, Barbara Mulder⁴, Connie Bezzina³

¹Academic Medical Center, Anatomy, Embryology & Physiology, Amsterdam, Netherlands, ²Academic Medical Center, Clinical Genetics, Amsterdam, Netherlands, ³Academic Medical Center, Experimental Cardiology, Amsterdam, Netherlands, ⁴Academic Medical Center, Cardiology, Amsterdam, Netherlands

Congenital heart disease (CHD) is the most frequent congenital disorder in newborns, affecting 7 out of 1000 live births. However, despite the improved treatment and prognosis of these patients, knowledge concerning the etiology and genetic underpinnings of CHD remains poor. The limited insight concerning the inheritance and recurrence risk of CHD hinders genetic counseling and the development of preventive interventions, such as pre-implantation genetic diagnostics. We focus on transposition of the great arteries (TGA), a severe CHD with an unknown cause in the great majority of patients. Most of TGA is sporadic and its recurrence risk is very low. This led us to hypothesize that TGA may be caused by the occurrence of de novo mutations in affected individuals or by the homozygous or compound heterozygous inheritance of genetic variants that are very rare in the general population.

The objective of our study is to identify novel genes causing TGA. We use whole exome and genome sequencing in parent-child trios, applying various strategies: (1) searching for de novo mutations in the coding and non-coding regions of the genome; (2) searching for rare mutations that are present in the coding regions of the genome homozygously or compound heterozygously in the affected child; (3) de novo copy-variant numbers in the affected child; (4) gene burden testing of candidate genes found by the above strategies in CONCOR probands. Genes identified by this approach will subsequently be validated by screening in a large set of additional TGA probands from the CONCOR database.

Pilot results: we performed WES on 15 TGA trios and identified two high priority genes. These genes were screened in hundreds of additional TGA probands. Various variants were identified and most of these had a significantly higher frequency in the population of TGA probands, than in the general population. This might suggest a possible role of these genes/variants in the pathogenesis of TGA. Our aim is to perform WGS in an additional 75 TGA trios.

AN AUTOMATED DATA MANAGEMENT SYSTEM FOR HEREDITARY CANCER ANALYSIS IN CLINICAL DIAGNOSTICS

Meera Prasad¹, Aijazuddin Syed¹, Yan Wang¹, Mustafa Syed¹, Zhen Y Liu¹, Donovan T Cheng², Marc Ladanyi¹, Liying Zhang¹, Michael F Berger¹, Ahmet Zehir¹

¹Memorial Sloan-Kettering Cancer, Pathology, New York, NY, ²Illumina Inc, San Diego, CA

Over the last few years, there has been a widespread adoption of the Next Generation Sequencing (NGS) technology by academic medical centers and research laboratories. The unprecedented throughput, reduced cost of sequencing, and detailed view of genomic information have steered NGS based molecular diagnostic tests to be offered in hospitals. The size and complexity of the data generated by NGS assays often makes it challenging to have a scalable, convenient and expandable automated management system that can provide smooth, uninterrupted analysis of the data, irrespective of the high throughput and short turn around times.

Furthermore, the sensitive nature of genetic data poses significant data privacy and security concerns and an elaborate informed consent process for patients.

We present an intelligent and expandable automated system that enables the analysis of highly penetrant cancer pre-disposition genes, shown to cause hereditary cancer syndromes in patients, which is fully integrated with data sensitivity and patient consenting. MSK-IMPACT is a clinical NGS test for detecting cancer-specific mutations, performed in the Molecular Diagnostics Service of Memorial Sloan Kettering Cancer Center, encompassing 410 genes clinically relevant to cancer. For patients with cancer, this assay typically involves the sequencing of tumor DNA alongside a patient-matched normal control; this germline DNA can reveal additional inherited cancer-predisposing mutations (“secondary” analysis). Alternatively, patients with a strong family history of cancer can opt for exclusive germline testing (“primary” analysis). The essence of the system includes the following: 1) Automatically detect new test orders from the clinical system and register the patients into a germline tracking database, grouped by the type of analysis and the cancer panel chosen for testing, 2) Gather necessary sequencing data and metadata for the patients from upstream systems in a secure fashion, 3) Perform variant calling and annotation of germline mutations, this includes de-identification of patient health information and predicting variant pathogenicity through i) in-house pathogenicity scoring scheme, ii) QIAGEN Clinical Insight, 4) Re-identify patient information and store germline analysis and mutation results in a secure and access restricted manner, 5) Load the patient results with the respective pathogenicity scores into MSK clinical variant results database. These results are then reviewed and curated by pathology attendings to generate patient specific clinical reports. The data management system (DMS) is implemented in Python and Perl with MySQL database.

BOILER: A COMPRESSION TOOL FOR BAM FILES SUPPORTING FAST, ACCURATE QUERIES

Jacob Pritt^{1,3}, Ben Langmead^{1,2,3}

¹Johns Hopkins University, Department of Computer Science, Baltimore, MD, ²Johns Hopkins University, Department of Biostatistics, Baltimore, MD, ³Johns Hopkins University, Center for Computational Biology, Baltimore, MD

RNA Sequencing (RNA-Seq) is an increasingly important tool in studying genome structure and gene expression. The Sequence Read Archive contains data for over 170,000 RNA-seq samples comprising trillions of reads. RNA-Seq reads must first be aligned by a tool such as Tophat, after which downstream tools such as Cufflinks, Stringtie, DESeq, and Derfinder can explore gene structure and expression. With the increasing availability of RNA-seq datasets covering hundreds or thousands of samples, there is a crucial need for methods that store results in a way that is both compact and easy to query. To make these public datasets as useful as possible, new methods are needed to (a) reduce the storage footprint of the data, (b) reduce the overhead associated with transferring the data to researchers, and (c) enable the sorts of queries that RNA-seq programs commonly pose.

We describe Boiler, a novel, lossy method for compressing RNA-seq alignments in SAM/BAM format. To compress, Boiler discards all but the information needed to ensure that downstream RNA-seq analyses like Cufflinks and Stringtie give substantially the same results. Reads are stored as a combination of coverage vectors and empirical read length distributions. We prove that reproducing the reads from this representation losslessly is NP-hard, but we also propose polynomial-time greedy algorithms that work well in practice. Compressed files are as small as 3% of the sorted BAM representation and 25% of the "stripped" sorted BAM with extraneous information (such as read names) removed. Boiler is "lossy" in an unusual sense: compression and decompression can cause alignments to shift along the genome, spuriously omit alignments, or even create new ones. But downstream results are substantially the same: for example, we show that transcripts assembled by Cufflinks and Stringtie are almost identical before and after compression.

Boiler compression is designed to facilitate regional queries for coverage levels, individual reads, and gene boundaries, all of which are commonly used by downstream analysis tools. These queries are not easily supported by the SAM/BAM format, but follow naturally from the coverage-based design of our compressed files. Boiler fulfills these queries quickly and accurately without the need to fully expand the compressed file.

ANNOTATING, MAINTAINING, AND CURATING REFSEQ PROKARYOTIC GENOMES.

Kim D Pruitt, Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Stacy Ciufu, Daniel Haft, Wenjun Li, Kathleen O'Neill, Tatiana Tatusova

National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD

The rate of sequenced bacterial genome submissions to the National Center for Biotechnology Information (NCBI) GenBank database, as well as to other members of the International Nucleotide Sequence Database Collaboration (INSDC), has increased rapidly in the last decade with no signs of slowing down. Historically, annotated bacterial genomes available in GenBank and in NCBI's reference sequence (RefSeq) database have been annotated by individual INSDC submitters or RefSeq pipelines using a variety of manual and automated annotation programs and differing levels of quality assurance testing. Annotation updates were infrequent, and occurred in the context of individual genomes. The variability among submitted annotation causes a number of issues, which affects downstream use of these data, including inconsistent structural and/or functional annotation (e.g., translation start sites, protein names). NCBI has developed a robust prokaryotic genome annotation pipeline (PGAP) which has been offered as a service for prokaryotic genome submissions to GenBank and is used to provide consistently annotated RefSeq prokaryotic genomes (<http://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>). In 2015 the RefSeq prokaryotic genomes data set was comprehensively re-annotated, with the exception of 122 curated reference genomes, using the PGAP software. Concordantly, RefSeq prokaryotic genomes completed a transition to a new protein data model whereby prokaryotic protein records are non-redundant across the entire prokaryotic kingdom (<http://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/>). This reannotation and new protein data model increased annotation consistency in structural and functional annotation, reduced the redundancy in the RefSeq protein data set, and spawned a more robust process for the provision and curation of functional annotation. This new effort is a multi-faceted approach that leverages functional evidence including HMMs, domain architecture, collaboration with research groups and expert databases, and NCBI staff curation. The presentation will summarize the current RefSeq genomes process flow and provide statistics and examples of annotation improvements, lowered protein redundancy, and improvements to functional annotation.

VARIATION IN TISSUE-SPECIFIC CODON USAGE ACROSS FOUR MEMBERS OF THE POACEAE.

Jane A Pulman, Megan J Bowman, Kevin L Childs

Michigan State University, Department of Plant Biology and Center for Genomics-Enabled Plant Science, East Lansing, MI

Codon usage bias occurs when synonymous codons are utilized at unequal frequencies. Although first thought to have no effect on resulting proteins or cellular function, it has been shown that changes in the use of synonymous codons can have diverse effects on RNA processing, protein translation and protein folding. Currently, two key models are proposed to explain codon usage bias: selection and mutational bias. The first suggests co-adaptation of synonymous codons and tRNA abundances in order to optimize translational efficiency. The second suggests codon usage is determined by local differences in mutational processes. The majority of the work thus far suggests that often a mixture of the two models best explains observed codon usage bias. Complicating our understanding of codon usage bias in multicellular organisms are the distinct differences in codon usage between tissue or organ type. For example, in *Arabidopsis thaliana* it has been shown that there is a strong relationship between codon usage of genes and their expression in specific tissues. Although a similar but slightly weaker relationship was shown in *Oryza sativa*, it has been determined that GC content is highly correlated with codon usage patterns in *O. sativa*. Unlike *A. thaliana*, which has a unimodal distribution and a relatively small range of GC variation, grasses have a bimodal GC distribution, which varies greatly. Very little work has been carried out to look for codon usage patterns specific to the grasses and could lead to important insights into gene evolution. Therefore, RNA-Seq data sets from five tissue types across four grass species (*O. sativa*, *Zea mays*, *Sorghum bicolor* and *Brachypodium distachyon*) were used to examine tissue-specific gene expression and codon usage. Orthologous groups of genes were determined within tissue specific gene sets and non-tissue specific gene sets to look for distinguishing patterns of codon usage.

USING ERCC SPIKE-INS AND ERCCDASHBOARD R PACKAGE TO ASSESS PERFORMANCE OF DIFFERENTIAL GENE EXPRESSION DETECTION BY ION AMPLISEQ™ TRANSCRIPTOME assays

Rongsu Qi, Srinka Ghosh, Tommie Lincecum

Thermo Fisher Scientific, Ion Torrent, South San Francisco, CA

External RNA controls, developed by External RNA Controls Consortium (ERCC), led by National Institute of Standards and Technology (NIST), are technology-independent controls used to evaluate differential gene expression experiments, including qPCR, microarray and next-generation sequencing by various platforms. Ratios of ERCCs can serve as the truth to benchmark methods. To facilitate standardization of such evaluations, NIST developed `erccdashboard` (Munro et al., 2014), an R package for data analysis and technical performance assessments from ERCC measurements. It generates a series of statistics, including ratio detection, variability, receiver operation characteristic (ROC) curves and bias estimations.

Ion AmpliSeq™ Transcriptome is a targeted transcript quantification assay based on the Ion AmpliSeq™ technology and Ion Proton™ instrument. It detects more than 20,000 RefSeq canonical transcripts. To allow this assay to be evaluated using ERCC controls, 92 ERCCs amplicons available from Ambion were included in the assay. Here we conducted experiments using universal human reference (UHR) and human brain reference (HBR) samples, each with one pool of ERCC spike-ins. We assessed the technical performance of Ion AmpliSeq™ Transcriptome using the `erccdashboard` R package, evaluating accuracy, reproducibility and limits of detection. We then compared its performance with whole transcriptome RNA-Seq experiments on Ion Proton™. We also examined the effects of sequencing depth and number of replicates per sample on AmpliSeq™ Transcriptome performance. We concluded that ERCC spike-ins combined with `erccdashboard`, is a valuable tool for standard and universal assessment of technical performance of existing and novel RNA-Seq methods. It is also very useful in assessing the confidence of individual experiments.

For Research Use Only. Not for use in diagnostic procedures.

THE ROLE OF ALTERNATIVE SPLICING AND GENE EXPRESSION IN DIFFUSE INTRINSIC PONTINE GLIOMAS

Arun K Ramani¹, Pawel Buczkowicz^{3,4}, Robert Siddaway^{3,4}, Man Yu^{3,4}, Yue Jiang¹, Patricia Rakopoulos^{3,4}, Cynthia Hawkins^{3,4}, Michael Brudno^{1,2}

¹Centre for Computational Medicine, Hospital for Sick Children, Toronto, Canada, ²Department of Computer Science, University of Toronto, Toronto, Canada, ³Division of Pathology, Hospital for Sick Children, Toronto, Canada, ⁴Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada

Alternative splicing plays a prominent role in tumour progression and regulation of splicing in cancers has provided insights into pathways that are frequently misregulated in cancers. Diffuse Intrinsic Pontine Glioma (DIPG) is a rare and fatal form of pediatric high-grade gliomas arising in the brainstem. Most pediatric gliomas exhibit complex genomic signatures with alterations in copy number, SNVs and structural rearrangements. While these genomic signatures have been well characterized, there has been very little focus on the underlying transcriptional consequences of DIPG. Here we address this deficiency by analyzing the transcriptome of 34 DIPG tumours and 17 normals sequenced to approximately 80 Million reads on average. Through differential expression analysis, we identified a number of splicing factors, such as SRSF1 and PTB that have previously been implicated in cancer regulation and progression. We compared the splicing patterns of the genes expressed in tumours and normal and identified ~600 events that are alternatively spliced. Nearly two-thirds of these alternative events are known to be neuron specific splicing changes. Furthermore, we identified a subset of events that affect known tumour suppressor genes and validated them using RT-PCR. Our results indicate that there is concerted regulation of alternative splicing in DIPGs and splicing plays a key role in the progression of DIPGs.

microRNA TARGET SITES ACT AS REGULATORY HOTSPOTS IN 3'UTRS

Simon H Rasmussen, Mireya Plass, Anders Krogh

University of Copenhagen, Biology, Copenhagen, Denmark

microRNAs (miRNAs) are endogenous short non-coding RNAs that target mRNAs leading to degradation of the mRNA. miRNAs associate with Argonaute (AGO) proteins to perform their regulatory function, and other RNA binding proteins (RBPs) modulate it. Recently, it has been proposed that mRNAs compete for miRNA regulators, this is known as the competing endogenous RNA (ceRNA) hypothesis. If RBPs can modulate target site accessibility and miRNA association, they can regulate the ceRNA networks that miRNAs form. CLIP-seq is a high-throughput assay, which is used to map binding sites of a particular RBPs transcriptome wide. In this study, we have analyzed 119 previously published CLIP-seq datasets of 49 RBPs in HEK293 cells with the aim of understanding the role of RBPs in miRNA-mediated regulation. We mapped all RBP binding sites to 3'UTRs of protein coding mRNAs and found that similarly to miRNAs, most RBPs analyzed preferentially bind to the edges of 3'UTRs. An analysis of CLIP-seq read distribution around miRNA target sites showed that most RBPs are highly enriched right on the miRNA target sites. Interestingly, the binding sites of many RBPs show a strong positional correlation not only with AGO2 binding sites but also with other RBPs. This result points towards the existence of regulatory hotspots where many RBPs, including AGO proteins, bind. In summary, our results suggest the presence of regulatory hotspots enriched on miRNA target sites. In an AGO2 knock down analysis of a ceRNA network with 961 genes we show that a subset of genes that has no hotspots overlapping miRNA target sites generally are more down-regulated by miRNAs. Our interpretation is that most RBPs act as miRNA competitors at most target sites adding an extra layer of regulation of ceRNA networks.

ACCURATE PREDICTION OF BREAKPOINTS IN SEQUENCES

Uma D Paila¹, Chun-Song Yang², Bryce Paschal², Aakrosh Ratan¹

¹University of Virginia, Center for Public Health Genomics, Charlottesville, VA, ²University of Virginia, Center for Cell Signaling, Charlottesville, VA

Structural variations have to be resolved to the level of precise nucleotide junctions if we want to understand the underlying mutational mechanisms. We present an algorithm and an accompanying implementation to predict exact breakpoints in a sample using clipped and unmapped reads in a sequencing dataset. We showcase its utility using simulated sequences replicating two common use-cases (a) human sample sequenced using short Illumina paired-end reads, both when the structural variant (SV) breakpoints are placed uniformly across the genome, and when a bias of breakpoint positioning towards repeat regions is assumed, (b) tumor sample with evidence of chromothripsis sequenced using longer PacBio reads. We use the implementation to identify structural variants in the tumors sequenced as part of the Prostate Cancer Genome Sequencing Project (dbGaP Study Accession: phs000447.v1.p1), and report on the concordance with the published results.

ISO-SEQ BIOINFORMATICS ANALYSIS WITH PACBIO LONG READS

Meisam Razaviyayn, David Tse

Stanford University, Electrical Engineering Department, Stanford, CA

The complexity of higher eukaryotic genomes imposes significant limitations on the assembly of transcript and splicing discrimination. In particular, it is known that in the presence of certain repeat structures, the RNA denovo assembly and splice product discrimination from short reads are impossible even when all constituent elements are identified. These limitations promotes the use of long read isoform sequencing (Iso-Seq) technology to discover novel splicings. In this work, we consider the denovo Iso-Seq problem using long PacBio reads. Unlike the initial natural discrete formulation of the problem, we propose an iterative convex reformulation of the problem. Then, based on the proposed reformulation, we develop a parallel multi-core software for the joint error correction and abundance estimation in the Iso-Seq problem. The numerical experiments on the heart tissue PacBio samples show that the proposed algorithm results in 10% improvement in the number of denoised reads as compared to the existing PacBio denovo Iso-Seq software TOFU.

HIGHLY ACCURATE READ MAPPING OF THIRD GENERATION SEQUENCING READS FOR IMPROVED STRUCTURAL VARIATION ANALYSIS

Philipp Rescheneder¹, Fritz J Sedlazeck², Maria Nattestad², Arndt von Haeseler^{1,3}, Michael C Schatz²

¹Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, A-1030 Vienna, Austria, ²Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ³Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

Characterizing genomic structural variations (SV) is vital for understanding how genomes evolve. Furthermore, SVs are known for playing a role in a wide range of diseases including cancer, autism, and schizophrenia. Nevertheless, due to their complexity they remain harder to detect and less understood than single nucleotide variations.

Recently, third-generation sequencing has proven to be an invaluable tool for detecting SVs. The markedly higher read length not only allows single reads to span a SV, it also enables reliable mapping to repetitive regions of the genome. These regions often contain SVs and are inaccessible to short-read mapping. However, current sequencing technologies like PacBio show a raw read error rate of 10% or more consisting mostly of insertions and deletions. Especially in repetitive regions the high error rate causes current mapping methods to fail finding exact borders for SVs, to split up large deletions and insertions into several small ones, or in some cases, like inversions, to fail reporting them at all. Furthermore, for complex SVs it is not possible to find one end-to-end alignment for a given read. The decision of when to split a read into two or more separate alignments without knowledge of the underlying SV poses an even bigger challenge to current read mappers.

Here we present NextGenMap-LR for long single molecule PacBio reads which addresses these issues. NextGenMap-LR uses a fast k-mer search to quickly find anchor regions between parts of a read and the reference and evaluates them using a vectorized implementation of the Smith-Waterman (SW) algorithm. The resulting high-quality anchors are then used to determine whether a read spans an SV and has to be split or can be aligned contiguously. Finally, NextGenMap-LR uses a banded SW algorithm to compute the final alignment(s). In this last step, to account for both the sequencing error and real genomic variations, we employ a non-affine gap model that penalizes gap extensions for longer gaps less than for shorter ones.

Based on simulated as well as verified human breast cancer SV data we show how our approach significantly improves mapping of long reads around SVs. The non-affine gap model is especially effective at more precisely identifying the position of the breakpoint, and the enhanced scoring scheme enables subsequent variation callers to identify SVs that would have been missed otherwise.

CLARIFY AND QUANTIFYING MECHANISMS OF DSB FORMATION USING MATHEMATICAL MODELLING AND BLESS SEQUENCING

Norbert Dojer¹, Jules Nde¹, Abhishek Mitra¹, Ji Li¹, Yea-Lih Lin², Anna Kubicka³, Magdalena Skrzypczak³, Nicola Crosetto⁴, Magda Bienko⁴, Ivan Dikic⁴, Krzysztof Ginalski³, Philippe Pasero², Maga Rowicka¹

¹UTMB, Biochemistry and Molecular Biology, Galveston, TX, ²CNRS, Institut de Genetique Humaine, Montpellier, France, ³University of Warsaw, Centre of New Technologies, Warsaw, Poland, ⁴Goethe University Medical School, Institut of Biochemistry II, Frankfurt, Germany

Double-stranded DNA breaks (DSBs) are most dangerous form of DNA damage. Despite many studies on the mechanisms of DSB formation, our knowledge of them is rather incomplete. A main reason for our limited knowledge of genome-wide importance of various mechanisms of DSB formation is that, to date, they have been extensively studied only at specific loci, due to lack of techniques to detect DSBs accurately genome-wide. We recently developed a method to label DSBs in situ followed by sequencing (BLESS), and used it to map DSBs in human cells with a resolution 2-3 orders of magnitude better than previously achieved. Here, we will show how mathematical modelling and numerical simulations can elucidate and quantify various mechanisms of DSB formation.

For example, there are many factors inducing DSBs, including replication stress, oxidative stress and irradiation. Most of them cause two-ended DSBs (having two free ends of DNA), the only exception is replication stress which usually induces one-ended DSBs (caused by replication fork stalling and collapse). We use this observation to infer DSBs resulting from replication stress and to analyse chromatin context and sequence features related to replication stress-induced DSBs. Moreover, we show how to reconstruct the direction of replication fork movement from BLESS read pattern. We apply this concept to infer replication domain boundaries for several cell lines and conditions and to analyse how they change upon treatments and vary between cell lines. We also provide experimental verification for the proposed computational method and show that purely computational methods can predict >80% of experimentally detected DSBs.

Even more interesting application of our approach is to discover whether replication-related DSBs are caused mostly just by replication-transcription collision or whether presence of R-loops plays key role. To this end, we simulate expected DSB distribution from each model and compare them with experimental data to calculate what percentage of the observed DSBs can be explained by each of these models. Such approach allows us also to estimate interesting model parameters, such as a probability of DSB upon replication-transcription collision or potential dependence of it on R-loop presence or time.

NORMALIZING VARIATION BETWEEN OPEN-ACCESS VARIATION DATASETS

Gary I Saunders, Dylan Spalding, Francisco J Lopez, Jag Y Kandasamy, Cristina Y Gonzalez, Justin Paschall

European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

There are a number of open-access variation datasets available for the genomics community. These are found on numerous FTP servers, various literature resources and probably one or two thousand hard-drives that are floating across the globe. We can consider this ‘data sharing’; however, it is difficult to conclusively track variant data shared in this manner, associate such variants with the correct sample(s) or indeed track these datasets in terms of updated releases. At the European Variation Archive (EVA), we aim to permit the sharing of normalized variant datasets, and the annotation of such, in a much more structured and accountable manner.

EVA contains variant data from more than 150,000 samples and our resource is growing at an increasing rate month on month due to submission of data from researchers interested in our service and the in-house addition of high-value datasets. Variants stored at EVA are annotated using a variety of methods, including the Variant Effect Predictor from Ensembl (<http://www.ensembl.org/info/docs/tools/vep/index.html>), for both the exhaustive and clinically relevant GENCODE genesets. Statistics are calculated both within and between studies, and we also permit views of submitter provided annotations. We shall present the ways in which users can mine these data via filters on the website to construct both study-centric and global queries, filtering on any combination of species, methodology, variant type, phenotype, consequence or allele frequency and how results from these queries can be downloaded in a variety of formats including VCF and CSV. Additionally, EVA provides a comprehensive RESTful web-service, to allow programmatic access, and hence the integration of these data with other resources such as Ensembl Variation and Uniprot, and this shall also be presented.

EVA works in collaboration with GA4GH to provide a Beacon against all loaded datasets and we are part of the push towards a federated system of global variation data sharing. We shall also present at the meeting our work in collaboration with ClinVar at NCBI to provide a clinically based browser, specifically for the ca. 135 thousand human variants that have been associated with at least one phenotype and a clinical significance from the ACMG classification.

THE GALAXY HiC BROWSER: AN INTERACTIVE MULTI-DIMENSIONAL GENOME TOPOLOGY BROWSER AND DATA REPOSITORY

Michael E Sauria, Carl Eberhard, James Taylor

Johns Hopkins University, Biology, Baltimore, MD

Chromatin topology is intimately linked to the mechanisms controlling gene expression, replication timing, and cell differentiation to name just a few. Along with our growing understanding of the importance of chromatin architecture, the number of publicly available chromatin interaction datasets is also rapidly expanding. In order to best utilize these data, they must be easily accessible and usable by the broader scientific community, not just labs with the resources needed to produce and analyze them. To that end, we present the HiC analysis tool suite HiFive, a visualization exploration engine, and library of raw and analyzed HiC data, all supported through the online bioinformatics platform Galaxy.

We have created a free repository of both raw and normalized HiC data from publicly-available HiC datasets, processed with HiFive. HiFive is a suite of tools efficiently implementing a number of published and novel normalization and data handling approaches for both HiC and 5C experimental data in a simple to use command-line or graphical interface. These data are available for use and download through Galaxy, making manipulation, subsampling, and integration into analysis pipelines quick and straightforward. To facilitate easy exploration and discovery using any of these datasets, we have also created a fast and dynamic multi-dimensional HiC browser that allows real-time interaction with HiC data. HiFive's browser can also integrate additional genomic data through Trackster, Galaxy's standard genome browser, to link annotations to topology. This approach allows real-time navigation of chromatin data at resolutions from kilobase to whole chromosomes. In order to achieve this, we make use of 'multi-resolution heat-mapping', an intelligent binning and indexing scheme that allows rapid access to data at resolutions that span multiple orders of magnitude.

For private or unreleased data not appearing in the HiC data repository, users can analyze and visualize their own data using HiFive through Galaxy. From raw reads to interactive heat-maps, users can process a complete HiC experiment in as little as a single click of a button. As a crucial part of the epigenome, chromatin interaction data should be available to every scientist, as well as usable without major infrastructure or dedicated bioinformatics support. We have tried to help realize this goal by creating a fully integrated, self-contained HiC analysis solution supported by the ease of use, power and resources of the Galaxy project.

QUANTITATIVE PROTEOGENOMICS APPLICATION TO PERSONALISED PROTEOMICS

Christoph Schlaffner¹, Graham Ritchie², Theodoros Roumeliotis¹, Julia Steinberg², Christine Le Maitre³, Mark Wilkinson⁴, Eleftheria Zeggini², Jyoti Choudhary¹

¹Wellcome Trust Sanger Institute, Proteomic Mass Spectrometry, Cambridge, United Kingdom, ²Wellcome Trust Sanger Institute, Department of Human Genetics, Cambridge, United Kingdom, ³Sheffield Hallam University, Biomolecular Sciences Research Centre, Sheffield, United Kingdom, ⁴University of Sheffield, Department of Human Metabolism, Sheffield, United Kingdom

The concept of using proteomics data for genome annotation was first introduced under the term “proteogenomics” in 2004. Today the term also refers to the quantitative integration of multi-omics data with the aim elucidate concordant and discordant regulation of mRNA and protein expression. Numerous studies have found modest correlation between transcript and protein expression and proteomics is therefore commonly used to confirm concordant regulation.

Recent genomics and transcriptomics efforts to identify variation between tissues and individuals have stressed the importance of uncovering determinants of phenotypic variation and disease susceptibility. Recent advancements in proteomics has enabled near complete quantitative proteome analysis, making it feasible to study variation at the protein level also. Furthermore, quantitative proteogenomics allows identification of alternative splicing, single amino acid variants, and population protein regulation. Proteomics is emerging as an important analytical technique for personalised genomics studies.

Here we propose a novel workflow for quantitative proteogenomics capable of detecting variation in the proteome originating from variation in the genome and transcriptome. To demonstrate the effectiveness of our workflow we deployed high-throughput RNA-seq and mass spectrometry to individually-matched samples of chondrocytes extracted from affected and relatively healthy articular knee cartilage in twelve osteoarthritis patients. We found peptide evidence for translation of multiple transcripts in 20 genes, including 5 transcripts for *PLEC* in 3 patients. We also identified *CFHR5* as translated in a single patient. Furthermore, we can detect single amino acid variant peptides with our workflow by integrating genotyping data. Overall, our results indicate that it is possible to identify protein regulation and inter-individual variation propagated from the genome and transcriptome in a multi-omics approach beyond cross verification.

Valerie A Schneider¹, Tina Graves-Lindsay², Kerstin Howe³, Paul Flicek⁴, Richard Durbin³

¹NCBI, IEB, Bethesda, MD, ²Washington University, McDonnell Genome Institute, St. Louis, MO, ³Wellcome Trust Sanger Institute, Genome Informatics, Hinxton, United Kingdom, ⁴European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

The human reference genome assembly plays a critical role in both basic and clinical research. It provides a coordinate system for feature annotation and communication and is the basis for variation analysis. The human reference genome assembly is the highest quality vertebrate genome available. It was assembled using a clone-based approach, rather than the whole genome assembly approach that is currently more prevalent. Its clone-based components were Sanger sequenced to finished quality, meaning they have an error rate of $< 10^{-4}$. As a result of these sequencing and assembly approaches, it continues to provide the most contiguous sequence and best representations of segmental duplications and other complex genomic features. The reference assembly also differs from the increasing number of human genome assemblies that are now publicly available because it represents multiple genomes, not that of a single individual or haplotype. Since 2007, the Genome Reference Consortium (GRC) has been responsible for updating and improving the assembly. It added further diversity with alternate loci, which provide additional sequence representations for genomic regions for where a single sequence representation is insufficient or poorly represented with mixed haplotypes. The current assembly, GRCh38, contains more than 200 such representations. The GRC continues to improve the reference assembly and alert users to genomic regions at which data interpretation may be comprised due to underlying assembly issues. Historically, the GRC has relied upon identification and sequencing of new genomic clones to add new sequence. This has ensured assembly updates are of the same or better quality than the preceding assembly. Additionally it has ensured that assembly components represent sequence found in actual human beings. However, clone-based resources are becoming scarcer, while sequences from whole genome shotgun (WGS) assemblies are increasing in length, quality and availability, as are modeled sequence representations. Thus, new approaches are needed to identify and assess sequence for use in the reference assembly and to evaluate and report local assembly quality. We will present our efforts on both fronts. We will discuss how changes in assembly resources affect GRC dataflows and procedures for assembly curation and the nature of the reference assembly and its global and regional quality. Lastly, we will present how assembly updates available in GRCh38 reflect this changing resource landscape.

REGULATORY VARIATION IN MICE WITH DIVERSE RESPONSES TO ENVIRONMENTAL STIMULI IS DRIVEN BY TRANSPOSABLE ELEMENT VARIATION AND ENVIRONMENTALLY INDUCED CHROMATIN REMODELING AT TISSUE SPECIFIC TRANSCRIPTION FACTOR BINDING SITES

Juan Du^{1,2}, Amy Leung¹, Candi Trac¹, Brian Parks³, Aldons J Lusis³, Rama Natarajan^{1,2}, Dustin E Schones^{1,2}

¹City of Hope, Department of Diabetes Complications and Metabolism, Duarte, CA, ²City of Hope, Irell & Manella Graduate School of Biological Science, Duarte, CA, ³University of California Los Angeles, Department of Medicine, Los Angeles, CA

Gene-environment interactions are involved in the susceptibility and progression of many types of complex diseases. Despite the importance of such interactions, the most relevant work to date investigating gene-environment interaction in complex diseases has been correlative, and a functional understanding of this interaction is lacking. We are investigating the interaction of environmental and genetic factors through modifications to chromatin. We have demonstrated that one manner by which environmental factors can influence molecular pathways is through modifications to chromatin, with high fat (HF) diet leading to chromatin remodeling in the liver. The genomic loci with the most dramatic diet-induced remodeling are liver regulatory regions, such as binding sites for HNF4 α , C/EBP α and FOXA1. Furthermore, the regions of greatest chromatin remodeling are largely dependent on the strain of mouse studied, indicating a genetic component in diet-induced chromatin remodeling. To further investigate the interplay of genetics and environmental factors, we are utilizing the natural genetic variation that exists between different strains of mice from the Hybrid Mouse Diversity Panel (HMPD), which display phenotypic variability in response to HF diet. The mice in this panel have been densely genotyped and have been profiled for a variety of metabolic markers, including insulin resistance and liver metabolites. We have now profiled chromatin accessibility genome-wide in liver tissue for various strains of mice from the HMDP. We have demonstrated that retrotransposon variation is a major source of chromatin variation across inbred strains of mice, indicating that transposable elements are a major driver of the genetic component of regulatory diversity across the strains. When we examined regions of chromatin variation that did not overlap with repetitive elements, we found liver specific transcription factor binding sites to be enriched in regions of greatest variation. Our results indicate that regulatory variation in mice with diverse responses to environmental stimulus is driven by transposable element variation and chromatin remodeling at tissue specific transcription factors.

DETECTION OF STRUCTURAL VARIANTS USING THIRD GENERATION SEQUENCING

Fritz J Sedlazeck, Maria Nattestad, Michael C Schatz

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology,
Cold Spring Harbor, NY

Structural Variants (SVs), which include deletions, insertions, duplications, inversions and chromosomal rearrangements, have been shown to effect organism phenotypes, including changing gene expression, increasing disease risk, and playing an important role in cancer development. Still it remains challenging to detect all types of SVs from high throughput sequencing data and it is even harder to detect more complex SVs such as a duplication nested within an inversion.

To overcome these challenges we developed algorithms for SV analysis using longer third generation sequencing reads from Pacific Biosciences. The increased read lengths, which currently average over 10kbp and some approach 100kbp, allow us to span more complex SVs and accurately assess SVs in repetitive regions, two of the major limitations when using short Illumina data. However, the long read length also comes with an increased error rate, which can cause false or poor alignments at SV sites by BWA-MEM or BLASR. These artifacts often hinder the detection of the SVs, leading to both a high false positive and a high false negative rate with current methods.

Our enhanced open-source analysis method Sniffles accurately detects structural variants based on split read mapping and assessment of the alignments. Sniffles uses a self-balancing interval tree in combination with a plane sweep algorithm to manage and assess the identified SVs. In experiments with simulated and real genomes ranging from small bacterial genomes to model organisms to human breast cancer sequenced with long reads, we find that Sniffles outperforms all other SV analysis approaches in both the sensitivity of finding events as well as the specificity of those events. In addition, Sniffles was able to detect all types of SV and even nested, more complex SVs. Central to its high accuracy is its advanced scoring model that can distinguish erroneous alignments from true breakpoints flanking SVs. Furthermore, Sniffles is highly optimized, and can detect all SVs in a *Saccharomyces cerevisiae* sample in less than 40 seconds and all SVs from a *Drosophila melanogaster* genome in less than 15 minutes. Future work will include a realignment step of the reads to further increase the accuracy of the predictions.

REFERENCE-MASKED RNA-SEQ ASSEMBLY IN HUMAN TISSUES RELEVANT FOR CARDIOVASCULAR DISEASE AND TYPE 2 DIABETES REVEALS UNANNOTATED TRANSCRIPTS AT THE CHR9P21 GWAS CANDIDATE LOCUS

Shurjo K Sen, Anna E Sappington, Cihan Oguz, Adam R Davis, Gary H Gibbons

NIH, NHGRI, Bethesda, MD

The chromosome 9p21 locus is of extreme interest for both cardiovascular disease (CVD) and Type 2 Diabetes (T2D) research. Despite genome-wide association study (GWAS) signals from a multitude of large CVD and T2D cohorts pointing to a causative genomic entity at this locus, the link between GWAS single nucleotide polymorphisms (SNPs) and disease etiology has not yet been established. Linkage disequilibrium regions harboring GWAS SNPs for both CVD and T2D are in gene desert regions. As an alternative to protein-coding genes, both non-coding RNAs and genomic enhancers have been suggested as being the effectors of disease in 9p21; however, the evidence in support of neither of these is conclusive, and hence the puzzle of 9p21 is still considered unsolved. Given recent discoveries of pervasive transcription in the human genome, we tested the hypothesis that transcriptome assembly in human tissue samples relevant for CVD and T2D would yield previously unannotated transcripts at the 9p21 GWAS locus.

As substrates for RNA-Seq assembly, we used publicly available Illumina RNA-Seq from the NCBI GEO database (<http://www.ncbi.nlm.nih.gov/geo/>). For CVD, we included reads from adult and fetal human hearts, carotid atherosclerotic lesions, non-diseased coronary and carotid arteries, CVD-associated cell lines and peripheral blood; for T2D, we focused on pancreatic islets (alpha and beta cells), adipose tissue, skeletal muscle and liver. First, we created an integrated database of currently known transcripts (including computational predictions) by merging reference transcriptome assemblies from the UCSC Genome Browser. Next, we assembled transcripts in CVD and T2D tissues after masking out the comprehensive set of known transcripts above. For putative novel transcripts arising from the above steps, we enforced presence in two replicates of each tissue and also absence of sequence homology elsewhere in the genome. Previously unannotated transcripts were found in all of the CVD and T2D tissues, which we are now evaluating for functionality based on evolutionary conservation, presence of microRNA or transcription factor binding sites. At this stage, the functional roles (if any) of these transcripts in CVD or T2D are unknown; however, given the high relevance of the 9p21 locus for both of these very common diseases, we suggest that these transcripts are worthy of independent follow-up studies as genomic agents in CVD and T2D etiology.

OPTIMIZATIONS OF PHYSICAL GENOME MAP CONTIGUITY BY *IN SILICO* LIGATION

Palak Sheth¹, Eva Chan², Alex Hastie¹, Andy Pang¹, Thomas Anantharaman¹, Erik Holmlin¹, Zeljko Dzakula¹, Xiang Zhou¹, Edward Cho¹, Vanessa Hayes², Han Cao¹

¹BioNano Genomics, Research, San Diego, CA, ²Garvan Institute of Medical Research, Research, Sydney, Australia

Genome assemblies based solely on short-read technologies are often fragmented due to structural complexities like tandem and interspersed repetitive segments, long-range structural variations, and dispersed segmental duplications. In diseases such as prostate cancer, it has been estimated that about 50% of all prostate cancers have recurrent gene fusions, structural variation events not observable with NGS1. Algorithms that improve automation, reduce manual curation time, and improve data quality and confidence are sorely needed.

The BioNano Genomics Irys® System linearizes and molecularly barcodes long DNA molecules, yielding single molecule information contiguous up to megabase lengths. The single molecule information is assembled into genome maps; once compiled, the Irys System uses this information to scaffold the sequence assemblies, validate the accuracy of the sequences, and anchor the adjacent sequences in the proper order and orientation. The long genome map contiguities are further improved by bridging the breaks at fragile sites: fragile site breaks are breaks that occur when modified restriction enzymes introduce nick sites too close in proximity to each other and an unintentional double-stranded break in the DNA molecule occurs.

We present a computational pipeline (fragileSiteRepair) that improves structural variation detection sensitivity; automating the process of fragile site prediction, scaffolding of genome maps across fragile sites, and applying confidence scores based on single molecule alignments. The resulting fragile site repaired genome maps are highly contiguous with 25-100% increase in overall N50s with a corresponding increase in structural variation sensitivity.

HIGHLY DYNAMIC EXPANSIONS OF ANTIMICROBIAL LOCI AMONG *MEDICAGO TRUNCATULA* ACCESSIONS ARE REVEALED BY INCLUSION OF SMRT SEQUENCING

Jason R Miller¹, Peng Zhou², Joann Mudge³, Thiru Ramaraj³, Brian Walenz⁴, Peter Tiffin⁵, Nevin D Young², Kevin A Silverstein⁶

¹J. Craig Venter Institute, Rockville, MD, ²University of Minnesota, Plant Pathology, Saint Paul, MN, ³National Center for Genome Resources, Santa Fe, NM, ⁴National Biodefense Analysis and Countermeasures Center, Frederick, MD, ⁵University of Minnesota, Plant Biology, Saint Paul, MN, ⁶University of Minnesota, Minnesota Supercomputing Institute, Minneapolis, MN

In order to deconvolute signatures of selection in multi-copy gene families within populations, highly contiguous genome assemblies are needed. This is due to the fact that many large receptor and secreted peptide family genes involved in adaptation appear in tandem and segmentally duplicated physical clusters of 100-250 Kb. We have developed a novel hybrid assembly pipeline called ALPACA that utilizes both Illumina and PacBio sequencing technologies to create contiguous, accurate assemblies with contig N50s in the 100-250 Kb range. We have applied this algorithm to assemble three accessions of the model legume, *Medicago truncatula*. Genome level comparisons among these three accessions and to the high-quality Sanger-based reference, A17, reveal a high level of structural and copy number variation, especially affecting genes encoding small cysteine-rich peptides (CRPs). In some cases large lineage-specific expansions have occurred in two or fewer accessions. Most lineage-specific tandem expansions were artefactually collapsed into single-copy representatives in corresponding short-read-only assemblies.

DUPLEX SEQUENCING FOR LOW ALLELE FREQUENCY DETECTION

Angad P Singh, Matt Hims, Alina Raza, Rebecca Leary, Wendy Winckler, Derek Chiang

Novartis Institutes for Biomedical Research, Next Generation Diagnostics, Cambridge, MA

Current sequencing technologies cannot detect variants reliably below ~ 5% allele frequency. Single strand nucleotide damage, PCR and sequencing errors generate false positive mutation calls that are challenging to distinguish from the true low frequency variants. Duplex sequencing has been developed by the Loeb Lab to retain information from both strands of template DNA and call only those mutations supported by a consensus family containing both strands of the template DNA. Duplex sequencing has been previously used to query mitochondrial genomes and 4 exons of the ABL1 gene. We applied a modified version of the Duplex Sequencing approach to query a 15 gene panel and developed a bioinformatics tool to analyze these data. This algorithm uses a combination of start and stop positions for each read, as well as dual indexing to create the molecular identifier (MI). The algorithm consists of 1) Identifying the unique molecules using the molecular identifier, 2) grouping MI families of reads to build consensus within a family and 3) identifying duplex-pairs to build high fidelity consensus. We generated sequencing data from a 99%:1% mixture of samples and called mutations using our duplex sequencing algorithm. These calls were compared against those made using other approaches (single-strand consensus building and Mutect). We analyzed sensitivity and specificity across the three methods to detect mutations that are exclusive to the sample mixed at 1%. These mutations were compared against calls made using Mutect on the original undiluted sample. Duplex sequencing gives similar sensitivity to alternative mutation calling methods, while significantly reducing the false positive rate.

SOMVARIUS: SOMATIC VARIANT IDENTIFICATION FROM UNPAIRED TISSUE SAMPLES

Kyle S Smith^{1,2}, Vinod K Yadav^{1,4}, Shanshan Pei³, Daniel A Pollyea^{1,3}, Craig T Jordan^{1,3}, Subhajyoti De^{1,2,4}

¹University of Colorado, Medicine, Aurora, CO, ²University of Colorado, Pharmacology, Aurora, CO, ³University of Colorado, Cancer Center, Aurora, CO, ⁴University of Colorado, Biostatistics and Informatics, Aurora, CO

Motivation: Assessment of somatic and germ line mutations to tailor personalized diagnosis and treatment is becoming a corner stone for precision medicine in oncology initiatives. Somatic mutations are typically detected by comparing sequencing data from target (e.g. tumor) and matched control tissues (e.g. benign tissue from the same patient). However, in many practical situations matched control tissues are not available, and it remains challenging to distinguish somatic and germ line variants in those cases. Popular variant calling tools are not designed to identify somatic and germ line variants from unpaired tissue samples, and there have been only limited efforts to detect somatic mutations from unpaired samples

Results: We present SomVarIUS, a computational method for detecting somatic variants using high throughput sequencing data from unpaired tissue samples. SomVarIUS identifies somatic variants in exome-seq data of ~150X coverage with at least 67.7% precision and 64.6% recall rates, when compared with paired-tissue somatic variant calls. We demonstrate utility of SomVarIUS by identifying somatic mutations in formalin-fixed samples, and tracking clonal dynamics of oncogenic mutations in targeted deep sequencing data from pre- and post-treatment leukemia samples.

Implementation and availability: SomVarIUS is written in Python 2.7 and is freely available at <http://www.sjdlab.org/resources/>

GRAMENE: COMPARATIVE PLANT GENOMICS AND PATHWAY RESOURCES

Joshua Stein¹, Wei Sharon¹, Justin Preece², Sushma Naithani², Andrew Olson¹, Yinping Jiao¹, Joseph Mulvaney¹, Sunita Kumari¹, Kapeel Chougule¹, Justin Elser², Bo Wang¹, James Thomason¹, Marcela K Tello-Ruiz¹, Peter D'Eustachio³, Robert Petryszak⁴, Paul Kersey⁴, Pankaj Jaiswal², Doreen Ware^{1,5}

¹CSHL, Plant Genomics, Cold Spring Harbor, NY, ²OSU, Botany & Plant Pathology, Corvallis, OR, ³NYU School of Medicine, Biochemistry & Molecular Pharmacology, New York, NY, ⁴EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, ⁵USDA ARS NEA, Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Gramene (www.gramene.org) is a curated resource for comparative functional genomics in crops and model plant species. It incorporates components produced in collaboration with Ensembl Plants and Reactome to process and integrate biological data for plant researchers and breeders to query, visualize, and analyse for specific purposes, such as identifying genomic regions with domestication signatures or breeding lines with desirable traits.

The current release (build 47) includes 39 complete reference genomes. Species added within the last year include cocoa, wild mustard, wild grasses, and a green unicellular alga. These build upon a foundation that includes rice, maize, sorghum, diploid and hexaploid wheats, barley, Brachypodium, banana, soybean, Arabidopsis, Brassicas, potato, tomato, grapevine, and several lower plants. Evolutionary histories are provided in phylogenetic gene trees classifying orthologous and paralogous relationships as speciation and duplication events. Orthologous genes inform synteny maps that enable inter-species browsing across ancestral regions. In addition, genome browsers from multiple species can be viewed simultaneously, with links showing homologous gene and whole-genome alignment mappings. SNP and structural diversity data, including individual genotypes, are available for 11 species, and are displayed in the context of gene annotation, along with the consequence of variation (e.g. missense variant). Visual displays can be downloaded as high-resolution, publication-ready, image files. Our Blast and BioMart interfaces enable complex queries of sequence, annotation, homology, and variation data.

Complementing our browser platform, Gramene provides plant pathway databases and tools. The Plant Reactome (<http://plantreactome.gramene.org/>) hosts 238 rice pathways (80% of which were manually curated) and orthology-based projections of the rice reference pathways to 33 plant species. Our pathway browser displays EBI-ATLAS baseline gene expression. Gramene is supported by an NSF grant IOS-1127112.

PROFILING PROTEIN OCCUPANTS OF THE GENOME: IS TF FOOTPRINTING READY FOR PRIME TIME?

Myong-Hee Sung¹, Songjoon Baek², Gordon Hager²

¹National Institutes of Health, National Institute on Aging, Baltimore, MD,

²National Institutes of Health, National Cancer Institute, Bethesda, MD

High-throughput sequencing technologies have allowed many gene locus-level molecular biology assays to become genome-wide chromatin profiling methods. DNA cleaving enzymes such as DNase I have been used to probe accessible chromatin. The accessible regions contain functional regulatory sites, including promoters, insulators, and enhancers. Chromatin mapping studies have revealed the dynamic and cell state-specific nature of accessibility *in vivo*. Deep sequencing of DNase-seq libraries and computational analysis of the cut profiles have been used to infer protein occupancy in the genome at the nucleotide-level, termed “digital genomic footprinting”. The approach has been proposed as an attractive alternative to ChIP-seq of hundreds of transcription factors, and overcomes antibody issues, poor resolution, and batch effects. Recent reports have uncovered some limitations of the DNase-based genomic footprinting approach that significantly reduce the scope of detectable protein occupancy, especially for transcription factors with dynamic chromatin binding. Moreover, transposase-accessible chromatin using sequencing (ATAC-seq) has recently been introduced as a new chromatin accessibility assay that can be performed on a small number of cells. Amid these new developments, the genomics community is grappling with issues concerning the utility of genomic footprinting and the distinction between the potential and robust deliverables of the proposed approaches. Here we summarize the consensus emerging and describe the remaining issues for genomic footprinting. We conclude that the enzyme-based protein occupancy profiling approach represents an evolving methodology that requires significant improvements to mature into a powerful tool for advancing the genomics of chromatin regulation.

DE NOVO MUTATIONS INDUCED BY MULTIPLE DNA DOUBLE STRAND BREAKS ARE REVEALED BY WHOLE-GENOME SEQUENCING OF *ARABIDOPSIS THALIANA*

Hidenori Tanaka¹, Nobuhiko Muramoto¹, Kazuto Kugou², Arisa Oda², Takahiro Nakamura², Kunihiro Ohta², Norihiro Mitsukawa¹

¹Toyota Central R&D Labs., Inc., Genome Eng. Prog., Nagakute, Aichi, Japan, ²The Univ. of Tokyo, Grad. Sch. of Arts and Sci., Meguro-ku, Tokyo, Japan

DNA double strand breaks (DSBs) induced by interruption in replicating and chemicals generate chromosomal rearrangement. During reduction division stage, most meiotic recombination is initiated by the formation of DSBs made by Spo11. Genome editing technologies using site-specific endonuclease, such as TALEN and CRIPR-Cas9, are growing rapidly in recent years. However, multiple DSBs' effect on genome in mitotic cells is still incompletely understood.

We are developing the method for inducing multiple DSBs by thermostable enzyme, such as *TaqI*, in living fungi and plant cells (TAQing system). In case of *Arabidopsis thaliana*, frequent rearrangement of GUS marker and up-regulation of DSB-responsive gene expression indicated that TAQing system trigger genome rearrangement. To check whether genome rearrangement induced in mitotic cells affect plant genome in the progeny, we conducted whole-genome sequencing using 100-bp pair-end reads. A number of mutations, such as SNV, deletion, and insertion, were detected. In addition, the mutations were distributed mainly within 200-bp from the potential *TaqI* recognition sequences in the genome. These results suggest that *de novo* mutations induced by TAQing system be inherited overcoming reproductive lineage.

INCREASING DISCOVERABILITY AND CONNECTIVITY OF SCIENTIFIC MEDIA THROUGH ANNOTATION WITH ICLIKVAL

Todd D Taylor, Naveen Kumar

RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan

Scientific media comes in a variety of languages and formats, including journal articles, books, images, videos, blog and database entries, and so on. In the case of textual media, there is often additional information, such as tables, figures and supplementary data, associated with or embedded in the text. While there are many good resources for browsing, searching and annotating some of this media, there is not one place where you can search them all, and generalized search engines such as Google, do not allow for the type of comprehensive and precise searches that researchers require. To address this, we created iCLiKVAL, an easy-to-use web-based tool that uses the power of crowdsourcing to accumulate annotation information for all scientific media found online (and potentially offline). Annotations in the form of key-relationship-value tuples (in any language), added by users with a variety of options, make information easier to find and allow for much richer data searches.

Since the introduction of iCLiKVAL at this meeting last year, we have implemented many additional features, in part thanks to feedback from the user community. Users can now create or join common interest groups, both public and private, to annotate related media together as a community. Users can now also create and edit their own controlled vocabulary lists, or import established vocabularies such as MeSH and GO. Within the user groups, vocabulary and bookmark lists can be shared to make it easy for everyone within the group to work together. In addition, we have implemented a notification center, some customization options, and a one-stop annotations feature where you can see and edit all of your own annotations. Most of the list-type pages, such as annotations, search history, reviews and vocabularies, are now searchable, sortable and filterable.

While the first version of iCLiKVAL, for development purposes, was intentionally limited to the annotation of PubMed articles, we knew that there was much more scientific media that can only be found from other sites and sources. Thus, we have now created a browser extension that allows any non-password protected online media to be bookmarked and annotated, creating unlimited possibilities for the linking of all scientific media. One could argue that any scientific media that is on the web is therefore connected, but much of it remains offline (e.g., books) or is inaccessible (not open source, only found in libraries, etc.) and is therefore not discoverable nor connected. With your help and iCLiKVAL, we hope to change that and put all scientific media at your fingertips.

EXPLORING DATABASE OPTIONS FOR STORAGE OF DIVERSE DNA SEQUENCE VARIATION DATASETS.

Jamie K Teer, Richard Z Liu, Guillermo Gonzalez-Calderon, Rodrigo Carvajal-Pelaez

H. Lee Moffitt Cancer Center and Research Institute, Department of Biostatistics and Bioinformatics, Tampa, FL

Massively parallel sequencing experiments produce large amounts of data, creating computational infrastructure challenges for many researchers. These challenges are multiplied when combining datasets using different target regions, as is often the case for institutional data collections and large consortium studies. Knowledge of the precise genotype at a given position (reference vs. variant) becomes critical when comparing samples across different platforms, but storing reference genotypes results in large storage and performance penalties. Relational database technologies have long been used to store large amounts of data, but precise reference/variant storage can quickly overwhelm capacity. New NoSQL databases promise to solve the problem of “Big Data” through novel strategies, but many different options exist, and suitability for genomic data is not well understood. To explore performance of different database strategies in storing large sequence variation datasets, we have evaluated relational SQL (MySQL), non-relational NoSQL (MongoDB, Redis), and distributed database systems (HBase) using an institutional dataset of 3,383 targeted gene sequencing tumor samples (1,321 genes), 367 TCGA melanoma whole exome somatic mutation samples, and 461 clinical tumor sequencing samples. We present performance metrics using different storage schema, including database size and common query time. In addition, we present a novel storage strategy specifically developed for precise but compact genotype storage across diverse sequencing target sets. We compare this novel strategy to more standard storage schemas. Our results will be of interest to investigators performing massively-parallel sequencing experiments, as well as bioinformaticians handling the storage of large, diverse variation datasets.

A COMPARATIVE STUDY OF METAGENOMIC ANALYSIS PIPELINES FOR ACCURATE QUANTIFICATION OF RELATIVE SPECIES ABUNDANCE

Yee Voan Teo¹, Alice Chu³, Ian Pan³, Andy Ly³, Nicola Neretti^{1,2}

¹Brown University, Department of Molecular Biology, Cell Biology and Biochemistry, Providence, RI, ²Brown University, Center for Computational Molecular Biology, Providence, RI, ³Brown University, Department of Computer Science, Providence, RI

Microorganisms detection methods such as lab culture and 16s rRNA sequencing pose limits to detection because many microbes are still “unculturable” and viruses that lack 16s RNA cannot be detected this way. The recent emergence of shotgun metagenomic sequencing has the potential to overcome these limitations and give us a more comprehensive view of the microbiome. Several metagenomic analysis pipelines have been developed to detect or quantify microbes from shotgun sequencing datasets. We tested different alignment and quantification strategies on both simulated and real datasets and assessed their ability to accurately quantify relative species abundance and discriminate between closely related species. In particular we compare and contrast count-based versus inference-based models of species abundance in terms of their accuracy and computational requirements when applied to the challenging problem of detecting and quantifying microbes within samples of human origin.

GALAXY METHYLATION TOOLKIT AS A GALAXY FLAVOR

Nitesh Turaga¹, Enis Afgan¹, Benjamin Berman^{2,3}, James Taylor¹, Galaxy Team^{1,4}

¹Johns Hopkins University, Biology, Baltimore, MD, ²University of Southern California, Keck School of Medicine, Los Angeles, CA, ³Cedars-Sinai Medical Center, Department of Biomedical Sciences, Los Angeles, CA, ⁴Pennsylvania State University, Biochemistry and Molecular Biology, State College, PA

A frequent challenge in bioinformatics-based research is supporting many different research domains, each with their own set of tools and requirements, without overwhelming researchers with choices. Ideally, this is done through a familiar interface such as the Galaxy platform. With so many tools that are not relevant to them, bioinformaticians can be swamped with an intractable and poorly managed (but needed) toolkits. We focus on DNA methylation domain as it is one of the most intensely studied epigenetic modifications in cancer research. This is because it is reversible, making it a likely target for therapeutic measures. DNA methylation is also known to play an essential role in the regulation of gene expression.

We assembled a tool suite for DNA methylation analysis that includes popular tools. This toolkit includes Bioconductor tools like *minfi* and *ELMER*, along with command line tools like *Bismark*, all available through the Galaxy interface. Earlier integration of these tools within Galaxy was not practical due to the required structure of input data. We have now implemented these tools using the newly available Galaxy dataset collections feature. The selection of tools facilitates quality control measures along with both methylation array and methylation sequencing analysis, offering a well-rounded toolkit for researchers. The toolkit is available as a ready-to-use platform via a Docker container and a cloud image.

The developed Methylation toolkit represent a *Galaxy Flavor* - an instance of Galaxy with a toolkit that has been tailored for a specific purpose. To facilitate easier generation of additional such flavors, we implemented an automated method for composing domain specific toolsets in a Galaxy server instance. Each Flavor's toolset is defined in a plain text file, which is then used to build a corresponding Galaxy instance - whether it is a local VM, a Docker based, or a cloud based instance. The file can easily be edited by domain researchers to compose a suitable toolset. Here we will showcase the newly available Methylation Flavor and describe how anyone can build their own Galaxy Flavor.

200 MAMMALS: SEQUENCE CONSERVATION AT THE SINGLE BASEPAIR LEVEL

Jason Turner-Maier¹, Jessica Alföldi¹, 200 Mammals Consortium⁷, Jeremy Johnson¹, Voichita Marinescu², Hyun Ji Noh¹, Ross Swofford¹, Eva Murén², Chris P Ponting³, Gill Bejerano⁴, Jussi Taipale⁵, Oliver Ryder⁶, Kerstin Lindblad-Toh^{1,2}

¹Broad Institute of MIT and Harvard, Vertebrate Genome Biology Group, Cambridge, MA, ²Uppsala University, Science for Life Laboratory, Uppsala, Sweden, ³University of Oxford, MRC Functional Genomics Unit, Oxford, United Kingdom, ⁴Stanford University, Department of Computer Science, Stanford, CA, ⁵Karolinska Institute, Department of Biosciences and Nutrition, Stockholm, Sweden, ⁶San Diego Zoo, Institute for Conservation Research, San Diego, CA, ⁷, MA

To determine the causes of a wide variety of diseases and traits, there is a pressing need to prioritize the candidate variants emerging from large-scale whole genome studies. Although reliable methods exist for prioritizing coding variants, it is still very difficult to accurately determine the significance or function of non-coding sequence where so many candidate variants reside. Recently, it has been shown that mammalian conservation is the human genome annotation that most enriches most highly for functional variants.

Sequence conservation among species provides a powerful tool for prioritizing candidate variants. Highly conserved sequence is likely to perform an important biological function, and variants that alter the conserved sequence may disrupt that function. Previous work has used 29 mammalian genomes to detect statistically significant levels of sequence conservation at a 12 base pair resolution. By taking advantage of improvements in sequencing and assembly technology, this project will achieve roughly single base pair resolution of constraint.

Toward this goal, we will make use of a new cost-effective *de novo* genome assembly method, DISCOVAR *de novo*. In contrast to previous methods, DISCOVAR *de novo* only requires a single read library to produce genome assemblies, greatly reducing sequencing costs. This will allow us to produce genomes for 150 placental mammals, a huge increase over the 29 genomes used for the original study. We will then create a human-centered multiple alignment over 200 mammals (50 existing + 150 novel).

Using the genome alignments, we will extract multiple conservation tracks: a high-resolution placental mammal conservation track for 200 mammals, as well as lineage-specific tracks (primate, rodent and canine). These data sets will allow comparisons across lineages, analyses of the evolutionary history of different variants or binding motifs, and correlations between candidate disease variants and different constraint patterns and elements. Our analyses will throw new light onto human evolution and will greatly assist the discovery of causative variants in traits and diseases.

IDENTIFICATION OF GLOBAL REGULATORS OF T-HELPER CELL LINEAGES SPECIFICATION

Kartiek Kanduri¹, Subhash Tripathi¹, Antti Larjo², Henrik Mannerström², Ubaid Ullah¹, Riikka Lund¹, David Hawkins³, Bing Ren⁴, Harri Lähdesmäki*², Riitta Lahesmaa*¹

¹University of Turku, Turku Centre for Biotechnology, Turku, Finland, ²Aalto University, Departments of Information and Computer Science, Espoo, Finland, ³University of Washington, School of Medicine, Seattle, WA, ⁴University of California, San Diego, Ludwig Institute for Cancer Research, La Jolla, CA

Background: Activation and differentiation of T-helper (Th) cells into Th1 and Th2 types is a complex process orchestrated by distinct gene activation programs engaging a number of genes. This process is crucial for a robust immune response and an imbalance might lead to disease states such as autoimmune diseases or allergy. Therefore, identification of genes involved in this process is paramount to further understand the pathogenesis and design intervention for immune mediated diseases.

Results: Here, we identified lineage-specific genes involved in early differentiation of T-helper 1 and 2 subsets by integrating transcriptional profiling data from multiple platforms. We have obtained a high confidence list of genes as well as new markers by employing more than one profiling platform. We showed that the density of lineage-specific epigenetic marks is higher around lineage-specific genes than anywhere else in the genome, which points to their specificity. Based on next generation sequencing data we identified lineage-specific lncRNAs involved in early T helper 1 and 2 differentiation and predicted their expected functions through Gene Ontology analysis. We showed a positive correlation between lineage-specific gene expression and lineage-specific lncRNA expression. We also found out that there is an enrichment of disease SNPs around a number of lncRNAs identified suggesting that these lncRNAs might play a role in etiology of autoimmune diseases.

Conclusion: The results presented here show the involvement of several new actors in the early differentiation of T-helper cells and will be a valuable resource for better understanding of autoimmune processes.

Keywords: Transcriptional profiles, microarrays, next-generation sequencing, T-helper cell differentiation, long non-coding RNAs, enhancers, promoters, and lineage specificity.

AN INTEGRATED ANALYSIS OF THE TRANSCRIPTIONAL RESPONSE OF HUMAN MONOCYTE-DERIVED MACROPHAGES TO LPS

Annalaura Vacca¹, Stuart Aitken¹, Kenneth J Baillie², David A Hume², Colin A Semple¹

¹MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom, ²The Roslin Institute, R(D)SVS, University of Edinburgh, Edinburgh, United Kingdom

Macrophages are characterized by high plasticity and are known to change their functional subtype depending on the environmental inputs. Here we present a rigorous analysis of genome-wide FANTOM5 CAGE time course data on the response of human macrophages to LPS. This time course is particularly densely sampled, with 22 samples over the first 24hrs post stimulus. The study is based on the classification of time course gene expression profiles to one of several predefined, broad patterns using a novel method where linear and piece-wise exponential functions are fitted to the data, given the variation among replicates. The patterns, termed kinetic signatures, are: linear, decay, dip and peak, which can be further specified as early or late peak, depending on the time of change (before or after the time course midpoint at 4hrs).

The algorithm assigned 66% of CAGE clusters, representing active TSSs, to one of the four kinetic signatures. Based on this classification, we were able to show that CAGE clusters with a sharp and temporary increment in expression during the first 4hrs, modelled by the early peak signature, are strongly enriched for known immediately early genes (IEGs) and TFs, and for GO terms specifically related to LPS stimulation. TSSs assigned to the early peak category are also significantly shorter in length (i.e. the site of transcription initiation is more precisely constrained) than others, and are relatively depleted in CpG islands; characteristics previously associated with cell type specific promoters. Genes assigned to the late peak class have an overrepresentation of terms related to the later stages of inflammation and immune processes. Integration of these results with DNase I data shows early and late peak genes occupy epigenomic landscapes characterized by chromatin accessibility, and that this cannot be explained simply by the magnitude of transcription. Genes assigned to the dip category form another interesting class of TSSs that are characterized by relatively high DNase I signal, and enrichment for TFs and GO terms related to transcriptional regulation. The expression of these genes decreases considerably but transiently after LPS stimulation. Finally, the class containing genes in which expression is seen to decay exponentially over time is enriched for terms related to fatty acid metabolism. Our analysis of the temporal profiles of CAGE expression, combined with chromatin data, reveals a rich interconnection between these layers of the macrophage response to LPS.

CRISPRSCAN: DESIGNING HIGHLY EFFICIENT SGRNAS FOR CRISPR-CAS9 TARGETING *IN VIVO*

Charles E Vejnar*¹, Miguel A Moreno-Mateos*¹, Jean-Denis Beaudoin¹, Juan P Fernandez¹, Emily K Mis², Mustafa K Khokha², Antonio J Giraldez¹

¹Yale University School of Medicine, Genetics, New Haven, CT, ²Yale University School of Medicine, Pediatrics, New Haven, CT

*These authors contributed equally to this work

CRISPR-Cas9 technology provides a powerful system for genome engineering. However, variable activity across different single guide RNAs (sgRNAs) remains a significant limitation. We analyzed the molecular features that influence sgRNA stability, activity and loading into Cas9 *in vivo* using >1000 sgRNAs. We observed that guanine enrichment and adenine depletion increased sgRNA stability and activity, whereas differential sgRNA loading, nucleosome positioning and Cas9 off-target binding were not major determinants. We also identified sgRNAs truncated by one or two nucleotides and containing 5' mismatches as efficient alternatives to canonical sgRNAs.

On the basis of these results, we modeled Cas9 nucleotide preferences to create a predictive sgRNA-scoring algorithm, CRISPRscan. This model effectively captures the sequence features affecting the activity of CRISPR-Cas9 *in vivo*. Specifically designed for *in vivo* applications where sgRNAs are directly delivered, we validated independently and demonstrated its higher performance compared to current methods. These results identify determinants that influence Cas9 activity and provide a framework for the design of highly efficient sgRNAs for genome targeting *in vivo*.

SAPLING: A TOOL FOR CUSTOMIZED NETWORK ANALYSIS FOCUSING ON PSYCHIATRIC GENETICS

Wim Verleyen, Jesse Gillis

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics,
Woodbury, NY

In the last decade, genetic screening studies for complex disorders have begun to allow us to compile gene lists containing the most ‘damaging’ mutations often derived from mutation callers and pathogenic prioritization. Typically, the gene list is analyzed for enrichment with reference gene lists such as the Gene Ontology (GO), biological pathways from KEGG and Reactome, or any other reference gene lists important for the biological function under investigation. Alternatively, a gene network analysis is also often applied – especially when there is no enrichment found for the gene list - to study the network topology of the members of the gene list.

At its most sophisticated, the network data for this analysis usually takes the form of characterization from condition-specific co-expression networks, varying by state in ways potentially relevant to disease. SAPLING is a web application that facilitates this downstream analysis of a gene list, both with user-specified network data from a wide variety of expression data and other data types. While the interface is simple from the user-perspective, a variety of high-level methods are implemented to characterize network properties and provide performance estimates. SAPLING combines all individual methods in order to construct a more robust predictor of novel candidate genes.

We focus on describing a number of use-cases for SAPLING, focusing on areas relevant to psychiatric genetics. We collect reference gene lists for recurrent proband mutations in autism, synaptic interactions, and attention deficit hyperactivity disorder; reports for these three use-cases are generated by SAPLING (<https://sapling.cshl.edu>). We report performances in terms of area under the ROC curve (AUROC) in cross-validation. Broadly, we find that aggregation across more network data adds to performance (15.5 % increase in AUROC performance). However, developing condition-specificity within the underlying data appears to be difficult; when generic data (e.g., unrelated to the conditions of interest) was used, co-expression networks provided information (AUROC ~ 0.666), whereas using brain-related data raised AUROC performance by only 4%. We find it is crucial to benchmark performance for a given disease against the network performance overall.

We show that while gene networks exhibit functional information, tailoring them in a more specific way may require non-automated choices. SAPLING is the first web application that allows the user to configure which data and algorithms they want to use for their downstream analysis.

UNVEILING THE COMPLEXITY OF MAIZE B73 TRANSCRIPTOME BY SINGLE MOLECULE LONG READ SEQUENCING

Bo Wang¹, Elizabeth Tseng², Michael Regulski¹, Tyson Clark², Ting Hon², Yinping Jiao¹, Andrew Olson¹, Joshua Stein¹, Doreen Ware^{1,3}

¹Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY, ²Pacific Biosciences, 1380 Willow Road, Menlo Park, CA, ³USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, USDA-ARS, Ithaca, NY

Zea mays is a leading model for elucidating transcriptional networks in plants, aided by increasingly refined studies of the transcriptome atlas across spatio-temporal, developmental, and environmental dimensions. Limiting this progress are uncertainties about the complete structure of mRNA transcripts, particularly with respect to alternatively spliced isoforms. Here, we use the single-molecule sequencing technology from PacBio to unveil the complexity of the maize transcriptome. Intact cDNAs from six tissues of the B73 inbred line were sequenced using the PacBio RS II platform with P6-C4 chemistry. From six size fractionated libraries of each tissue, this generated more than 1.5 million full length reads. In parallel, the same samples were used for mRNA sequencing with the Illumina HiSeq2000 PE101 platform to comprehensively validate and quantitate gene/isoform expression. Combined, these data revealed tremendous previously unidentified transcript isoforms as well as previously unrecognized protein-coding genes. In addition, novel long non-coding RNA (lncRNA) and fusion transcripts were captured in long-read sequences. Result has shown that mechanisms of alternative splicing are differentially employed between different tissues. This work has provided basis for extensive improvements to the genome annotation of maize, with added dimension of tissue-dependent transcript isoforms. This work was funded by NSF grant #1127112 and NSF #1032105.

BUILDING A DISTRIBUTED SYSTEM FOR SEQUENCE ANALYSIS

Liya Wang^{1,3}, Peter Van Buren^{1,3}, Doreen Ware^{1,2,3}

¹Cold Spring Harbor Labs, Ware Lab, Cold Spring Harbor, NY, ²USDA ARS, NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY, ³iPlant Collaborative, Thomas W. Keating Bioresearch Building, Tucson, AZ

We have successfully set up a computing system at Cold Spring Harbor Laboratory by leveraging the Cyber-infrastructures developed by the iPlant Collaborative project. The CSHL-iPlant system is the first distributed system that can execute iPlant's reproducible scientific workflows on a local computing cluster through Agave API. Besides the computing cluster, we have also set up a CSHL resource server (data server) that can hold large scale of sequencing data locally for analysis. The resource server is synchronized with iPlant Data Store via iRODS to enable metadata management via iPlant's Graphical User Interface - Discovery Environment. Comparing with the centralized main iPlant system powered by the national computing resources (XSEDE) and UA's Condor cluster, the distributed system like the CSHL-iPlant system minimizes cross-US data traffic and thus improves computing efficiency. In the mean time, the distributed system enabled easily sharing of data, workflow, and hardware among one or more research groups, institutes, or to the general public.

RAPID, DYNAMIC, AND INTERACTIVE VISUALIZATION FRAMEWORK FOR PATHOGEN IDENTIFICATION IN COMPLEX RESPIRATORY SPECIMENS FROM UNEXPLAINED RESPIRATORY DISEASE OUTBREAK RESPONSES

Yuanbo Wang^{1,2,3}, S S Morrison², H P Desai^{1,2,4}, J M Winchell²

¹APHL, Silver Spring, MD, ²CDC, NCIRD, Atlanta, GA, ³GaTech, Ctr for Operations Res. in Medicine and HealthCare, Atlanta, GA, ⁴GSU, CS, Atlanta, GA

Background

Using a targeted sequencing approach, potential pathogens can be identified within clinical specimens, but complex algorithms are required to explore the patterns within these data. Existing tools for visualizing such clinically-associated metagenomic data are static, non-interactive, or lack the ability to display hierarchical relationships. Here we present a fast, dynamic, and interactive visualization framework, which allows identification and estimation of the abundance of potential pathogens at each taxonomy level.

Methods

Illumina next generation sequencing reads were assigned taxonomic labels using Kraken. This output was used as the basis to analyze the metagenomic data. We implemented an in-house web service for visual interpretation of these data. Its back end consists of a MySQL database, which houses the parent-child relationships for the entire bacteria and virus taxonomy lineages classified in NCBI, the Kraken results, and PHP scripts for communication between two ends of the web service. The front-end was built with HTML5 and JavaScript, using D3 for flexible manipulation and display of data.

Results

The current resulting framework is an in-house web application. It allows users to upload up to 20 Kraken output files at once and instantly generate independent visualization for each as well a summary view that is a composite of all upload samples for inter-comparisons. Bubble charts were displayed in two-view windows. The top-view bubble can be clicked to display additional bubbles representing its children (taxonomic classification one rank lower) in the bottom view. This set up allows for a many-to-many comparison to be visualized. The size of the bubble reflects the number of reads collected for the classified target. The user can hover on a bubble for metadata including the taxonomic id and the number of reads associated with its classification. Transitions between visualizing components are fast and smooth, providing users with intuitive knowledge for further analysis.

Conclusions

This web visualization framework represents a powerful approach for exploring pathogen detection data and is applicable to other metagenomic data. Its implementation harnesses scalable technologies, making it compatible with multiple browsers and adoptable into existing pipelines. Its rapid, dynamic, and interactive displays enhance users' ability to interpret and analyze complex large-scale metagenomic datasets. This approach may have high utility when investigating outbreaks.

A NOVEL APPROACH TO DETERMINING NULL MODELS AND CONTROLS FOR CO-EXPRESSION NETWORKS

Melanie Weber, Jesse Gillis

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Woodbury, NY

Co-expression networks are a common and meaningful way of representing the associations between genes inferred from experimental data. In a graph theoretic setting, nodes describe genes and edges the respective correlations between gene expression levels in a set of samples. Genes associated by correlated expression levels are more likely to share functions, making co-expression networks a popular means of determining novel commonalities among sets of genes, as for disease candidate genes.

However, networks built from experimental data can be highly biased due to small sample sizes and the presence of noise. A major challenge of bioinformatics and computational data analysis lies in identifying artifacts in biased data and separating them from biological meaningful information. As in any statistical assessment, choice of the null can be critical to ensure results are meaningful. Our work introduces two novel approaches for null-models to analyze and estimate the amount of biological and technical confounds in co-expression data.

In a data-driven approach we develop a series of highly efficient tools to calculate functional properties in networks. These allow rapid controlled comparisons and analyses. Two of the most useful methods are: a function prediction algorithm which is fully vectorized, allowing network characterization across even thousands of functional groups to be accomplished in minutes in cross-validation and an analytic determination of the optimal prior to guess candidate genes across multiple functional sets. We demonstrate the methods by tracing the effects of selection biases arising in the transfer of function predictions for orthologous genes from humans to model organisms, focusing on autism candidate genes.

The second approach describes a novel null-model by evaluating the transitivity of correlations in co-expression networks from a mathematical perspective. By analyzing the network topology of modules around so-called hub genes, we determine constraints on connectivity not typically accounted for when nulls are constructed through link permutation.

DECISION TREE-BASED METHOD FOR INTEGRATING MULTI-DOMAIN DATA TO IDENTIFY CHILDHOOD OBESITY DISEASE ENDOTYPES

ClarLynda R Williams-DeVane^{1,2}, Michele Josey²

¹North Carolina Central University, Department of Biological and Biomedical Sciences, Durham, NC, ²North Carolina Central University, Biomedical Biotechnology Research Institute, Durham, NC

As the cost of ‘omic technologies have decreased, the amount of data produced has increased exponentially. However, these data have not lead to a corresponding increase in knowledge about complex human disease. We have previously shown multi-domain data integration to be integral in the further understanding of complex human diseases such as childhood asthma. We previously developed a multi-step decision tree method to identify disease endotypes of childhood asthma. We identified these disease endotypes, data-driven subtypes of human disease, leading to the understanding of complex disease etiology that facilitated the discovery of new mechanisms underlying childhood asthma. The multi-step decision tree method significantly exceeded the performance of other single domain methods and other multi-domain methods by providing superior disease group segregation and easy access to all genes and clinical covariates that distinguished the endotypes. Here, we present an extension of our multi-step decision tree method to integrate epigenetic, epidemiologic, electronic medical record (EMR), and disease status data to identify endotypes of childhood obesity. This serves as an extension to our previous method based on gene expression data. We will specifically present the maternal epigenetic etiology of childhood obesity. By considering multiple data types at each step, we are able better able to capture the heterogeneity of diseases such as childhood obesity as compared to traditional analysis methods based on disease status alone. The extension of the multi-step decision tree method to epigenetic data increases our ability to understand heritable and modifiable disease associations not possible with gene expression data alone.

SCALING CANCER SUBPOPULATION PHYLOGENY RECONSTRUCTION TO THOUSANDS OF TUMORS

Jeff A Wintersinger¹, Amit G Deshwar², Quaid Morris³

¹University of Toronto, Department of Computer Science, Toronto, Canada, ²University of Toronto, Edward S Rogers Sr Department of Electrical and Computer Engineering, Toronto, Canada, ³University of Toronto, The Donnelly Center, University of Toronto, Toronto, Canada, Toronto, Canada

Tumors are composed of distinct, heterogeneous cellular populations, each containing a different selection of mutations. By observing the frequencies of mutations across all cells in a tumor, we can infer the existence of these subpopulations, as well as their evolutionary histories relative to one another. Past studies have relied solely on semi-manual methods for reconstructing tumor phylogenies, limiting their application to only tens of cancers. Moreover, with few experts possessing the requisite expertise to perform such work, subclonal reconstruction could only be confidently performed by a handful of labs.

We recently published PhyloWGS, an automated method for reconstructing tumor phylogenies. Subsequently, we have made numerous improvements to help non-experts to perform phylogenetic reconstructions at large scales. PhyloWGS uses Markov Chain Monte Carlo methods to probabilistically sample from a Bayesian posterior over phylogenies consistent with observed mutation frequencies; as such, our results consist not of a single phylogeny for each tumor, but of thousands of phylogeny estimates. Given this multitude of evolutionary history reconstructions, we must understand which phylogeny portions are consistent across estimates, and which portions vary, so as to communicate areas of certainty and uncertainty. Moreover, we require consensus representations that illustrate the primary structures observed in each sample of evolutionary histories, ignoring minor discrepancies. To realize these goals, we have begun clustering sampled phylogenies for each tumor. Rather than presenting the researcher with thousands of trees, we instead return only a handful of clusters, each of which represents a plausible phylogeny. Guided by scores indicating our confidence in each cluster, the non-expert researcher can use additional data to determine the likely correct phylogeny.

We have also made substantial other improvements to PhyloWGS. These include a means of determining how heavily to weigh copy-number variations (CNVs) relative to single nucleotide variations (SNVs) when reconstructing tumor phylogenies; an input parser that supports a variety of popular applications for calling SNVs and CNVs, simplifying the integration of PhyloWGS into existing tumor analysis pipelines; and a web-based visualization tool that allows interactive exploration of reconstructed tumor phylogenies for hundreds of different datasets.

To utilize PhyloWGS' ability to perform phylogenetic reconstructions for thousands of tumors, we joined the Pan-Cancer Analysis of Whole Genomes project, where we are analyzing 2500 tumors drawn from 23 cancer types. By adapting PhyloWGS to a cluster-computing environment, we have been able to reconstruct phylogenies for thousands of tumors in parallel, demonstrating the method's ability to operate at massive scales.

GENETIC INSIGHTS INTO JUVENILE IDIOPATHIC ARTHRITIS DERIVED FROM DEEP WHOLE GENOME SEQUENCING

Laiping Wong¹, Kaiyu Jiang², James N Jarvis³

¹University at Buffalo, Department of Pediatrics and Genetics, Genomics, & Bioinformatics Program, Buffalo, NY, ²University at Buffalo, Department of Pediatrics and Genetics, Genomics, & Bioinformatics Program, Buffalo, NY, ³University at Buffalo, Department of Pediatrics and Genetics, Genomics, & Bioinformatics Program, Buffalo, NY

Deep whole genome sequencing (WGS) is an unprecedented opportunity to comprehensively study genetic landscapes at finer resolution than can be achieved with chip-based methods. We are particularly interested juvenile idiopathic arthritis (JIA), a complex trait that represents one of the most common chronic disease conditions in children. Although GWAS have identified regions associated with this disease, such studies were limited in scope (only regions represented by the Illumina Immuchip were surveyed) and resolution. Thus, finer mapping is needed to clarify genetic landscapes of children with JIA to gain insights into pathogenesis and treatment responses.

We conducted deep whole genome sequencing on 29 JIA individuals using the Illumina HiSeq x10 system at an average sequencing depth of 39x. Samples included a pair of biological replicates giving a total of 30 sets of paired-end sequencing data. We compared the discovered variants with 1000 Genomes Project and NCBI dbSNP141. We identified 798,504 novel SNPs. In addition, we detected 186,625 novel indels (64,595 insertions and 122,030 deletions between 1 and 50 basepairs in size). We found 356 novel SNPs and 82 novel indels within LD blocks comprising known JIA risk loci as identified by Hinks et al (Nature Genet, 2013). We next examined how many of the novel variants mapped to known binding sites of the transcription factor, MYC, which we have previously identified as an important regulator of JIA-associated gene co-expression networks. Intersection of novel variants with MYC binding sites yielded overlapping of 12,677 SNPs, 1,012 insertions and 1,793 deletions. We also identified 191 novel SNPs and 63 novel indels situated within differentially methylated regions that we have previously observed in the comparison of JIA neutrophils with those from healthy children. We next surveyed genetic variants situated within histone marked regions as a first step toward interrogating the influence of variants on histone modifications. We associated novel variants with H3k4me1 and H3K27ac marks in neutrophils, and identified 47,389 and 17,014 SNPs located within 20,493 H3k4me1 and 9,684 H3K27ac histone marks respectively. We also identified 4,301 and 1,461 novel insertions within 3,269 H3k4me1 and 1,253 H3K27ac marks. Finally, 5,432 H3k4me1 marks overlapped with 8,026 novel deletions and 1,933 H3K27ac marks encompassed 2,557 novel deletions.

Preliminary results revealed previously unreported genetic variations in children with JIA, including private mutations at the individual patient level. These data also provide a vantage point from which to understand the complex genetic-epigenetic interactions that are likely to underlie complex illness such as JIA.

KRAKEN 2: FASTER AND MORE SENSITIVE METAGENOMIC CLASSIFICATION

Derrick E Wood^{1,2}, Ben Langmead^{1,2}

¹Johns Hopkins University, Department of Computer Science, Baltimore, MD, ²Johns Hopkins University, Center for Computational Biology, Baltimore, MD

Through the use of exact k-mer alignment and pre-computing lowest common ancestor queries, Kraken introduced the ability to quickly and accurately classify the species found in metagenomic samples. Kraken requires considerable memory - nearly 80 GB - to be available to perform its classification, and the memory required increases as more genomes are added to its reference database. Although we introduced the ability to run Kraken with a reduced database ("MiniKraken"), such databases provide lower sensitivity when compared to Kraken's full database.

By reorganizing Kraken's data structures, Kraken 2 provides the ability to build a database using a much larger set of reference data while using less than 16 GB of memory. The inclusion of a more diverse reference set also provides users with increased sensitivity in classification. In addition, Kraken 2's new structure allows for classification speeds 1.5 to 3 times faster than Kraken 1. We will also describe some of the improvements made in targeting 16S classification and in preventing misclassification due to reference contamination.

NANOSIM: NANOPORE SEQUENCING READ SIMULATOR BASED ON STATISTICAL CHARACTERIZATION

Chen Yang^{1,2}, Justin Chu^{1,2}, René L Warren², Inanç Birol^{2,3,4}

¹University of British Columbia, Faculty of Science, Vancouver, Canada, ²British Columbia Cancer Agency, Genome Science Centre, Vancouver, Canada, ³University of British Columbia, Department of Medical Genetics, Vancouver, Canada, ⁴Simon Fraser University, School of Computing Science, Burnaby, Canada

Background:

In 2014, Oxford Nanopore Technologies (ONT) announced a new, portable single-molecule sequencing platform called MinION. As the first pre-commercial nanopore sequencer, MinION stands out among existing sequencing technologies due to its ability to generate ultra-long read lengths (E.g., the *S. cerevisiae* dataset has an average read length of 5473bp, and the maximum is reaching 147 kbp), though at the expense of high error rate. The particular features of the reads show great potential in genome analysis, while posing new challenges in algorithm design. However, as of yet, there exists no software that simulates MinION platform reads with genuine ONT characteristics.

Results:

In this work, publicly available datasets utilizing R7 and R7.3 chemistry were characterized statistically. Through sequence alignment, we observed unique patterns of correct base calls and errors (mismatches and indels), which can be described by statistical mixture models. The structures of the models are consistent between both chemistries and data from two organisms (*E. coli* and *S. cerevisiae*). Building on these features, we have developed a read simulator, NanoSim, which generates synthetic ONT reads with empirical quality profile or customized parameters. The lengths of intervals between errors (stretches of correct bases) are modeled by a Markov chain, and the lengths of errors are drawn from the mixed statistical models. Empirical distributions of read length and alignment ratio are incorporated in our algorithm to improve the fidelity of the simulation.

Conclusion:

Simulation of ONT reads is essential for developing and benchmarking genomic applications, including read alignment, *de novo* assembly, and genomic variation discovery. Here, we introduce NanoSim, a fast and lightweight read simulator that captures the technology-specific features of ONT data and allows for adjustments upon improvement of nanopore sequencing technology.

CLOUD-BASED VARIANT DISCOVERY USING GENOMEVIP

R. Jay Mashl, Kai Ye, Li Ding

Washington University in St Louis, The Genome Institute, St Louis, MO

Identifying genomics variants is a fundamental step in the understanding of inherited traits and acquired mutations that may precede the development of disease. The ever-increasing body of next-generation sequencing data poses a significant challenge to researchers who use the traditional “download-and-analyze” model. As a result, cloud computing has emerged as an attractive analysis platform as it enables access to vast amounts of storage and processing power. However, analysis pipelines must be ported to the cloud to take advantage of these resources. We have therefore developed the Genome Variant Investigation Platform (GenomeVIP), a lightweight, web-driven, extensible genomics pipeline that enables germline, somatic, and de novo variant analysis to be performed on Amazon’s cloud. GenomeVIP deploys multiple computationally intense discovery tools capable of analyzing whole-genome and exome sequence data using best practices and customizable parameter sets.

PROBABILISTIC MODEL FOR DETECTING MRNA TRANSLATION EFFICIENCY CHANGES FROM RIBOSOME PROFILING

Yi Zhong¹, Theofanis Karaletsos*¹, Philipp Drewe*², Vipin Sreedharan¹, Kamini Singh³, Hans-Guido Wendel³, Gunnar Rättsch¹

¹Memorial Sloan Kettering Cancer Center, Computational Biology Program, New York, NY, ²Max Delbrück Center for Molecular Medicine, Computational Regulatory Genomics, Berlin, Germany, ³Memorial Sloan Kettering Cancer Center, Cancer Biology Program, New York, NY

Deep sequencing based ribosome footprint profiling allows the identification of mRNA fragments that are bound by the ribosome complex. It provides valuable information on ribosome occupancy and protein synthesis activity. This recently established technology can be leveraged by combining the measurements from RNA-Seq estimates in order to determine a gene's translation efficiency, which provides insights into the hidden level of translational regulation. However, the observed ribosome profile is fundamentally confounded by transcriptional activity. In order to decipher principles of translational regulation, tools that can reliably detect changes in translation efficiency in case-control studies are needed. In this study, we developed a statistical framework and analysis tool, *RiboDiff*, to detect genes with changes in translation efficiency across experimental treatments with biological replicates (<http://bioweb.me/ribodiff>). *RiboDiff* uses generalized linear models to estimate the over-dispersion of RNA-Seq and ribosome profiling measurements separately, accounting for the fact that two different sequencing protocols having distinct statistical characteristics. *RiboDiff* robustly performs a statistical test for differential translational efficiency while taking the variability of transcription as the "confounding factor" into account. We prove that *RiboDiff* can reliably discern the translational effect in case-control experiments using both simulated and real biological data. The release of this tool enables proper statistical analyses of data from any ribosome/RNA-Seq profiling experiments and, hence, facilitates to decipher the unknown translational regulation mechanisms that lead to variety of phenotypes and diseases.

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Mr. Robert Aboukhalil
Cold Spring Harbor Laboratory
raboukha@cshl.edu

Dr. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Dr. Zaky Adam
Agriculture and Agri-Food Canada
Zaky.Adam@AGR.GC.CA

Dr. Jan Aerts
KU Leuven, Belgium
Jan.Aerts@esat.kuleuven

Dr. Bahman Afsari
Johns Hopkins University
bahman.afsari@gmail.com

Dr. Joo Wook Ahn
Guy's and St Thomas' NHS Foundation
Trust
jwahnchris@gmail.com

Mr. Gediminas Alzbutas
Thermo Fisher Scientific
gediminas.alzbutas@thermofisher.com

Mr. David Amar
Tel-Aviv University
ddam.am@gmail.com

Mr. Lin An
Pennsylvania State University
lua137@psu.edu

Ms. Joann Arce
Brigham Young University
jdirayarce@byu.edu

Dr. Francois Artiguenave
CEA - IG - CNG
artiguenave@cng.fr

Mr. Luay Aswad
Nanyang Technological University
luay001@e.ntu.edu.sg

Ms. SENA BAE
Duke University
sb168@duke.edu

Mr. Taejeong Bae
Mayo Clinic
bae.taejeong@mayo.edu

Dr. Gnanaprakash Balasubramanian
German Cancer Research Institute (DKFZ)
g.balasubramanian@dkfz.de

Dr. Sara Ballouz
CSHL
sballouz@cshl.edu

Dr. Riyue Bao
University of Chicago
rbao@uchicago.edu

Dr. Christophe Battail
CEA / Institut de Génomique
christophe.battail@cea.fr

Dr. Alexis Battle
Johns Hopkins University
ajbattle@cs.jhu.edu

Dr. Brady Bernard
Institute for Systems Biology
bbernard@systemsbiology.org

Ms. Gail Binkley
Stanford University
gail.binkley@stanford.edu

Dr. Daniel Blankenberg
Penn State University / Galaxy Team
dan@bx.psu.edu

Mr. Emil Bouvier
The Pennsylvania State University
dave@bx.psu.edu

Dr. Megan Bowman
Michigan State University
mbowman@msu.edu

Dr. Florian Breitwieser
Johns Hopkins University
florian.bw@gmail.com

Dr. Michael Brudno
Hospital for Sick Children
brudno@cs.toronto.edu

Mr. Riccardo Brumm
Center of human genetics
info@medizinische-genetik.de

Dr. Scott Cain
OICR
scott@scottcain.net

Dr. Brandi Cantarel
Baylor Health Care System
brandi.cantarel@baylorhealth.edu

Mr. Stefan Canzar
Toyota Technological Institute at Chicago
canzar@ttic.edu

Dr. Xia Cao
Bayer CropScience
xia.cao@bayer.com

Mr. Martin Cech
Penn State University
marten@bx.psu.edu

Ms. Sylvia Chang
Thermo Fisher Scientific
sylvia.chang@thermofisher.com

Dr. Tingwen Chen
Chang Gung University
s18909032@ym.edu.tw

Dr. Ye Chen
National Institutes of Health
ye.chen@nih.gov

Ms. Jenny Chen
University of Massachusetts Medical
jjenny@mit.edu

Dr. Zhenxia Chen
National Institutes of Health
zhen-xia.chen@nih.gov

Mr. Yuping Chen
Stony Brook University
yuping.chen@stonybrook.edu

Mr. Colby Chiang
Washington University
cchiang3@gmail.com

Mr. John Chilton
Pennsylvania State University
jmchilton@gmail.com

Dr. Jason Chin
Pacific Biosciences, Inc
jchin@pacb.com

Mr. Kapeel Chougule
Cold Spring Harbor Laboratory
kchougul@cshl.edu

Dr. Deanna Church
Personalis
deanna.church@personalis.com

Dr. Alain Coletta
InSilico DB
alaincoletta@insilicodb.com

Mr. Daniel Cooke
Wellcome Trust Centre for Human
Genetics
dcooke@well.ox.ac.uk

Mr. Chris Cremer
University of Toronto
chris.a.cremer@gmail.com

Mr. Steven Criscione
Brown University
steven_criscione@brown.edu

Dr. Shareef Dabdoub
The Ohio State University
dabdoub.2@osu.edu

Dr. Ryan Dale
NIH
dalerr@nidk.nih.gov

Ms. Charlotte Darby
Carnegie Mellon University
cdarby@andrew.cmu.edu

Mr. Rishi Das Roy
University of Helsinki
rishi.dasroy@helsinki.fi

Ms. Harriet Dashnow
Murdoch Childrens Research Institute
harriet.dashnow@mcri.edu.au

Dr. Jaime Davila
Mayo Clinic
davila.jaime@mayo.edu

Dr. Adam Davis
NHGRI
adam.davis2@nih.gov

Dr. Subhajyoti De
University of Colorado
subhajyoti.de@ucdenver.edu

Dr. Laura DeMare
Genome Research/Molecular Case Studies
ldemare@cshl.edu

Mr. Alan Derr
University of Massachusetts Medical
alan.derr@umassmed.edu

Mr. Jairav Desai
Eli Lilly and Company
jairav@gmail.com

Mr. Amit Deshwar
University of Toronto
amit.deshwar@utoronto.ca

Dr. Alexander Dobin
CSHL
dobin@cshl.edu

Dr. Jason Dobson
Novartis Institutes for BioMedical Research
jason.dobson@novartis.com

Ms. Tamsen Dunn
Illumina
tamsen.dunn@gmail.com

Dr. Nathan Dunn
Lawrence Berkeley Labs
nathandunn@lbl.gov

Dr. Daniel Ence
University of Utah
dence@genetics.utah.edu

Mr. Simon Eng
University of Toronto
simon.eng@mail.utoronto.ca

Mr. Bertrand Escaliere
IB2
bescalie@ulb.ac.be

Mr. Han Fang
Cold Spring Harbor Laboratory
hanfang.cshl@gmail.com

Prof. Chris Fields
U of Illinois at Urbana-Champaign
cjfields@illinois.edu

Mr. Steven Foltz
Washington University
sfoltz@genome.wustl.edu

Dr. Mallory Freeberg
Johns Hopkins University
mfreebe2@jhu.edu

Dr. Iddo Friedberg
Iowa State University
idoerg@iastate.edu

Mr. Alexander Frieden
Claritas Genomics
alexander.frieden@claritasgenomics.com

Dr. Nick Fulcher
Nature Protocols
nick.fulcher@nature.com

Mr. Brendan Gallagher
Sentieon
brendan.gallagher@sentieon.com

Dr. Shouguo Gao
National Institutes of Health
gaos2@nih.gov

Dr. Shouguo Gao
National Institutes of Health
gaos2@nih.gov

Prof. Manuel Garber
University of Massachusetts Medical
School
manuel.garber@umassmed.edu

Mr. Gonazlo Garcia
The Genome Analysis Centre
business.support@tgac.ac.uk

Mr. Andrea Gazzo
IB2
andrea.gazzo.86@gmail.com

Dr. Mark Gerstein
Yale University
paaa@gersteinlab.org

Dr. Daniel Gilchrist
NHGRI
gilchristd@mail.nih.gov

Dr. Jesse Gillis
Cold Spring Harbor
jgillis@csHL.edu

Dr. Jesse Gillis
Cold Spring Harbor
jgillis@csHL.edu

Dr. Thomas Gingeras
Cold Spring Harbor Laboratory
gingeras@csHL.edu

Dr. Hani Girgis
University of Tulsa
hani-girgis@utulsa.edu

Dr. Tamara Goldfarb
National Center for Biotechnology
Information
tamara.goldfarb@nih.gov

Dr. Leonard Goldstein
Genentech, Inc.
goldstein.leonard@gene.com

Dr. Ryan Golhar
Bristol-Myers Squibb
ryan.golhar@bms.com

Dr. Giorgio Gonnella
University of Hamburg, Germany
gonnella@zbh.uni-hamburg.de

Ms. Aparna Gorthi
Univ of Texas Health Science Center at
San Antonio
gorthi@livemail.uthscsa.edu

Ms. Anaïs Guoin
Inria
anais.gouin@irisa.fr

Ms. Ewa Grabowska
New York Genome Center
egrabowska@nygenome.org

Dr. Casey Greene
University of Pennsylvania
csgreene@upenn.edu

Mr. Tom Groot Kormelink
University of Applied Sciences Leiden
info@hsleiden.nl

Dr. Theresa Guo
Johns Hopkins Hospital
tguo5@jhmi.edu

Dr. Simone Gupta
Eli Lilly
gupta_simone@lilly.com

Mr. James Gurtowski
Cold Spring Harbor Laboratory
gurtowsk@cshl.edu

Ms. Melissa Gymrek
Whitehead Institute
mgymrek@mit.edu

Mr. Kevin Ha
University of Toronto
k.ha@mail.utoronto.ca

Dr. Lukas Habegger
Regeneron
lukas.habegger@regeneron.com

Mr. Thomas Hackl
Max-Planck-Institute for medical Research
thackl@mpimf-heidelberg.mpg.de

Dr. Nancy Hansen
NHGRI/NIH
nhansen@mail.nih.gov

Ms. Yuan Hao
Cold Spring Harbor Laboratory
yhao@cshl.edu

Dr. Jason Harris
Personalis
eliane.sousa@personalis.com

Dr. Stephen Hartley
NHGRI
stephen.hartley@nih.gov

Dr. Ron Hause
University of Washington
hauser@uw.edu

Dr. Fritz Hauser
Aerzte-Gesundheitszentrum Laegern AG
segmenta@gmx.net

Mr. James Havrilla
University of Utah
semjaavria@gmail.com

Dr. Miao He
Illumina Cambridge Ltd
mhe1@illumina.com

Dr. Kun He
Monsanto
kun.he@monsanto.com

Dr. Maxime HEBRARD
RIKEN Yokohama - IMS - LIB
maxime.hebrard@riken.jp

Dr. Raphael Helaers
de Duve Institute (Universite catholique
Louvain)
raphael.helaers@uclouvain.be

Mr. Sunghoon Heo
Yonsei University
shhuh@yonsei.ac.kr

Dr. Mark Hickman
Rowan University
hickmanm@rowan.edu

Dr. Stephanie Hicks
Dana-Farber Cancer Institute / Harvard
SPH
shicks@jimmy.harvard.edu

Prof. Winston Hide
Sheffield Institute for Translational
Neuroscience
winhide@sheffield.ac.uk

Dr. Benjamin Hitz
Stanford University
hitz@stanford.edu

Dr. Michael Hoffman
Princess Margaret Cancer Centre
michael.hoffman@utoronto.ca

Dr. Celine Hong
NIH/NHGRI
celine.hong@nih.gov

Dr. Pingsha Hu
Syngenta
pingsha.hu@syngenta.com

Mr. ByungJin Hwang
Yonsei University
bjhwang113@yonsei.ac.kr

Dr. Yuval Itan
The Rockefeller University
yitan@rockefeller.edu

Dr. Pierre-Etienne Jacques
Université de Sherbrooke
Pierre-Etienne.Jacques@USherbrooke.ca

Dr. Andrew Jaffe
Lieber Institute for Brain Development
andrew.jaffe@libd.org

Dr. Pankaj Jaiswal
Oregon State University
jaiswalp@science.oregonstate.edu

Dr. Aaron Jex
University of Melbourne
ajex@unimelb.edu.au

Dr. Ying Jin
Cold Spring Harbor Laboratory
yjjin@cshl.edu

Dr. Jyoti Joshi
Dalhousie University
jyotijoshi111@gmail.com

Prof. Haja Kadarmideen
University of Copenhagen
hajak@sund.ku.dk

Dr. Andre Kahles
Memorial Sloan Kettering Cancer Center
akahles@cbio.mskcc.org

Ms. Gurmanna Kalra
Rowan University
hickmanm@rowan.edu

Mr. VENKATESWARA KANAPARTHI
THERMO FISHER SCIENTIFIC
VENKATESWARA.KANAPARTHI@LIFETE
CH.COM

Dr. John Karro
Miami University,
karroje@miamiOH.edu

Dr. Laura Kavanaugh
Syngenta Biotechnology
laura.kavanaugh@syngenta.com

Mr. Keffy Kehrli
Stony Brook University
keffy.kehrli@stonybrook.edu

Dr. David Kelley
Harvard University
dkelley@fas.harvard.edu

Dr. Janet Kelso
Max Planck Institute for Evolutionary
Anthropology
kelso@eva.mpg.de

Mr. Mohd Khan
University of Turku
mkhan@btk.fi

Mr. Hamza Khan
BC Cancer Agency
lclarke@bcgsc.ca

Dr. Jihye Kim
University of Colorado School of Medicine
Jihye.Kim@UCDenver.edu

Dr. Daehwan Kim
Center for Computational Biology
infphilo@gmail.com

Dr. Seok-Won Kim
RIKEN
seokwon.kim@riken.jp

Dr. Junho Kim
Yonsei University College of Medicine
kimjh607@yuhs.ac

Dr. Sangtae Kim
Illumina
skim2@illumina.com

Dr. Hyunmin Kim
University of Colorado, School of Medicine
hyun.kim@ucdenver.edu

Dr. Paul Kitts
NIH/NLM
kitts@ncbi.nlm.nih.gov

Dr. James Knnight
Yale University
j.knight@yale.edu

Dr. Sergey Koren
National Institutes of Health
sergek@umd.edu

Dr. Toshinori Kozaki
Tokyo University of Agriculture and
Technology
kozakit@cc.tuat.ac.jp

Dr. Arjun Krishnan
Princeton University
arjunk@princeton.edu

Mr. Muhammet Erdi Kucuk
BC Cancer Agency
lclarke@bcgsc.ca

Dr. Alper Kucukural
University of Massachusetts Medical
School
alper.kucukural@umassmed.edu

Mr. Naveen Kumar
RIKEN Center for Integrative Medical
Sciences
naveen.kumar@riken.jp

Dr. V Kumar
CSHL
vkumar@cshl.edu

Dr. Kasper Lage
MGH / Harvard / Broad Institute
lage.kasper@mgh.harvard.edu

Mr. Sangeet Lamichhaney
Uppsala University
sangeet.lamichhaney@imbim.uu.se

Dr. Ben Langmead
Johns Hopkins University
langmea@cs.jhu.edu

Mr. Young-suk Lee
Princeton University
youngl@princeton.edu

Ms. Hayan Lee
Lawrence Berkeley National Laboratory
hayan.lee@lbl.gov

Mr. Marc-André Legault
Université de Montréal
marc-andre.legault.1@umontreal.ca

Dr. Louis-Philippe Lemieux Perreault
Beaulieu-Saucier Pharmacogenomics
Centre
louis-
philippe.lemieux.perreault@statgen.org

Dr. Florian Lenz
University Medical Center Tuebingen
Florian.Lenz@med.uni-tuebingen.de

Dr. Suzanna Lewis
Lawrence Berkeley National Laboratory
selewis@lbl.gov

Dr. Jun Li
Bristol-Myers Squibb
jun.li3@bms.com

Ms. Qing Li
University of Utah
liqing850104@gmail.com

Dr. Yong Li
Janssen Research & Development
yli280@its.jnj.com

Prof. Shoudan Liang
Pacific Biosciences
sliang@pacb.com

Mr. Eun-Cheon Lim
Max Planck Institute for Developmental
Biology
euncheon.lim@tue.mpg.de

Dr. Michael Livstone
Celgene
mlivstone@celgene.com

Mr. Tharvesh Moideen Liyakat Ali
Université Paris Diderot - Paris 7
tharvesh.moideen@univ-paris-diderot.fr

Dr. Leslie Low
Malaysian Palm Oil Board
lowengti@mpob.gov.my

Mr. Zhenyuan Lu
CSHL
luj@cshl.edu

Dr. Gerton Lunter
University of Oxford
gerton.lunter@well.ox.ac.uk

Mr. Zunping Luo
New York University
zunping@nyu.edu

Dr. Gholson Lyon
Cold Spring Harbor Laboratory
gholsonjlyon@gmail.com

Dr. Daniel MacArthur
Massachusetts General Hospital
macarthur@atgu.mgh.harvard.edu

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Mr. Samuel Macleon
Rowan University
hickmanm@rowan.edu

Mr. AMINE MADOU
CEA - GENOSCOPE
amadoui@genoscope.cns.fr

Mr. Anup Mahurkar
University of Maryland Baltimore
amahurkar@som.umaryland.edu

Mr. David McKean
Marquette University
david.mckean@marquette.edu

Prof. Aoife McLysaght
Trinity College Dublin
mclysaga@tcd.ie

Dr. Pall Melsted
University of Iceland
pmelsted@gmail.com

Dr. Karyn Meltz Steinberg
Washington University School of Medicine
kmeltzst@genome.wustl.edu

Dr. Alison Meynert
University of Edinburgh
ameynert@gmail.com

Mr. Aleksandar Mihajlovic
Seven Bridges Genomics
aleksandar.mihajlovic@sbgenomics.com

Ms. Sanja Mijalkovic
Seven Bridges Genomics
sanja.mijalkovic@sbgenomics.com

Dr. Christopher Miller
Washington University
cmiller@genome.wustl.edu

Dr. David Molik
Cold Spring Harbor Lab
dmolik@cshl.edu

Dr. Quaid Morris
University of Toronto
quaid.morris@utoronto.ca

Mr. Joseph Mulvaney
CSHL
jmulvane@cshl.edu

Dr. Adriana Munoz
Cold Spring Harbor Laboratory
amunoz@cshl.edu

Dr. Ryuichiro Nakato
Univ. of Tokyo
rnakato@iam.u-tokyo.ac.jp

Dr. Maria Nattestad
Cold Spring Harbor Laboratory
mnattest@cshl.edu

Dr. Abhinav Nellore
Johns Hopkins University
anellore@jhu.edu

Dr. Nicola Neretti
Brown University
nicola_neretti@brown.edu

Mr. Tobias Neumann
Research Institute of Molecular Pathology
tobias.neumann@imp.ac.at

Dr. Ekaterina Nevedomskaya
Netherlands Cancer Institute
e.nevedomskaya@nki.nl

Ms. Kun Nie
University of Toronto
kunnie525@gmail.com

Dr. Zemin Ning
Wellcome Trust Sanger Institute
zn1@sanger.ac.uk

Ms. Anke Nissen
Medical Genetics Center (MGZ)
nissen@mgz-muenchen.de

Dr. Barbara Novak
Agilent Technologies
barbara_a_novak@agilent.com

Dr. Cihan Oguz
National Institutes of Health
cihan.oguz@nih.gov

Dr. Laura Okagaki-Vraspir
North Carolina State University
laura.okagaki@gmail.com

Mr. Andrew Olson
Cold Spring Harbor Laboratory
olson@cshl.edu

Mr. Tien-chi Pan
University of Pennsylvania
tspan@mail.med.upenn.edu

Dr. Dhruv Pant
Univ. of Pennsylvania
dpant@mail.med.upenn.edu

Dr. Meeyoung Park
University of Michigan
mparkbio@med.umich.edu

Dr. YoSon Park
Perelman School of Medicine Univ of
Pennsylvania
ypar@upenn.edu

Mr. Akshay Paropkari
The Ohio State University
paropkari.1@buckeyemail.osu.edu

Dr. Justin Paschall
EMBL-EBI
paschall@ebi.ac.uk

Mr. Matt Paul
University of Pennsylvania
mattpaul@mail.med.upenn.edu

Mr. Nathaniel Pearson
New York Genome Center
npearson@nygenome.org

Dr. Brent Pedersen
University of Utah
bpederse@gmail.com

Mr. William Pembroke
MRC Functional Genomics Unit, Oxford
University
william.pembroke@new.ox.ac.uk

Dr. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Dr. Lon Phan
NIH/NLM/NCBI
lonphan@ncbi.nlm.nih.gov

Mr. Adam Phillippy
National Institutes of Health
aphillippy@gmail.com

Dr. Stephen Piccolo
Brigham Young University
stephen_piccolo@byu.edu

Mr. Harold Pimentel
UC Berkeley
haroldpimentel@gmail.com

Dr. Luca Pinello
Dana-Farber Cancer Institute
lpinello@jimmy.harvard.edu

Mr. Milos Popovic
Seven Bridges Genomics
milos.popovic@sbgenomics.com

Dr. Alex Postma
Academic Medical Center
a.v.postma@amc.uva.nl

Ms. Meera Prasad
Memorial Sloan Kettering Cancer Center
prasadm@mskcc.org

Mr. Jacob Pritt
Johns Hopkins University
jacobpritt@gmail.com

Dr. Kim Pruitt
NCBI/NLM/NIH
pruitt@ncbi.nlm.nih.gov

Dr. Jane Pulman
Michigan State University
pulmanj@msu.edu

Dr. Rongsu Qi
Thermo Fisher Scientific
rongsu.qi@thermofisher.com

Dr. Jean Qin
Bayer
jean.qin@bayer.com

Dr. Aaron Quinlan
University of Utah
aaronquinlan@gmail.com

Ms. Michal Rabani
Harvard University
michalra@gmail.com

Dr. Goran Rakocevic
Seven Bridges Genomics
goran.rakocevic@sbgenomics.com

Dr. Arun Ramani
Hospital for Sick Children
arun.ramani@sickkids.ca

Mr. Simon Rasmussen
University of Copenhagen
simras@binf.ku.dk

Mr. Aakrosh Ratan
University of Virginia
ratan@virginia.edu

Dr. Meisam Razaviyayn
Stanford University
meisamr@stanford.edu

Dr. Aviv Regev
Broad Institute
aregev@broadinstitute.org

Mr. Claudio Reggiani
IB2
claudio.reggiani@ulb.ac.be

Mr. Philipp Rescheneder
Max F. Perutz Laboratories
philipp.rescheneder@univie.ac.at

Dr. Diogo Ribeiro
Wellcome Trust Sanger Institute
dr7@sanger.ac.uk

Dr. Maga Rowicka
University of Texas Medical Branch
merowick@utmb.edu

Ms. Yulia Rubanova
University of Toronto
rubanova@cs.toronto.edu

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Dr. Gary Saunders
EMBL-EBI
garys@ebi.ac.uk

Dr. Michael Sauria
Johns Hopkins University
crookpotamus@gmail.com

Ms. Florentine Scharf
Medical Genetics Center (MGZ)
florentine.scharf@mgz-muenchen.de

Dr. Michael Schatz
Cold Spring Harbor Laboratory
mschatz@cshl.edu

Mr. Christoph Schlaffner
Wellcome Trust Sanger Institute
cs25@sanger.ac.uk

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@ncbi.nlm.nih.gov

Dr. Dustin Schones
City of Hope
dschones@coh.org

Dr. Fritz Sedlazeck
Cold Spring Harbor Laboratory
fritz.sedlazeck@gmail.com

Dr. Chris Seidel
Stowers Institute For Medical Research
cws@stowers.org

Dr. Shurjo Sen
NIH
sensh@mail.nih.gov

Ms. Rachel Sherman
Johns Hopkins University
rsherman@jhu.edu

Mr. Palak Sheth
BioNano Genomics
jcampione@bionanogenomics.com

Dr. Han Si
Molecular Characterization & CLinical
Assay Dev
han.si@fnlcr.nih.gov

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Kevin Silverstein
University of Minnesota
kats@umn.edu

Dr. Meromit Singer
Broad Institute
msinger@broadinstitute.org

Mr. Angad Singh
Novartis Institutes for Biomedical Research
angad.singh@novartis.com

Ms. Charlotte Siska
University of Colorado Anschutz Medical
Campus
charlotte.siska@ucdenver.edu

Mr. Kyle Smith
University of Colorado
kyle.s.smith@ucdenver.edu

Dr. Joshua Stein
Cold Spring Harbor Laboratory
steinj@cshl.edu

Dr. Tomasz Stokowy
University of Bergen
tomasz.stokowy@k2.uib.no

Dr. Myong-Hee Sung
National Institutes of Health
sungm@mail.nih.gov

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Dr. Hidenori Tanaka
TOYOTA CENTRAL R&D LABS., INC.
e1613@mosk.tytlabs.co.jp

Mr. Mayank Tandon
NIH/NIDCR/MPTB
Mayank.Tandon@nih.gov

Mr. Yin Tang
Pennsylvania State University, University
Park
yxt148@psu.edu

Dr. Todd Taylor
RIKEN Center for Integrative Medical
Sciences
taylor@riken.jp

Dr. Jamie Teer
Moffitt Cancer Center
Jamie.Teer@moffitt.org

Dr. Marcela Tello-Ruiz
COLD SPRING HARBOR LABORATORY
mmonaco@cshl.edu

Ms. Yee Voan Teo
Brown University
yee_voan_teo@brown.edu

Ms. Uduak Thomas
GenomeWeb
uthomas@genomeweb.com

Ms. Ngoc Tran
Cold Spring Harbor Laboratory
ntran@cshl.edu

Mr. Cole Trapnell
University of Washington
coletrap@uw.edu

Dr. Ilker Tunc
National Institutes of Health
ilker.tunc@nih.gov

Dr. Ilker Tunc
National Institutes of Health
ilker.tunc@nih.gov

Mr. Nitesh Turaga
Johns Hopkins University
nitesh.turaga@gmail.com

Dr. Stephen Turner
University of Virginia
sdt5z@virginia.edu

Mr. Jason Turner-Maier
The Broad Institute
jturner@broadinstitute.org

Dr. Ubaid Ullah
University of Turku
uullah@btk.fi

Ms. Annalaura Vacca
MRC Institute of Genetics and Molecular
Medicine
annalaura.vacca@gmail.com

Mr. Karel van Duijvenboden
Academic Medical Center
k.vanduijvenboden@amc.uva.nl

Dr. Charles Vejnar
Yale University
charles.vejnar@yale.edu

Mr. Daniel Vera
The Florida State University
vera@genomics.fsu.edu

Dr. Wim Verleyen
Cold Spring Harbor Laboratory
wverleye@cshl.edu

Dr. Bo Wang
Cold Spring Harbor Lab
bwang@cshl.edu

Dr. Liya Wang
Cold Spring Harbor Labs
wangli@cshl.edu

Mr. Yuanbo Wang
Center for Disease Control - Fellow
mcdy143@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory USDA ARS
ware@cshl.edu

Dr. Susanne Warrenfeltz
University of Georgia
swfeltz@uga.edu

Ms. Melanie Weber
Cold Spring Harbor Laboratory
mweber@cshl.edu

Dr. Owen White
University of Maryland Baltimore
owhite@som.umaryland.edu

Dr. ClarLynda Williams-DeVane
North Carolina Central University
clarlynda.williams@nccu.edu

Mr. Jeff Wintersinger
University of Toronto
jeff@wintersinger.org

Dr. Lai Ping Wong
University at Buffalo
LAIPINGW@BUFFALO.EDU

Dr. Derrick Wood
Johns Hopkins University
dwood@cs.jhu.edu

Mr. Adam Wright
OICR
frances.dirnbeck@oicr.on.ca

Dr. Chao Wu
Zhejiang University
wuchao1984@zju.edu.cn

Dr. Chunlin Xiao
NIH/NLM/NCBI
xiao2@ncbi.nlm.nih.gov

Dr. Ethan Xu
Infinity Pharmaceuticals
ethan.xu@infi.com

Mr. Tao Yang
The Pennsylvania State University
txy146@psu.edu

Ms. Chen Yang
BC Cancer Agency
lclarke@bcgsc.ca

Dr. Kai Ye
Washington University
kye@genome.wustl.edu

Dr. William Young
Genentech
youngw8@gene.com

Dr. Chengfeng Zhao
DuPont Pioneer
chen.zhao@pioneer.com

Dr. Yi Zhong
Memorial Sloan Kettering Cancer Center
zhongy@cbio.mskcc.org

Dr. Sai Zhou
Stony Brook University
sai.zhou@stonybrook.edu



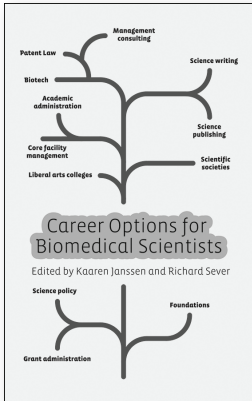
bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

A nonprofit resource from
Cold Spring Harbor Laboratory
for all the biosciences

More details at bioRxiv.org

New book from Cold Spring Harbor Laboratory Press



Career Options for Biomedical Scientists

Edited by Kaaren Janssen, *Cold Spring Harbor Laboratory Press*
and Richard Sever, *Cold Spring Harbor Laboratory Press*

The majority of PhDs trained in biomedical sciences do not remain in academia. They are now presented with a broad variety of career options, including science journalism, publishing, science policy, patent law, and many more. This book examines the numerous different careers that scientists leaving the bench can pursue, from the perspectives of individuals who have successfully made the transition. In each case, the book sets out what the job involves and describes the qualifications and skill sets required.

2015, 232 pp., illustrated, index

Hardcover \$45

ISBN 978-1-936113-72-9

Visit cshlpress.org for special offers



VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	(9) 742-3300	(9) 692-4747
Ambulance	(9) 742-3300	(9) 692-4747
Poison	(9) 542-2323	(9) 542-2323
Police	(9) 911	(9) 549-8800
Safety-Security	Extension 8870	

Emergency Room Huntington Hospital 270 Park Avenue, Huntington	631-351-2300 (1037)
Dentists Dr. William Berg Dr. Robert Zeman	631-271-2310 631-271-8090
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	631-423-5400 (1034)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	631-549-9400 (1039)

Free Speed Dial

Dial the four numbers (****) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers

BOOKSTORE 367-8837 (hours posted on door)

Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers

Photocopying – Main Library

Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri

10:00 a.m. – 6:00 p.m. Saturday

Helpful tips – Use PIN# 62450 to enter Library after hours.

See Library staff for photocopier code.

Computers, E-mail, Internet access

Grace Auditorium

Upper level: E-mail and printing

STMP server address: mail.optonline.net

To access your E-mail, you must know the name of your home server.

Dining, Bar

Blackford Hall

Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00

Bar 5:00 p.m. until late (Cash Only)

Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes, ATM

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m.

Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, east wing, lower level

PIN#: Press 62450 (then enter #)

Concierge

On duty daily at Meetings & Courses Office.

After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	9-1-800-321-0288
MCI	9-1-800-674-7000

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station (\$9.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (1031)
US Limousine Service	800-962-2827, ext:3 (1047)
Super Shuttle	800-957-4533 (1033)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (1030)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (1032)

Trains

Long Island Rail Road	822-LIRR
<i>Schedules available from the Meetings & Courses Office.</i>	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 (1036)
Orient Point/ New London	631-323-2525 (1038)

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322