Abstracts of papers presented
at the 2015 meeting on

THE BIOLOGY OF GENOMES

May 5–May 9, 2015

Cold Spring Harbor Laboratory
1890 125 2015

Abstracts of papers presented
at the 2015 meeting on

THE BIOLOGY OF GENOMES

May 5–May 9, 2015

Arranged by

Ewan Birney, *EBI/EMBL, UK*
Michel Georges, *University of Liege, Belgium*
Elaine Mardis, *Washington University School of Medicine*
Molly Przeworski, *Columbia University*

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

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.

Front Cover: "Genome Mosaicism."

Lynn Fellman is an independent multimedia artist from Washington, D.C. specializing in human evolution and genomic science. She works with scientists to communicate the beauty and benefit of their research. She is currently a Fulbright Senior Scholar to Israel producing an animation about evolutionary genetics in the mitochondria.

Fellman believes that art and narrative are a gateway to appreciating science. Her signature style pairs lyrical story and digital paintings with explanations of genomics. Working in a variety of media from animated videos and interactive eBooks to digital paintings and DNA portraits, Fellman endeavors to find a balance between accurate information and poetic interpretation.

The cover image are portions from a large digital painting in the collection of BioBusiness Alliance in Minneapolis, Minnesota. See the original image and more of Fellman's work on her web site http://fellmanstudio.com. Contact the artist at Lynn@Fellmanstudio.com to discuss communicating your research.

THE BIOLOGY OF GENOMES

Tuesday, May 5 – Saturday, May 9, 2015

Tuesday	7:30 pm	**1** Functional Genomics
Wednesday	9:00 am	**2** Cancer and Medical Genomics
Wednesday	2:00 pm	**3** Poster Session I
Wednesday	4:30 pm	*Wine and Cheese Party**
Wednesday	7:30 pm	**4** Genetics of Complex Traits
Wednesday	10:30 pm	*Happy Hour*
Thursday	9:00 am	**5** Computational Genomics
Thursday	2:00 pm	**6** Poster Session II
Thursday	4:30 pm	**7** ELSI Panel and Discussion
Thursday	7:30 pm	**8** Population Genomics
Thursday	10:30 pm	*Happy Hour*
Friday	9:00 am	**9** Evolutionary and Non-human Genomics
Friday	2:00 pm	**10** Poster Session III
Friday	4:30 pm	GUEST SPEAKERS
Friday	6:00 pm	Banquet
Saturday	9:00 am	**11** Translational Genomics and Genetics

Lunchtime Workshop: Illumina, Wednesday, 12:30 pm
See abstract following NOTES section.

* *Airslie 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

TUESDAY, May 5—7:30 PM

SESSION 1 FUNCTIONAL GENOMICS

Chairpersons: **R. Darnell,** New York Genome Center, New York
G. Petukhova, Uniformed Services University of the Health Sciences, Bethesda, Maryland

Hotspots of recombination initiation in the mouse genome
Fatima Smagulova, Kevin Brick, R. Daniel Camerini-Otero, Galina V. Petukhova.
Presenter affiliation: Uniformed Services University of the Health Sciences, Bethesda, Maryland. 1

Evolution of gene regulation in 20 mammals
Camille Berthelot, Diego Villar, Duncan T. Odom, Paul Flicek.
Presenter affiliation: European Molecular Biology Laboratory, Hinxton, United Kingdom. 2

Spatial transcriptomics—A method for gene expression analysis of multiple regions within whole tissue sections
Fredrik Salmén, Patrik Ståhl, Sanja Vickovic, Anna Lundmark, Stefania Giacomello, José Fernandez, Michaela Asp, Emelie Berglund, Annelie Mollbrink, Pelin Sahlén, Jens Magnusson, Joel Sjöstrand, Erik Sjölund, Mikael Huss, Jakub Westholm, Jonas Frisén, Joakim Lundeberg.
Presenter affiliation: Science for Life Laboratory/Royal Institute of Technology (KTH), Solna, Sweden. 3

Thousands of novel translated open reading frames and dually coded regions accurately inferred using ribosome footprinting data
Anil Raj, Sidney Wang, Heejung Shim, Yang Li, Matthew Stephens, Yoav Gilad, Jonathan K. Pritchard.
Presenter affiliation: Stanford University, Stanford, California. 4

Robert Darnell.
Presenter affiliation: New York Genome Center, New York, New York.

WEDNESDAY, May 6—9:00 AM

SESSION 2 CANCER AND MEDICAL GENOMICS

Chairpersons: **C. Dive,** University of Manchester, United Kingdom
E. Mardis, Washington University School of Medicine, St. Louis, Missouri

SESSION 3 POSTER SESSION I

A simple and powerful new approach for generating and improving genome assemblies
Nicholas H. Putnam, Jonathan C. Stites, Brendan L. O'Connell, Brandon J. Rice, Jarrod A. Chapman, Charles W. Sugnet, Tomas Marques-Bonet, Wesley C. Warren, Andrew Fields, Paul D. Hartley, David Haussler, Daniel S. Rokhsar, Richard E. Green.
Presenter affiliation: Dovetail Genomics, LLC, Santa Cruz, California; University of California, Santa Cruz, Santa Cruz, California. 122

Identifying signatures of paternal transgenerational genetic effects on mouse transcriptomes
Rodrigo Gularte Mérida, Audrey Tromme, Fabien Ectors, Benoit Hennuy, Wouter Coppieters, Carole Charlier, Michel Georges.
Presenter affiliation: GIGA - Research, Liège, Belgium. 123

Error correction and de novo assembly of Oxford Nanopore Sequencing
James Gurtowski, Sara Goodwin, Scott Ethe-Sayers, Panchu Deshpande, Michael C. Schatz, W. Richard McCombie.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 124

WEDNESDAY, May 6—4:30 PM

Wine and Cheese Party

WEDNESDAY, May 6—7:30 PM

SESSION 4 GENETICS OF COMPLEX TRAITS

Chairpersons: **J. Flint,** University of Oxford-Wellcome Trust Centre for Human Genetics, United Kingdom
N. Soranzo, Wellcome Trust Sanger Institute, Cambridge, United Kingdom

The interplay of genomes and epigenomes in hematopoietic development and cardiovascular disease
Nicole Soranzo.
Presenter affiliation: Wellcome Trust Sanger Institute, Cambridge, United Kingdom; University of Cambridge, Cambridge, United Kingdom. 125

Detecting gene-by-environment interactions using allele specific expression
David A. Knowles, Joe R. Davis, Stephen B. Montgomery, Alexis Battle.
Presenter affiliation: Stanford University, Stanford, California. 132

WEDNESDAY, May 6—10:30 PM

Happy Hour
Sponsored by Swift Biosciences

THURSDAY, May 7—9:00 AM

SESSION 5 COMPUTATIONAL GENOMICS

Chairpersons: **J. Marioni,** EBI-EMBL, Hinxton, Cambridge, United Kingdom
F. Michor, Harvard School of Public Health/Dana Farber Cancer Institute, Boston, Massachusetts.

Computational challenges in single-cell biology and applications in mammalian development
John Marioni.
Presenter affiliation: European Molecular Biology Laboratory, Cambridge, United Kingdom; Wellcome Trust Sanger Institute, Cambridge, United Kingdom. 133

Exploration of genetic variation and genotypes among millions of genomes
Ryan M. Layer, Konrad J. Karczewski, Exome Aggregation Consortium (ExAC), Aaron R. Quinlan.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 134

A DNA code governs chromatin accessibility
Tatsunori Hashimoto, Richard Sherwood, Daniel Kang, Amira Barkal, Haoyang Zeng, Bart Emons, Sharanya Srinivasan, Nisha Rajagopal, Tommi Jaakkola, David Gifford.
Presenter affiliation: MIT, Cambridge, Massachusetts; Harvard University and Medical School, Cambridge, Massachusetts. 135

DNA methylation dynamics in pigment cell development
Rebecca F. Lowdon, Hyung Joo Lee, Stephen L. Jonshon, Ting Wang.
Presenter affiliation: Washington University in St. Louis, Saint Louis, Missouri. 209

A genomic view of local adaptation
David B. Lowry.
Presenter affiliation: Michigan State University, East Lansing, Michigan. 210

Assembling maize inbred CML247—The maize pan-genome takes off
Fei Lu, Robert Bukowski, Qi Sun, Edward S. Buckler.
Presenter affiliation: Cornell University, Ithaca, New York. 211

Systematic identification of GxE determinants of gene expression
Roger Pique-Regi, Gregory Moyerbrailean, Chris Harvey, Omar Davis, Donovan Watza, Xiaoquan Wen, Francesca Luca.
Presenter affiliation: Wayne State University, Detroit, Michigan. 212

The brachiopod genome of *Lingula anatina* provides insight into the evolution of lophotrochozoans and calcium-phosphate-based biomineralization
Yi-Jyun Luo, Takeshi Takeuchi, Ryo Koyanagi, Lixy Yamada, Miyuki Kanda, Mariia Khalturina, Manabu Fujie, Shinichi Yamasaki, Kazuyoshi Endo, Noriyuki Satoh.
Presenter affiliation: Okinawa Institute of Science and Technology Graduate University, Onna, Japan. 213

Discovery and genetic characterization of new neuropsychiatric syndromes from family-based studies
Gholson J. Lyon, Jason O'Rawe, Yiyang Wu, Han Fang, Laura Jimenez Barron, Giuseppe Narzisi, Michael Schatz, Min He, Kai Wang.
Presenter affiliation: CSHL, Cold Spring Harbor, New York. 214

Integrated analysis of protein-coding variation in over 60,000 individuals
Daniel G. MacArthur, Exome Aggregation Consortium.
Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts; Broad Institute of Harvard and MIT, Cambridge, Massachusetts. 215

SESSION 7 ELSI PANEL AND DISCUSSION

Genomic Data Sharing: Past, Present And Future

Moderator: **Lawrence Brody,** National Human Genome Research Institute, National Institutes of Health

Panelists

Laura Rodriguez, National Human Genome Research Institute
Eric Campbell, Harvard University
Daniel O'Connor, The Wellcome Trust

One extraordinary aspect of the human genome project was the commitment of all parties to free and broad data sharing. The rapid release of human sequence data allowed for external data validation and accelerated the study of human genetics and genomics. Recognizing that access to large, information-rich genomic data sets can increase our understanding of the factors that contribute to health and disease, the NIH enacted a policy mandating the sharing data from genome-wide association studies. Most recently, the NIH issued a new Genomic Data Sharing (GDS) Policy (which became effective in January 2015). This policy is broad in scope and applies to all NIH-funded research that generates large-scale human or non-human genomic data.

Under the new policy, special attention is given to the conditions under which data derived from humans are to be shared. The ability to share data generated from individuals enrolled in studies before 2015 will need to be interpreted in the context of the extant consent documents. For studies commenced in 2015, the policy encourages investigators to seek the broadest possible informed consent from research participants for the use of their de-identified data (as well as de-identified cell lines or clinical specimens). Timelines for data submission and access established in the policy are intended to promote timely and broad sharing, while being mindful of the significant effort required by investigators to recruit research participants and generate and prepare the data for release.

This session will place current approaches to genomic data sharing – both as embodied in the recently issued GDS policy in the U.S. and as practiced more globally – into their broader historical context. Discussants will address the ways that data sharing in other areas of science has been approached over the years and how genomic data sharing practices, in particular, have evolved since the inception of the Human Genome Project. Practical, political and publication issues associated with this new policy will be discussed. The panel will also take up the increasingly challenging privacy and related ethical issues associated with the broad sharing of genomic data, especially with the continuing proliferation of massive biobanks and biorepositories.

SESSION 8 POPULATION GENOMICS

Chairpersons: **K. Bomblies,** Harvard University, Cambridge, Massachusetts
A. Price, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

Frequency of mosaicism points towards mutation prone early cleavage cell divisions
Chad Harland, Carole Charlier, Latifa Karim, Nadine Cambisano, Wouter Coppieters, Michel Georges.
Presenter affiliation: University of Liège, Liège, Belgium. 256

Inter-individual variation in epigenetic marks between human induced pluripotent stem cell lines
Angela Goncalves, Natsuhiko Kumasaka, Andrew Knights, Francesco Casale, Jose Garcia-Bernardo, Daniel Gaffney, on behalf of the HipSci Consortium.
Presenter affiliation: Wellcome Trust Sanger Institute, Cambridge, United Kingdom. 257

Linking immune responsive regulatory variation and population adaptation to pathogen pressure
Maxime Rotival, Helene Quach, Eddie Loh, Julien Pothlichet, Etienne Patin, Guillaume Laval, Nora Zidane, Christine Harmant, Marie Lopez, Geert Leroux-Roels, Frédéric Clément, Jean-François Deleuze, Lluis Quintana-Murci.
Presenter affiliation: Human Evolutionary Genetics Unit CNRS URA3012, Paris, France. 258

THURSDAY, May 7—7:30 PM

Happy Hour
Sponsored by Illumina

FRIDAY, May 8—9:00 AM

SESSION 9 EVOLUTIONARY AND NON-HUMAN GENOMICS

Chairpersons: **D. Bachtrog,** University of California, Berkeley
J. Willis, Duke University, Durham, North Carolina

Genetic analysis of parallel local adaptation to serpentine and mine soils in *Mimulus*
Kevin M. Wright
evin M. Wright, Jessica Selby, Annie Jeong, Uffe Hellsten, Daniel S. Rokhsar, John H. Willis.
Presenter affiliation: Duke University, Durham, North Carolina. 259

SESSION 10 POSTER SESSION III

Investigating the molecular underpinnings of human hippocampal neurogenesis and the effects of antidepressants
Timothy R. Powell, Tytus Murphy, Simone de Jong, Jack Price, Sandrine Thuret, Gerome Breen.
Presenter affiliation: King's College London, London, United Kingdom. 277

Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript
Marta Puig, David Castellano, Lorena Pantano, Carla Giner-Delgado, David Izquierdo, Magdalena Gayà-Vidal, José Ignacio Lucas-Lledó, Tõnu Esko, Chikashi Terao, Fumihiko Matsuda, Mario Cáceres.
Presenter affiliation: Institut de Biotecnologia i Biomedicina, Bellaterra, Barcelona, Spain. 278

Real-time monitoring of disease progression by longitudinal analysis of tumor subclone structure in refractory breast cancer patients
Yi Qiao, Sam W. Brady, Andrea Bild, Gabor T. Marth.
Presenter affiliation: University of Utah School of Medicine, Salt Lake City, Utah. 279

Denisovan ancestry in East Eurasian and Native American populations
Pengfei Qin, David Reich, Mark Stoneking.
Presenter affiliation: Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. 280

Assessing cell-to-cell DNA methylation variability on individual long reads
Wei Qu, Hideaki Yurino, Shin-ichi Hashimoto, Tatsuya Tsukahara, Hiroyuki Takeda, Shinichi Morishita.
Presenter affiliation: University of Tokyo, Tokyo, Japan. 281

Identification of genes involved in functional recovery after stroke through exome sequencing of extreme phenotypes
Raquel Rabionet, Marina Mola, Carolina Soriano, Caty Carrera, Georgia Escaramís, Stephan Ossowski, Israel Fernandez-Cadenas, Jordi Jimenez-Conde, Xavier Estivill.
Presenter affiliation: Center for Genomic Regulation (CRG), UPF and CIBERESP, Barcelona, Spain. 282

Characterizing subclonal evolution in lymphoma
Deepthi Rajagopalan, Jenny Zhang, Andrea Moffitt, Anupama Reddy, Casandra Love, Tiffany Tzeng, Sandeep Dave.
Presenter affiliation: Duke University, Durham, North Carolina. 283

Non-diploid indel discovery via *de novo* assembly
Nicholas Stoler, Boris Rebolledo-Jaramillo, Marcia Shu-Wei Su, Kateryna D. Makova, Anton Nekrutenko.
Presenter affiliation: Penn State University, State College, Pennsylvania. 324

A statistical model for signal detection and bias correction in ChIP-Seq data
Alexander Engelhardt, Georg Stricker.
Presenter affiliation: Gene Center Munich, Munich, Germany. 325

Functional genetic variants in the Chromogranin A gene promoter govern plasma protein levels by differential transcription regulation
Lakshmi Subramanian, Prasanna K R. Allu, Bhavani S. Sahu, Abrar A. Khan, Malapaka Kiranmayi, Ajit S. Mullasari, Nitish R. Mahapatra.
Presenter affiliation: Indian Institute of Technology Madras, Chennai, India. 326

***De novo* metagenome assembly using PacBio long reads**
Yoshihiko Suzuki, Junko Taniguchi, Jun Yoshimura, Kenshiro Oshima, Masahira Hattori, Shinichi Morishita.
Presenter affiliation: The University of Tokyo, Kashiwa, Japan. 327

Characterizing DNA methylation of living LINE/L1 transposons in human genomes using long SMRT reads
Yuta Suzuki, Shoji Tsuji, Shinichi Morishita.
Presenter affiliation: The University of Tokyo, Kashiwa, Chiba, Japan. 328

Nanopore sequencing for genotyping pathogens of tropical diseases
Yutaka Suzuki, Arthur E. Mongan, Josef Tuda, Junya Yamagishi.
Presenter affiliation: University of Tokyo, Kashiwa, Japan. 329

Aberrant pre-mRNA splicing due to mutations in *RNU4atac*, a minor spliceosomal snRNA, results in severe developmental phenotypes in new mouse models
David E. Symer, Dandan He, Jingfeng Li, Katherine A. Yates, Keiko Akagi, Christopher J. Hlynialuk, Zhengqiu Zhou, Xiaomei Meng, Yanqiang Wang, Chelsea A. Moherman, Tanvi V. Joshi, Huiling He, Albert de la Chapelle, Brad N. Bolon, Richard A. Padgett.
Presenter affiliation: Ohio State University, Columbus, Ohio. 330

The secrets of GWAS are written in the reads
Claes Wadelius, Marco Cavalli, Gang Pan, Helena Nord, Ola Wallerman, Emelie Wallen Arzt, Olof Berggren, Ingegerd Elvers, Maija-Lena Eloranta, Lars Rönnblom, Kerstin Lindblad Toh.
Presenter affiliation: Science for Life Laboratory, Uppsala , Sweden. 349

GO-PCA—An unsupervised method to explore biological heterogeneity based on gene expression and prior knowledge
Florian Wagner, Sandeep Dave.
Presenter affiliation: Duke University, Durham, North Carolina. 350

Whole-genome bisulfite sequencing of acute lymphoblastic leukemia cells
Per Wahlberg, Anders Lundmark, Jessica Nordlund, Stephan Busche, Erik Forestier, Gudmar Lönnerholm, Tomi Pastinen, Ann-Christine Syvänen.
Presenter affiliation: Uppsala University, Uppsala, Sweden. 351

Characterization of maize B73 transcriptome by hybrid sequencing
Bo Wang, Michael Regulski, Andrew Olson, Joshua Stein, Tyson Clark, Yinping Jiao, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 352

Genome-wide crossover distribution in male and female of maize
Minghui Wang, Qi Sun, Pawlowski Wojtek.
Presenter affiliation: Cornell University, ithaca, New York. 353

Principles of epigenome conservation
Jia Zhou, Xiaoyun Xing, Bo Zhang, Daofeng Li, Renee L. Sears, Nicole B. Rockweiler, Rebecca F. Lowdon, Hyung Joo Lee, Ting Wang.
Presenter affiliation: Washington University, St. Louis, Missouri. 354

Rapid, comprehensive, whole-genome interrogation of medical sequencing data
Barry Moore, Alistair Ward, Carson Holt, Shawn Rynearson, David Nix, Brett Milash, Aaron Quinlan, Mark Yandell, Gabor Marth.
Presenter affiliation: USTAR Center for Genetic Discovery, Salt Lake City, Utah. 355

Annotating non-genic regions in Ensembl
Daniel R. Zerbino, Nathan Johnson, Thomas Juettemann, Steven P. Wilder, David Richardson, Avik Datta, Laura Clarke, Paul R. Flicek.
Presenter affiliation: European Molecular Biology Laboratory, Hinxton, United Kingdom. 374

Dynamic DNA methylation change of transposable elements in human cancer
Bo Zhang, Rebecca Lowdon, Xiaoyun Xing, Daofeng Li, Joseph Costello, Ting Wang.
Presenter affiliation: Washington University School of Medicine, St. Louis, Missouri. 375

Genomic, epigenomic, and gene expression analysis reveals the connection between mutational pattern and lineage of B cell lymphomas
Jenny Zhang, Andrea Moffitt, Cassandra Love, Sandeep Dave.
Presenter affiliation: Duke University, Durham, North Carolina. 376

Novel IDEAS for detecting epigenetic variation in multiple human cell types
Yu Zhang, Marta Byrska-Bishop, Feng Yue, Ross C. Hardison.
Presenter affiliation: The Pennsylvania State University, University Park, Pennsylvania. 377

FRIDAY, May 8—4:30 PM

GUEST SPEAKERS

George Davey Smith
University of Bristol, UK

Francis S. Collins
National Institutes of Health

FRIDAY, May 8

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SESSION 11 TRANSLATIONAL GENOMICS AND GENETICS

Chairpersons: **B. Hayes,** Department of Environment and Primary Industries, Melbourne, Australia
D. Lo, Chinese University of Hong Kong

HOTSPOTS OF RECOMBINATION INITIATION IN THE MOUSE GENOME

Fatima Smagulova*[1], Kevin Brick*[2], R. Daniel Camerini-Otero[2], Galina V Petukhova[1]

[1]Uniformed Services University of the Health Sciences, Dept. of Biochemistry and Molecular Biology, Bethesda, MD, [2]National Institute of Diabetes, Digestive and Kidney Diseases, Genetics and Biochemistry Branch, Bethesda, MD

The vast majority of meiotic recombination takes place in recombination hotspots – regions of the genome with recombination frequency significantly above the frequency in adjacent regions. The location of recombination hotspots is defined by the meiosis-specific methyltransferase PRDM9, which has highly polymorphic DNA binding specificity. We have recently mapped the distribution of recombination hotspots in males of the two congenic mouse strains carrying different *Prdm9* alleles and have demonstrated that PRDM9 defines the location of at least 99% of hotspots. F1 hybrids of these strains formed hotspots derived from either one parent or another with preferentially the stronger parental hotspots being utilized. We have now extended our studies to additional mouse strains to examine how divergent genetic background affects the placement of recombination hotspots in hybrid strains. We found that up to 30% of hotspots in some crosses are new, i.e., are not present in either of the parental strains. We show that *Prdm9* alleles exhibit consistent patterns of incomplete dominance in their ability to define hotspot locations in the hybrids. Finally, we found that up to 7% of hotspots in the hybrids can be defined by *Prdm9*-independent mechanisms. We will present the analysis of our data and discuss the possible mechanisms that can explain our observations.

EVOLUTION OF GENE REGULATION IN 20 MAMMALS

Camille Berthelot*[1], Diego Villar*[2], Duncan T Odom[2], Paul Flicek[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom, [2]University of Cambridge, Cancer Research UK, Cambridge, United Kingdom

Despite long-standing efforts from the genomics community, our understanding of gene expression regulation in mammalian genomes remains incomplete. This is in a large part because unlike coding sequences, non-coding regulatory elements cannot be fully uncovered using sequence characteristics and conservation alone, so that traditional sequence-based approaches have proved inadequate to describe the regulatory network of mammalian genomes. Previous work on the evolutionary dynamics of transcription factor binding and histone modifications has shown that regulatory sequences seem to be fast-evolving and exhibit high turn-over rates. However, the underlying properties that rule these dynamics and their consequences on gene expression are poorly understood. We report here the results of a large-scale functional genomics study in liver tissue from 20 species spanning the breadth of the mammalian phylogenetic tree from monotremes to primates, including many species that have not been previously studied with genome-wide methods such as cetaceans. We used ChIP-seq to produce a high-resolution description of the genomic landscape marked by two types of histone modifications: H3K4me3 (associated with active promoters) and H3K27ac (associated with active promoters and enhancers), as well as RNA-seq to document gene expression levels. The project explores the conservation, birth, loss and turn-over of promoters and enhancers and their consequences on gene expression in mammalian liver. We observe that while the total number of active sites remains largely unchanged over mammalian evolution, the location of these sites and especially of active enhancers has evolved markedly. Using statistical models, we describe the rates and patterns of evolutionary divergence of these regulatory regions as well as the emergence of new regulatory elements. We then investigate the genomic properties and functional annotations that underlie these evolutionary dynamics in an effort to connect regulatory evolution and function. Lastly, we explore how changes in local regulatory context affect gene expression, and how evolutionary maintenance and loss can be used to infer regulatory relationships between enhancers and target genes.

* Equal contribution

SPATIAL TRANSCRIPTOMICS – A METHOD FOR GENE EXPRESSION ANALYSIS OF MULTIPLE REGIONS WITHIN WHOLE TISSUE SECTIONS

Fredrik Salmén*[1], Patrik Ståhl*[2], Sanja Vickovic[1], Anna Lundmark[3], Stefania Giacomello[1], José Fernandez[2], Michaela Asp[1], Emelie Berglund[1], Annelie Mollbrink[1], Pelin Sahlén[1], Jens Magnusson[2],Joel Sjöstrand[2], Erik Sjölund[1], Mikael Huss[4], Jakub Westholm[4], Jonas Frisén[2], Joakim Lundeberg[1]

[1]Science for Life Laboratory/Royal Institute of Technology (KTH), Division of Gene Technology, Solna, Sweden, [2]Karolinska Institutet (KI), Department of Cell and Molecular Biology, Stockholm, Sweden, [3]Karolinska Institutet (KI), Department of Dental Medicine, Huddinge, Sweden, [4]Science for Life Laboratory, Bioinformatics Long-term Support (WABI), Solna, Sweden

Here we present a novel method, Spatial Transcriptomics (ST), which combines gene expression analysis of multiple regions in intact tissue sections, with staining and imaging of the same sections. The protocol starts by placing a thin tissue section onto a glass slide, containing several thousands of small features. Each feature is built up by millions of nucleic acid capture probes, which all contain a specific sequence. The specific sequence is shared by all probes within a feature but varied between features. We refer to this sequence as a spatial barcode, something that is used to determine the location of each feature. The tissue section is stained and imaged in order to get the morphological information and to determine the location of each tissue region, in relation to the features. The imaging step is followed by a specific tissue treatment and complementary DNA (cDNA) synthesis. This enables the generation of cDNA libraries, where each transcript is attached to a capture probe. The finished cDNA libraries are sequenced using paired end sequencing, allowing us to get the spatial barcode sequence from one end and the transcript sequence from the other end. Since the location of all features and tissue regions are known, the spatial barcode sequence can be used to determine which region each transcript was derived from. Finally, the spatially located transcripts can be visualized in our software tool by combining the analyzed sequence data with the tissue image. The result is a high-resolution pattern of gene expression information across the analyzed tissue section, where the resolution is limited only by the size of the underlying features.

THOUSANDS OF NOVEL TRANSLATED OPEN READING FRAMES AND DUALLY CODED REGIONS ACCURATELY INFERRED USING RIBOSOME FOOTPRINTING DATA.

Anil Raj[1], Sidney Wang[2], Heejung Shim[2], Yang Li[1], Matthew Stephens[2,3], Yoav Gilad[2], Jonathan K Pritchard[1,4,5]
[1]Stanford University, Genetics, Stanford, CA, [2]University of Chicago, Human Genetics, Chicago, IL, [3]University of Chicago, Statistics, Chicago, IL, [4]Stanford University, Biology, Stanford, CA, [5]Stanford University, Howard Hughes Medical Institute, Stanford, CA

Understanding the functional effects of gene expression critically depends on the accurate and comprehensive annotation of sequence elements that are translated in each gene. Traditionally, phylogenetic approaches have been used to estimate coding potential in transcribed regions conserved across species. More recently, mass-spectrometry (MS) based proteomic assays have been used to directly identify and quantify highly expressed proteins in a cell-specific manner. Proteomic assays, however, have poor coverage to be useful for accurately annotating the translation start and end sites in a given transcript. In contrast, ribosome footprinting provides an unbiased, high coverage measurement of translation in a cell type-specific manner. Although ribosome footprints have been used to identify translated transcripts, these data have not been used to elucidate the precise set of open reading frames (ORF) that are being translated in each transcript in a given cell type.

In this work, we used hidden Markov models to analyze ribosome footprinting data along with sequence information in the cell-specific transcriptome. Our analyses accurately resolved the ORFs that are being translated in human lymphoblastoid cell lines (LCL). We identified thousands of transcripts with previously unannotated translated ORFs, including hundreds of short translated ORFs and hundreds of transcripts with dually coded translated regions. Nearly 70% of previously undiscovered translated regions predicted to have a peptide match were validated using MS data generated in the same cell line and using peptides identified from MS data in related cell types available in public proteomics databases. Approximately 90% of these validated novel translated regions consisted of processed pseudogenes and dually coded transcripts, and a substantial number of these regions are strongly conserved in vertebrates. Finally, using ribosome footprinting data in LCLs from a panel of 72 Yoruba individuals, we find strong evidence for inter-individual variation in the translation of the pair of ORFs for most dually coded transcripts, suggesting regulatory control of ORF usage. The application of our method to ribosome footprinting data across cell types and organisms may prove useful in teasing apart functional and evolutionary differences in gene expression between tissues and species.

A 3D MAP OF THE HUMAN GENOME AT KILOBASE RESOLUTION REVEALS PRINCIPLES OF CHROMATIN LOOPING

Miriam H Huntley*[1,2,3], Suhas S Rao*[1,2], Neva C Durand[1,2], Elena K Stamenova[1,2], Ivan D Bochkov[1], James T Robinson[1,2], Adrian L Sanborn[1], Ido Machol[1], Arina D Omer[1], Eric S Lander[2], Erez Lieberman Aiden[1,2]

[1]Baylor College of Medicine, The Center for Genome Architecture, Houston, TX, [2]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA, [3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

*These authors contributed equally.

We use in situ Hi-C to probe the three-dimensional architecture of genomes, constructing maps of ten cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1-kilobase resolution. We find that genomes are partitioned into domains (median length, 185kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs ‘facing’ one another.

This work was funded by NSF grants DGE0946799 and DGE1144152, an NIH New Innovator award (1DP2OD008540-01), an NIH CEGS (P50HG006193), an NVIDIA Research Center award, an IBM University Challenge Award, a Google Research Award, a Cancer Prevention Research Institute of Texas Scholar Award (R1304), a McNair Medical Institute Scholar Award, the President’s Early Career Award in Science and Engineering, and an NHGRI grant (HG003067).

SUPER-RESOLUTION IMAGING OF CHROMATIN NANO-STRUCTURE REVEALS TIGHT COUPLING OF EPIGENETIC STATE AND 3D GENOME ORGANIZATION

Alistair N Boettiger[1,2], Bogdan Bintu[2], Jeffrey Moffitt[1,2], Brian Beliveau[3], Chao-ting Wu[3], Xiaowei Zhuang[1,2,4]

[1]Harvard University, Chemistry and Chemical Biology, Cambridge, MA, [2]Harvard University, Physics, Cambridge, MA, [3]Harvard Medical School, Genetics, Boston, MA, [4]Howard Hughes Medical Institute, Investigator, Cambridge, MA

The spatial organization of the genome has the potential to affect many aspects of gene expression; the relative locations of genes and regulatory sequences in three-dimensional (3D) space strongly affects the probability these sequences can interact; chromatin packaging may occlude access of large regulatory proteins to the DNA sequence and unfolding of domains may facilitate binding; some chromatin structures may serve as scaffolds to recruit regulatory complexes. Unfortunately, current methods provide little information about 3D organization of the genome at the length scale of genes (kilobases) and regulatory domains (hundreds of kilobases) in single cells. I will present a new super-resolution imaging approach to study the structural organization of the genome at the kilobase to megabase scale in individual cells at 20 nm resolution. These domains largely occupy diffraction limited volumes and thus their structures cannot be resolved with conventional imaging approaches. From thousands of images of dozens of epigenetic domains from across the Drosophila genome, we have discovered, within a single cell type, a substantial diversity of structural patterns: compact and diffuse domains, branched and linear domains, domains that are highly entangled with one-another and domains which are strictly segregated. These different structural features are closely correlated to certain differences in the epigenetic state of the chromatin. I will focus on the organization of Polycomb bound domains, which exhibit a surprising, entangled structure and length-dependent compaction. Computational models suggest this organization could contribute to the repressive nature of the domains. This work suggests further super-resolution imaging studies of chromatin structure at this scale may greatly aid our understanding of the 3D genome.

ANALYSIS OF RNA DECAY FACTOR MEDIATED RNA STABILITY CONTRIBUTIONS ON THE RNA ABUNDANCE

Sho Maekawa[1], Naoto Imamachi[2], Takuma Irie[1], Hidenori Tani[3], Kyoko Matsumoto[1], Rena Mizutani[2], Katsutoshi Imamura[2], Miho Kakeda[2], Tetsushi Yada[4], Sumio Sugano[1], Yutaka Suzuki[1], Nobuyoshi Akimitsu[2]

[1]The University of Tokyo, Graduate School of Frontier Sciences, Kashiwa, Japan, [2]The University of Tokyo, Radioisotope Center, Tokyo, Japan, [3]National Institute of Advanced Industrial Science and Technology, Research Institute for Environmental Management Technology, Tsukuba, Japan, [4]Kyushu Institute of Technology, Department of Bioscience and Bioinformatics, Iizuka, Japan

Recent large-scale genomic project have collected a large number of histone epigenome data, where they could infer the eventual RNA expression levels to certain level. Chromatin states, measured by chromatin immunoprecipitation sequencing (ChIP-seq) and RNA expression, measured by RNA sequencing (RNA-seq) does not necessarily correlate with each other. One of the components of this inconsistency may arise from the variation in RNA stability, where RNA that are more stable may accumulate more RNA than predicted by ChIP-seq data, and vice versa. To test the hypothesis, we used transcriptome-wide stability data generated by 5'-bromouridine immunoprecipitation chase sequencing (BRIC-seq), allowing a better estimation on the eventual RNA expression levels and better understanding of into the importance of post-transcriptional regulation in determining the RNA expression levels.

As postulated, we identified discrepancies between RNA abundance and histone H3 lysine 4 tri-methylation (H3K4me3) ChIP-seq intensity, a major active chromatin mark, in HeLa cells. we identified 865 genes where their RNA expression was predominantly controlled at the level of RNA stability. Additionally, ENCODE and DBTSS data analyses supported the proposal that the RNA stability control aids to determine transcript levels in multiple cell types. We analysed the contributions of UPF1, EXOSC5 and STAU1, RNA degradation factors, in determining the RNA abundance and we found that RNA abundance of 8% of genes were attributable to these three factors. Furthermore, we propose a feedback control circuit consisting of regulated degradation of mRNA encoding transcription factors maintaining steady-state level of RNA abundance. Moreover, these regulatory mechanisms seem to vary between mRNA and lincRNA.

Integrative analysis of ChIP-seq, RNA-seq and my BRIC-seq showed that transcriptional regulation and RNA degradation are independently regulated. In addition, RNA stability is an important determinant of eventual transcript levels. RNA binding proteins, such as UPF1, EXOSC5 and STAU1 may play active roles in such controls.

THE VERSATILITY OF CIRCULATING TUMOUR CELLS IN LUNG CANCER - BIOMARKERS, BIOLOGY AND MOUSE MODELS

Caroline Dive[1], Ged Brady[1], Christopher Morrow[1], Fiona H Blackhall[1,2,3], Cassandra Hodgkinson[1], Crispin Miller[1], Kathryn Simpson[1], Dominic Rothwell[1], Francesca Trappani[1], Robert Metcalf[1,2], Louise Carter[1,2]

[1]University of Manchester, Cancer Research UK Manchester Institute, Manchester, United Kingdom, [2]Christie Hospital NHS Foundation Trust, Oncology, Manchester, United Kingdom, [3]UNiversity of Manchester, Institute of Cancer Sciences, Manchester, United Kingdom

Minimally invasive biomarkers to stratify and monitor patients receiving targeted therapies have potential benefit. Routine monitoring of ctDNA with improving sensitivity edge closer as companion diagnostics. Remaining challenges lie with optimising sample collection and method standardisation. What then is the added utility of CTCs? The rarity and heterogeneity of CTCs make them technically demanding but they hold great promise. Marker independent technology platforms in development better assess CTC subpopulations but none are fully validated for clinical utility. CTC enumeration with CellSearch (EpCam and CK positive CTCs) has prognostic significance in many epithelial tumours. In diseases with prevalent CTCs such as small cell lung cancer (SCLC), the dynamic range of CTC number allows pharmacodynamic evaluation. Multiplex protein analysis and FISH are also possible if sufficient CTCs are detected. Single CTC isolation and DNA profiling allow insights to the genomic landscape of disseminating tumour cells. Transcriptomics of single cells is now feasible. We are molecular profiling single SCLC CTCs to inform on tumour heterogeneity and mechanisms of drug resistance.

We developed lung cancer patient CTC derived mouse models (termed CDX). SCLC CDX models faithfully recapitulate patient drug responses. Paired CDX can be generated at patient presentation and again at relapse. CDX allow comprehensive analysis of acquired drug resistance, the discovery of new drug targets and testing of targeted therapies. Viable disaggregation of CDX tumours coupled with removal of dead cells and contaminant mouse cells allows assessment of how many CDX derived cells are needed to regrow a CDX and whether regrowth is from a distinct sub-population which may correspond to human SCLC cells with 'stem-like' characteristics.

CTCs in NSCLC are less easy to find, <35% Stage IV NSCLC patients have CellSearch detetected CTCs, although their number is prognostic. Marker independent technolgoies however, reveal increased numbers and heterogeneity in EMT phenotypes. I will present a case report of a NSCLC patient whose blood sample contained no CellSearch CTCs yet generated a CDX. CTC filtration revealed a high proportion of mesenchymal CTCs consistent with EMT.

UNDERSTANDING CARDIAC STRUCTURE AND FUNCTION IN HUMANS USING 4D IMAGING GENETICS.

Hannah V Meyer[1], Antonio De Marvao[2], Timothy J Dawes[2], Wenzhe Shi[2], Tamara Diamond[2], Daniel Rueckert[2], Enrico Petretto[2], Leonardo Bottole[2], Declan P O'Regan[2], Stuart A Cook[2], Ewan Birney[1]

[1]European Bioinformatics Institute (EMBL-EBI), Birney Research Group, Cambridge, United Kingdom, [2]Imperial College London, MRC Clinical Sciences Centre, London, United Kingdom

Human health is dependent on the long lasting function of many organ systems; these in turn develop due to complex genetic programs and are maintained over a lifespan. Many human diseases are related to cardiac structure and function, from relatively common cardiac infarctions through to more rare but serious diseases such as different cardiomyopathies. Understanding the biology of the human heart is informative for both basic and translational research.

We have created the first at scale cohort of 1,500 detailed cardiac images from healthy volunteers. We used a 1.5T Philips MRI scanner to acquire detailed 4D images of the heart in a single breath hold. This provides a far more detailed and consistent cardiac measurement than the traditional combination of 2D planar cardiac images. We are able to map all these 4D images into a consistent volumetric reference, and derive over 27,000 measurements per individual representing the heart. The individuals were also genotyped on a modern SNP array and imputed using a combination of 1000 Genomes and UK10K known variants, leading to 9.4 million variants for use in association studies.

A major challenge in imaging genetics is handling the large number of correlated dimensions present in these images, even when placed in a common reference framework. We used Bayesian latent factor analysis to project this high dimensional phenotype space into a lower dimensional factor space as a representation of the underlying structure of the heart. Using this lower-dimensional projection, we are able to find a large number of genetic loci which show strong association with the heart structure; this is in contrast to the traditional single scalar measures of cardiac structure commonly in use in cardiac research such a left ventricular mass. Interestingly, some of the discovered cardiac components correspond to known features of the human heart. We explore the relationship of the genetics of these cardiac structures to large scale physiological measurements such as blood pressure.

This work shows that imaging genetics provides an invaluable, unbiased discovery process for exploring the underlying biology of human organs, with an impact on our understanding of both healthy and disease physiology.

ANALYSIS OF THE GENETIC PHYLOGENY OF MULTIFOCAL PROSTATE CANCER IDENTIFIES MULTIPLE INDEPENDENT CLONAL EXPANSIONS IN NEOPLASTIC AND MORPHOLOGICALLY NORMAL PROSTATE TISSUE

Colin S Cooper*[1,2], Rosalind Eeles*[1,3], David C Wedge*[4], Peter Van Loo*[4,5,6], Gunes Gundem[4], Ludmil B Alexandrov[4], Barbara Kremeyer[4], Andrew G Lynch[7], Adam Butler[4], Charlie E Massie[8], Jorge Zamora[4], Vincent Gnanapragasam[8], Anne Y Warren*[9], Christopher S Foster*[10], Hayley C Whitaker*[8], Ultan McDermott*[4], Daniel S Brewer*[1,2,11], David E Neal*[9,12]

[1]The Institute Of Cancer Research, Division of Genetics and Epidemiology, London, United Kingdom, [2]University of East Anglia, Norwich Medical School, Norwich, United Kingdom, [3]Royal Marsden NHS Foundation Trust, London, United Kingdom, [4]Wellcome Trust Sanger Institute, Cancer Genome Project, Hinxton, United Kingdom, [5]VIB and KU Leuven, Department of Human Genetics, Leuven, Belgium, [6]Cancer Research UK London Research Institute, London, United Kingdom, [7]Cancer Research UK Cambridge Research Institute, Statistics and Computational Biology Laboratory, Cambridge, United Kingdom, [8]Cancer Research UK Cambridge Research Institute, Urological Research Laboratory, Cambridge, United Kingdom, [9]Cambridge University Hospitals NHS Foundation Trust, Department of HistopathologyCambridge, United Kingdom, [10]HCA Pathology Laboratories, London, United Kingdom, [11]The Genome Analysis Centre, Norwich, United Kingdom, [12]University of Cambridge, Department of Surgical Oncology, Cambridge, United Kingdom

Whole genome DNA sequencing was used to decrypt the phylogeny of multiple samples from distinct areas of cancer and morphologically normal tissue taken from the prostates of three men. Mutations were present at high levels in morphologically normal tissue distant from the cancer reflecting clonal expansions, and the underlying mutational processes at work in morphologically normal tissue were also at work in cancer. Our observations demonstrate the existence of on-going abnormal mutational processes, consistent with field-effects, underlying carcinogenesis. This mechanism gives rise to extensive branching evolution and cancer clone mixing as exemplified by the coexistence of multiple cancer lineages harboring distinct *ERG* fusions within a single cancer nodule. Subsets of mutations were shared either by morphologically normal and malignant tissue or between different *ERG*-lineages, indicating earlier or separate clonal cell expansions. Our observations inform on the origin of multifocal disease and have implications for prostate cancer therapy in individual cases.

* authors contributed equally

INTEGRATIVE GENOMICS WITH EXOME, TRANSCRIPTOME AND WHOLE GENOME SEQUENCING OF HUMAN AND MURINE T CELL LYMPHOMAS REVEAL NOVEL SUBTYPES ASSOCIATED WITH CLINICAL OUTCOME

Andrea B Moffitt, Matthew McKinney, Cassandra Love, Jenny Zhang, Jyotishka Datta, Sandeep S Dave

Duke University, Center for Genomics and Computational Biology, Durham, NC

Introduction:
Hepatosplenic T cell lymphoma (HSTL) is an aggressive form of non-Hodgkin lymphoma that is characterized by a young age of onset and an unusually dismal prognosis. The most frequent known genetic abnormality in HSTL is isochromosome 7q (iso7q), but the molecular implications and clinical significance of this abnormality are unknown. Studies in other T cell lymphomas have revealed their genetic landscape to be comprised of novel gene mutations including RHOA and TET2. Similar investigations into the role of mutations in HSTL have been lacking. Additionally, a mouse model of HSTL appears to recapitulate key aspects of human tumors, but the degree to which these mouse tumors genetically overlap with human HSTLs is unknown.
Methods:
In this study, we sought to define the genetic features of HSTL through exome sequencing of 30 HSTL tumors and germline DNA from the same patients, as well as several murine HSTL tumors. Low coverage whole genome sequencing and transcriptome sequencing provided further characterization the tumors. Candidate HSTL driver genes were identified based on the characteristics of known cancer genes, including frequency of non-synonymous events in HSTL, rate of variation in healthy controls and predicted functional impact of the mutation. Cox proportional hazards regression was used to model survival associations between clinical covariates such as age and gender, as well as molecular covariates.
Results:
The histone methyltransferase SETD2 was found to be the most frequently mutated gene in HSTL, with predominantly frameshift and nonsense mutations. In addition, HSTLs were characterized by frequent mutations in STAT5B, SMARCA2, PIK3CD and TET1. Interestingly, we found the mutational profile of HSTL to be distinct from other T cell lymphomas. Copy number analysis identified recurrent chromosomal alterations, and transcriptome analysis showed key pathways, such as DNA repair, to be correlated with recurrent mutations. Our investigation of the only known mouse model of HSTL found significant overlap between human and murine disease in mutations and pathways related to chromatin modification, signal transduction, DNA repair and cell cycle progression. Finally, an integrated analysis of genetic features with clinical data found that iso7q was associated with a poorer prognosis in HSTL patients while SETD2 mutations suggest better outcomes. Our data thus describes in detail the genetic landscape of HSTL and implicate a number of novel genes and molecular pathways in this disorder.

THE SPECTRUM OF HUMAN DISEASE MUTATIONS IN > 5,000 CLINICAL EXOME SEQUENCING CASES

Christine M Eng[1], Yang Yaping[1], Sharon E Plon[1,2], Donna M Muzny[3], James R Lupski[1,2,3], Arthur L Beaudet[1], Eric Boerwinkle[4], Richard A Gibbs[3]

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, [2]Baylor College of Medicine, Department of Pediatrics, Houston, TX, [3]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [4]University of Texas, Human Genetics Center, Houston, TX

The BCM CLIA/CAP certified clinical sequencing laboratory (Whole Genome Laboratory: WGL), as of 01/31/2015, has analyzed more than 5,000 cases of childhood disease by whole exome sequencing (WES). As methods improve in the research laboratory we progressively adapt them for clinical application: the current procedures include the deployment of rapid exome hybridization, more complete gene coverage and robust routines to identify deleted exons in critical genes. The 'pretests' currently include DNA Capture/Pacific Biosystems combinations and whole genome sequencing. Distributed 'cloud' computing methods that are robustly functioning for research applications are also being applied to this diagnostic application. We previously reported an ~25% 'solution rate' including: multiple instances of rare events in this cohort; uniparental disomy; and the role of copy number variants. We also now confirm the prevalence of phenotypes confounded by mutations at multiple loci as ~4.5% are revealed to be individuals harboring two variants that otherwise would each be predicted to cause disease. We also identify multiple cases of de novo variation that lead to disease and many examples where we simultaneously identify 'new' Mendelian loci while improving the diagnostic rate. This clinical application effort is now the 'hub' of Mendelian discovery and so the complete model represents both the demonstration of a full translation of genomics into the clinic, and integration of research and clinical agendas.

SINGLE CELL PORTRAITS OF BREAST CANCER HETEROGENEITY

Timour Baslan[1,2], James Hicks[1]

[1]Cold Spring Harbor Laboratory, Cancer Center, Cold Spring Harbor, NY, [2]Memorial Sloan Kettering Cancer Center, Program in Cancer Biology and Genetics, New York, NY

The fundamental operative unit of a cancer is the genetically innovative single cell. Whether in a detectably growing tumor, or as quiescent cells disseminated elsewhere in the body, single cells govern the parameters that dictate the biology of tumors. Thus, genome-wide single-cell analysis provides the ultimate level of resolution and consequently, holds enormous potential to further our understanding of tumor biology. Recent advances in whole genome amplification technologies, next generation sequencing, and single cell data analysis have enabled such investigations. Here, we will present on recent methodological developments, on both the bench side as well as the algorithmic side, that enabled the profiling and analysis of thousands of single cell genomes from a cohort of breast cancer patient samples representing all gene expression subtypes. We will discuss, among other observations, (1) genetic heterogeneity in normal cells present within biopsies (ex. Infiltrating T-cells), (2) the landscape of clonal/sub-clonal copy number variantion, (3) the sub-population structure and phylogeny of sub-clones, and (4) the association of genetic heterogeneity with clinical parameters.

SINGLE CELL ALLELE-SPECIFIC EXPRESSION (ASE) IN TRISOMY 21

S E Antonarakis[1], G Stamoulis[1], C Borel[1], F Santoni[1], A Letourneau[1], P Makrythanasis[1], Michel Guipponi[2]

[1]University of Geneva, Genetic Medicine, Geneva, Switzerland, [2]University Hospitals of Geneva, Genetic Medicine, Geneva, Switzerland

Trisomy 21 is a model disorder (aneuploidy) of altered gene expression. We have previously used a pair of monozygotic twins discordant for T21 to study the global dysregulation of gene expression, without the noise of the genetic variation among individuals (Nature:508; 345-350; 2014). Studies on gene and allelic expression in single cells, may reveal important biological insights regarding the cellular impact of aneuploidy and elucidate the fundamental mechanisms of gene dosage. We employed here allele specific expression (ASE as described in AJHG 96: 70-80, 2015) using RNASeq from 352 single cell fibroblasts (172 Normal and 180 T21 cells) from the pair of monozygotic twins discordant for T21. A considerable number of heterozygous sites throughout the non-chr21 genome were expressed monoallelically (Normal: 73.2% monoallelic in 559,134 observations, and T21: 78.8% monoallelic in 573,670 observations). There was also considerable monoallelic expression for chr21 genes in Normal and surprisingly in T21 cells as well (Normal: 67.2% monoallelic in 4,985 observations, and T21: 76.1% monoallelic in 6,723 observations). This metric was used to classify expressed genes on chr21 according to the level of monoallelic expression (9 monoallelic, 29 intermediate, 2 biallelic). We hypothesize that different classes of genes contribute with different dosage mechanisms to the phenotypic variability of DS. Furthermore we have made a preliminary observation that genomewide the T21 cells showed more monoallelic expression than the normal cells, but more analysis is needed to confirm these results. This study provides a fundamental understanding of the allele specific expression behavior of T21 single cells.

COMPONENTS OF BREAST CANCER HERITABILITY IN A MULTI-ETHNIC TARGETED SEQUENCING STUDY

Akweley Ablorh[1,2], Alexander Gusev[1,2], Brad Chapman[3,4], Gary Chen[5], Constance Chen[1,2], Sara Lindstroem[1,2], Brian E Henderson[5], Loic Le Marchand[6], Oliver Hofmann[3,4], Christopher A Haiman[5], Peter Kraft[1,2,3], Alkes Price[1,2,3]

[1]Harvard T.H. Chan School of Public Health, Program in Genetic Epidemiology, Boston, MA, [2]Harvard T.H. Chan School of Public Health, Epidemiology, Boston, MA, [3]Harvard T.H. Chan School of Public Health, Biostatistics, Boston, MA, [4]Harvard T.H. Chan School of Public Health, HSPH Bioinformatics Core, Boston, MA, [5]Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Preventive Medicine, Los Angeles, CA, [6]University of Hawai'i, Cancer Center, Honolulu, HI

Much of the heritability of sporadic breast cancer (BC) remains unexplained despite dozens of common, reproducible and genome-wide significant single nucleotide variant (SNV) associations. Variance component methods may explain more trait heritability than top genome-wide association SNV associations alone. We applied variance component methods that consider all SNVs together to determine what proportion of BC heritability can be explained by common and rare variants at 12 GWAS-identified BC loci that we selected for targeted sequencing in 2,316 cases and 2,295 controls from four self-reported ethnicities. In each ethnicity, we partitioned the heritability explained by each variant class for GWAS index SNVs, broadly-defined DNase Hypersensitivity (DHS), promoter, or coding variants. We also explored the relative contribution of rare and common variation to determine the main refuge of hidden heritability at twelve BC-associated loci.

GINKGO: UNCOVERING COPY-NUMBER VARIATIONS IN SINGLE-CELL SEQUENCING DATA

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

Cold Spring Harbor Laboratory, Quantative 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, neuronal mosaicism, and gamete development. 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 that we developed for the analysis and quality assessment of single-cell copy-number alterations. Ginkgo automates and standardizes the computation required to go from raw reads to copy-number profiles of individual cells, to trees of entire populations. 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/WGA4) through comparative analysis of 9 different single-cell datasets. We find that both MALBAC and DOP-PCR vastly outperform MDA in terms of data quality. As previously reported, MDA displays poor coverage uniformity and low signal-to-noise ratios. Coupled with overwhelming GC biases, MDA is unreliable for accurately determining CNVs compared to the other two techniques. Furthermore, we show that while both DOP-PCR and MALBAC data can be used to generate CNV profiles and identify large variants, DOP-PCR data has substantially 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, and is more reliable for accurate absolute copy-number calls.

SPECTRUM OF SOMATIC VARIATIONS IN HEALTHY SKIN FIBROBLASTS

Alexej Abyzov[1], Livia Tomasini[2,3], Jessica Mariani[2,3], Mariangela Amenduni[2,3], Anahita Amiri[2,3], Flora M Vaccarino[2,3,4]

[1]Mayo Clinic, Department of Health Sciences Research, Rochester, MN, [2]Yale University, Program in Neurodevelopment and Regeneration, New Haven, CT, [3]Yale University, Child Study Center, New Haven, CT, [4]Yale University, Department of Neurobiology, New Haven, CT

Multiple studies have been performed on the analysis of somatic genomic alterations in cancer, but only a few have been conducted to understand natural somatic mosaicism, that is post-zygotic accumulation of mutations in cells of human body. Fundamental knowledge about somatic mosaicism is not only crucial for finding determinants of cancer development and progression, but also for an understanding of various diseases and aging. We have compared genomes of over ten clonally derived human induced pluripotent stem cell (hiPSC) lines to the genomes of four primary skin fibroblast samples, parental to the hiPSC lines. The clonal nature of hiPSC lines allows the discovery of somatic genomic variants present in the founder cell, but not in all fibroblast cells, thereby providing a mean for a high-resolution analysis of single cell genomes. With this approach we found that, on average, an iPSC line manifests 400 single nucleotide variants (SNVs) not apparent in the parental fibroblasts. We next performed an in-depth re-evaluation of these candidate somatic variants in the fibroblasts, by applying the digital droplet PCR technique as well as by conducting ultra-deep sequencing coupled with a precise estimation of sequencing errors. These experiments confirmed that at least 20%, and up to 70%, of the candidate SNVs are mosaic somatic variants in fibroblast cells. Thus, our analysis revealed that a single human skin fibroblast cell has from 80 up to 270 somatic SNVs. The allele frequency of these SNVs ranged from 0.1% to a dozen percent in the fibroblast cell population, and the mutation spectrum was surprisingly similar to that observed in some cancers (Alexandrov et al., Nature, 2013; Lawrence et al., Nature, 2013). These new discoveries emphasize a large degree of somatic mosaicism existing in healthy human tissues, and provide the first evidence that mutational signatures observed in cancers could be attributed to a background somatic mosaicism in normal cells.

BENCHMARKING OF SPLICE ISOFORM QUANTIFICATION METHODS FOR RNA SEQUENCE DATA

Francois Aguet[1], David S DeLuca[1], Tim Sullivan[1], The GTEx Project Consortium -[2], Gad Getz[1,3], Kristin Ardlie[1]

[1]Broad Institute, Cambridge, MA, [2]The Genotype-Tissue Expression (GTEx) Project Consortium, DC, [3]Massachusetts General Hospital, Department of Pathology, Boston, MA

A majority of human genes, including ~80% of protein coding genes, undergo alternative splicing to produce functionally distinct transcripts. How this process is regulated at a transcriptome-wide scale, and how genetic variants affect the expression of specific splice isoforms remains poorly understood due to challenges in accurately quantifying transcript abundance. While RNA-seq can in principle capture the full range of expressed transcripts in a sample, several factors limit its accuracy in practice, including small fragment sizes, sampling noise, RNA degradation, and reduced coverage for low-abundance transcripts. Several approaches for the deconvolution of isoform abundance from RNA-seq reads have been proposed (e.g., Cufflinks, Flux Capacitor, eXpress, RSEM, Sailfish/Salmon), with different models to correct for error-inducing biases. Comparisons between these approaches often produce diverging results, and more objective evaluations are hindered by the absence of ground truth data.

RNA-seq data from over 40 human tissues generated by the GTEx Project Consortium provide a unique opportunity to assess alternative splicing across the human transcriptome and to identify quantitative trait loci that modulate alternative splicing. To this end, high confidence in the estimates of transcript isoform abundance is critical. Here, we present a rigorous framework for the evaluation of isoform quantification methods, based on a combination of simulations, partial ground truth derived from a subset of long read (2x250bp) data from GTEx, and measures of isoform complexity, and apply this to the evaluation of the aforementioned methods.

A STATISTICAL FRAMEWORK FOR MODELING GENETIC DATA AS HAPLOTYPE CLUSTER GRAPHS WITH APPLICATION TO HAPLOTYPE PHASING, ASSOCIATION MAPPING, AND WHOLE-GENOME COMPRESSION

Derek C Aguiar[1], Lloyd T Elliott[2], Yee Whye Teh[2], Barbara E Engelhardt[1]

[1]Princeton University, Department of Computer Science, Princeton, NJ,
[2]University of Oxford, Department of Statistics, Oxford, United Kingdom

Haplotype-based analyses, including missing data imputation, multilocus association mapping, and inferring evolutionary history, suffer from two problems that reduce their efficacy in genomic research. First, haplotypes are often not known precisely: experimental methods that type haplotypes are unscalable and expensive, whereas computational approaches trade accuracy for efficiency, use complex models that are intractable for large samples, or use overly inexpressive models. Second, haplotype-based analyses often are agnostic to the particular haplotype sequences but instead use them as a proxy for the latent genealogy. These haplotype-based analyses would greatly benefit from a representation that captures the rich coalescent ancestry of the sample in a computationally tractable framework. Here, we present a statistical framework based on Bayesian nonparametric fragmentation-coagulation processes (FCP) that explicitly models coalescent processes and haplotype demography, improving on the PHASE model by adding exchangeability (Stephens *et al.* 2001), and has the computational tractability of HMM-based models such as SHAPEIT (Delaneau *et al.* 2012). Our method, *FCP-Phase*, improves on previous methods by (1) estimating the haplotype clustering at each locus with an expressive Bayesian nonparametric model that, at each locus, has the marginal structure of a coalescent process, (2) using reference haplotype data with an exchangeable model, (3) producing a biologically useful inferred latent structure of the evolutionary relationships between haplotypes, namely a *haplotype cluster graph*, where haplotype blocks may overlap, and (4) exploiting the condensed latent structure to make inference of haplotype phase, reconstruction of ancestry, and multilocus association efficient. In particular, FCP-Phase is able to use a Bayesian nonparametric model that captures the full complexity of haplotype structure among genomic samples by inferring this haplotype cluster graph a priori from high quality reference data; subsequently, new genomic samples can be efficiently phased using the existing haplotype cluster graph. We show applications of FCP-phase to phasing and imputation in the 1000 Genomes data, compare against the leading haplotype phasing and imputation software, and demonstrate how the haplotype cluster graph provides an invaluable data structure for genomics research through applications to association mapping, inferring genealogy, and whole-genome sample compression.

EXPRESSION AND EQTL MAPPING OF HLA GENES IN LARGE-SCALE RNASEQ ASSAYS

Vitor C Aguiar[1], Jonatas E Cesar[1], Emmanouil T Dermitzakis[2], Diogo Meyer[1]

[1]University of Sao Paulo, Dept. of Genetics and Evolutionary Biology, Sao Paulo, Brazil, [2]University of Geneva, Dept. of Genetic Medicine and Development, Geneva, Switzerland

The HLA genes (Human Leukocyte Antigens) are well-documented targets of balancing selection, and variation at these loci is associated to many disease phenotypes. Variation in expression levels also influences disease susceptibility and resistance, but relatively little information exists about regulatory variation and population-level expression patterns of HLA genes. The paucity of information stems from the difficulty in mapping short reads to these highly polymorphic loci, and in accounting for the existence of several paralogues in the HLA family. In this study we develop a computational pipeline to accurately estimate expression for five classical HLA genes based on RNAseq, and to use this data to map eQTLs.
Our method initially aligns reads to an index containing all known HLA sequences. We use available information on the individual's HLA genotype to parse mapped reads and assign them to each of the alleles the individual carries. For the cases in which reads map to multiple alleles or loci (multireads), we use an Expectation-Maximization algorithm to estimate the contribution from that read to the expression of each locus or allele. We applied our pipeline to 300 European individuals with whole-transcriptome quantification of expression, made available from the Geuvadis Consortium. Due to the lack of an external gold standard we carried out two forms of internal validation. First, we assessed the correlation between the proportion of multireads and the degree of differentiation among the two alleles which the individual carries, and this correlation was significant, as expected ($r = -0.96$). Second, co-expression patterns between the 5 HLA genes were strong and significant, and more intense between loci which are physically close ($r = 0.40$-0.90), as previously documented. We then compared our quantifications to those obtained using the standard approach of mapping reads to the reference genome, and found a weak correlation (average 0.35), indicating substantial differences in estimates of expression. We next evaluated the impact of expression quantification method on eQTL detection. Our method and the standard approach show a low overlap of significant associations (25% of SNPs with $p<10^{-4}$), with the standard approach detecting a larger number of significant eQTLs and more extreme p-values. These results suggest that mapping bias may have inflated p-values, and highlights the importance of comparing methods of quantifying HLA expression.

HUMAN PAPILLOMAVIRUS INDUCES FOCAL GENOMIC INSTABILITY AND DISRUPTS CANCER-CAUSING GENES IN PRIMARY ORAL CANCERS

Keiko Akagi[1,4], Jingfeng Li[1,2], Weihong Xiao[1,2], Tatevik Broutian[1,2], Bo Jiang[1,2], Robert Pickard[1,2], Amit Agrawal[3], Anne-Katrin Emde[5], Nora Toussaint[5], André Corvelo[5], Giuseppe Narzisi[5], Karen Bunting[5], Maura L Gillison[1,2], David E Symer[1,2,4]

[1]Ohio State University, Human Cancer Genetics Program, Columbus, OH, [2]Ohio State University, Internal Medicine, Columbus, OH, [3]Ohio State University, Otolaryngology-Head and Neck Surgery, Columbus, OH, [4]Ohio State University, Molecular Virology, Immunology and Medical Genetics, Columbus, OH, [5]New York Genome Center, Sequencing Project, New York, NY

Approximately 5% of cancers worldwide are caused by human papillomavirus (HPV) infection. While HPV infection is necessary for cancer development, it is insufficient, as the genetic factors that promote cancer progression among HPV-infected individuals are largely unknown. Recently, we comparatively analyzed HPV-positive cell lines and primary tumors. Whole genome sequencing and other molecular methods revealed that HPV integrants frequently flank various forms of host genomic structural variation, including amplifications and rearrangements such as deletions, inversions and chromosomal translocations (1). We proposed a looping model by which HPV integration events may result in focal host genomic DNA replication, resulting in viral-host DNA concatemers (1). Similar focal genomic instability was described in HeLa cells and in other primary oral cancers, confirming this model of HPV looping (2,3). To compare somatic mutational profiles and to determine the frequency of HPV looping events in primary cancers, we now have sequenced the genomes and transcriptomes from more than 50 HPV-positive oral cancer specimens, together with matched normal controls. We also used innovative long-range sequencing methods to characterize the detailed, complex genomic structures induced by HPV integrants in human cancers. These high-resolution results have shed new light on a catastrophic process, distinct from chromothripsis and other mutational processes, by which HPV directly promotes focal genomic instability and alters the expression of genes, thereby potentially contributing to cancer formation.

(1) Akagi, K., Li, J. et al., Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. 24(2): 185-99, 2014.
(2) Adey, A. et al., The haplotype-resolved genome and epigenome of the aneuploidy HeLa cencer cell line. Nature 500: 207-211, 2013.
(3) Parfenov, M. et al., Characterization of HPV and host genome interactions in primary head and neck cancers. Proc. Natl. Acad. Sciences 111: 15544-15549, 2014.

CANINE LYMPHOMA AND MELANOMA SOMATIC ANALYSIS REVEALS THE POWER OF DOG BREED STRUCTURE TO INFORM HUMAN DISEASE

Jessica Alföldi[1], Ingegerd Elvers[1,2], Christophe Hitte[3], Jason Turner-Maier[1], Ross Swofford[1], Jeremy Johnson[1], Chip Stewart[1], Cheng-Zhong Zhang[1,4], Mara Rosenberg[1,4], Clotilde De Brito[3], Edouard Cadieu[3], Marc Gillard[3], Rachael Thomas[5], Catherine André[3], Jaime Modiano[6], Matthew Breen[5,7], Kerstin Lindblad-Toh[1,2]

[1]Broad Institute, Cambridge, MA, [2]SciLifeLab, Uppsala University, Uppsala, Sweden, [3]University Rennes, Institut Genetique et Developpement de Rennes, Rennes, France, [4]Dana Farber Cancer Institute, Boston, MA, [5]North Carolina State University, Raleigh, NC, [6]University of Minnesota, Minneapolis, MN, [7]University of North Carolina, Lineberger Comprehensive Cancer Center, Chapel Hill, NC

Lymphoma is the most common form of hematological malignancy in developed countries. Curability is strongly determined by molecular subtype, reflecting a need for new and improved treatment options. Dogs spontaneously develop both lymphoma and melanoma, and the predisposition of certain breeds has allowed the study of inherited risk factors. Here, we examine the spectrum of somatic mutations in B- and T-cell lymphoma in multiple breeds with different predispositions. We see strong similarities between the golden retrievers and cocker spaniel breeds, both with high B-cell lymphoma frequency. The POT1, FBXW7, and TRAF3-MAP3K14 genes are commonly mutated. The FBXW7 mutations recurrently occur in a specific codon; the corresponding codon is recurrently mutated also in human cancer. In contrast, golden retrievers and boxers, the two T-cell lymphoma predisposed breeds, show little overlap in their mutation pattern. Boxers, who develop aggressive T-cell lymphomas, are typically mutated in the PTEN-mTOR pathway. T-cell lymphomas in Golden retrievers are often less aggressive and the studied tumors typically show mutations in genes involved in cellular metabolism.

The incidence of melanoma in humans has been increasing worldwide faster than any other cancer, and while there is a generally low mortality rate for this cancer, there is a significant subset of cases that do not respond to currently available treatment. Here, we investigated somatic mutations and differential skin expression in poodles, golden retrievers and Labrador retrievers. Little overlap of significantly mutated genes was seen between breeds, although poodles and Labrador retrievers showed significant numbers of mutations in the same pathways.

For both cancers, we identify genes with known involvement in equivalent human cancers, genes implicated in other human cancers, as well as novel genes.

EXOME DATA SHOWS THAT DEMOGRAPHY AND MATING BEHAVIOR SHAPE THE ACCUMULATION OF DELETERIOUS ALLELES IN BONOBOS.

Aida M Andrés[1], Cesare de Filippo[1], Genís Parra[1], Juan Ramón Meneu[1], Romain Laurent[1], Gottfried Hohmann[2], Martin Surbeck[2], Linda Vigilant[2], Svante Pääbo[1], Sergi Castellano[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]Max Planck Institute for Evolutionary Anthropology, Department of Primatology, Leipzig, Germany

Bonobos and chimpanzees are our closest living relatives. The two species are genetically and phenotypically similar, but their societies differ in important ways. Compared with chimpanzees, bonobos live in peaceful and egalitarian societies; in bonobo groups, for example, male competition is moderate and reproduction is loosely determined by dominance rank. The genetic consequences of bonobo's particular social and mating behavior are unknown. We sequenced to high coverage the exome of 20 chimpanzees and 20 bonobos from sanctuaries in Africa, as well as 20 Yoruba humans. We produced and analyzed all data in parallel to obtain fully comparable catalogs of genetic variation in the three species. We then compared the patterns of genetic variation in each species focusing on the X chromosome, which is especially sensitive to sex-biased behaviors. We show that the X chromosome accumulates, compared with the autosomes, a higher proportion of non-synonymous and putatively deleterious alleles in bonobos than in chimpanzees (and humans), likely due to weaker natural selection. This is best explained by the mating behavior of bonobos, which makes them particularly sensitive to the recent reduction in population size that the species has experienced. This study shows how genomic studies help understand the effects that demography and social patterns play in the accumulation of phenotypically relevant variation in natural populations of endangered species.

CELL SURFACE INTERACTOR SEQUENCING (CSI-SEQ) REVEALS NOVEL FEATURES ABOUT INVASIVE CANCER CELL PHENOTYPES

Simeon Andrews, Joel Malek

Weill Cornell Medical College - Qatar, Genetic Medicine, Doha, Qatar

The interactions of cancer cells with their environments are critical to determining their invasiveness. Yet, our understanding of cancer biology and development of new biomarkers is hampered by our lack of knowledge about the cancer cell surface. Therefore, we have developed cell-surface interactor sequencing (CSI-seq) to determine the complement of binding partners that distinguish between cell types.

Using phage libraries encoding millions of fragments from cDNA libraries, we have probed the surface of breast cancer cell lines. Cell lines MDA-MB-231 and MCF-7 model more and less aggressive breast cancer lines, respectively. By exposing these cells to the phage cDNA libraries, followed by brief washing, we can select the phage particles that bind selectively to the cells. By using next generation sequencing, we can quantify the differences in the phage binding to one cell line vs. the other. This allows us to analyze cell surface markers of aggressiveness in cancer, a process we have termed cell-surface interactor sequencing (CSI-seq)

We have shown that phage display, paired with next generation sequencing, can provide quantitative information on the complement of proteins or peptides that bind selectively to particular cancer phenotypes. As an unbiased approach, cDNA phage display on cancer cells promises to uncover substantial new biology and new markers to direct therapy.

MASSIVE SPATIALLY RESOLVED IN SITU GENE EXPRESSION ANALYSIS IN DEVELOPING HEART TISSUE SECTIONS.

Michaela Asp[1], Matthias Corbascio[3], Fredrik Salmén[1], Eva Wärdell[4], Elin Johansson[1], Sanja Vickovic[1], Stefania Giacomello[1], Emelie Berglund[1], José Fernandez Navarro[2], Jonas Frisén[2], Patrik Ståhl[2], Joakim Lundeberg[1]

[1]Royal Institute of Technology, SciLifeLab, Gene Technology, Stockholm, Sweden, [2]Karolinska Institute, Department of Cell and Molecular Biology, Stockholm, Sweden, [3]Karolinska Institute, Department of Molecular Medicine and Surgery, Stockholm, Sweden, [4]Karolinska Institute, Department of Medicine, Stockholm, Sweden

The heart is the first organ that becomes functional in the human embryo, beginning to beat and pump blood already three weeks after fertilization. Although functional at this early stage, the developmental process that the nascent heart must follow is complex. The primitive heart tube will eventually be transformed into a fully operational adult heart with four individual chambers and two separate outflow tracts. Today we lack a complete understanding of how mature cardiac myocytes develop and there is no defined stem cell that can give rise to all the components of the adult heart. Standard bulk gene expression analyses produce an average expression pattern across the entire tissue, masking the expression patterns of individual cells. The developing human heart is a network of interactive but distinct primary cardiac myoblasts, and a spatial view of which genes each cell expresses during the different developmental stages would provide new insights in the field of human embryology and cardiovascular medicine. We are developed a method by which gene expression on the single-cell level and within the context of an entire tissue section can be obtained. By using microarrays covered with poly-T surface probes, we aim to capture the entire mRNA population of every cell within a single fetal heart tissue section. Clusters of surface probes are individually barcoded, giving every cell a unique positional tag. This allows the mRNA material captured on the microarray surface to be pooled while still retaining information about spatial disposition. The final mRNA libraries are analyzed by Massively Parallel cDNA Sequencing, and the obtained sequencing data can then be superimposed back onto the histological image using an in-house software developed for this purpose. By utilizing this method, stem cells that give rise to the heart can be followed over time by analyzing fetal hearts at different gestational points. Furthermore, all gene activity that gives rise to the human heart will be possible to study at a single cell level for the first time.

METAPLASTIC BREAST CANCER: GENEALOGY OF INTERTWINED TUMOR SUBTYPES

Bracha Avigdor(Erlanger)[1], Ashley Cimino-Mathews[2], Roisin Connolly[1], Sarah J Wheelan[1], Ben H Park[1]

[1]The Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, Oncology, Baltimore, MD, [2]The Johns Hopkins Hospital, Pathology, Baltimore, MD

Pathological analysis of Breast Cancer tumors becomes challenging when presented with co-existing tumors with different histologies. Such tumors can arise via multiple mechanisms, each with unique clinical implications. Furthermore, these may be multiple primary tumors that arose in close proximity or a primary tumor that evolved into a different subtype.

One such example is Metaplastic Breast Cancer. MBC is a rare heterogenous subtype of breast cancer prevalent in less than 1% of all Breast Cancer cases, with an increasing annual incidence. MBC presents as epithelial neoplasms that differentiate into mesenchymal and/or Squamous Epithelial lineages with differing histologies, often displaying triple negative (ER-/PR-/HER2-) receptor status with poor prognosis, large tumor size, rapid growth and high recurrence rate.

This study aims to determine whether tumors with different histologies arise from the same clone, and further if Metaplastic Carcinoma (MC) stems from more common ductal tumors or if they arise independently. The hypothesis is that independently arising lesions would share some driver mutations and few passenger mutations, since those occur randomly during tumorigenesis. However, tumors that evolved from each other would share driver mutations and many more passenger mutations.

To that end, whole exome sequencing was performed on Formalin Fixed Paraffin Embedded (FFPE) Invasive Ductal Carcinoma (IDC) and MC tumor pairs from 8 patients. Variant calling and CNV analysis is used to answer these questions by way of comparing variants and their allele frequencies between samples and tracking the mutation rate between and within IDC-MC tumor pairs.

ASSESSMENT OF WHOLE GENOME CAPTURE METHODOLOGIES ON SINGLE- AND DOUBLE-STRANDED ANCIENT DNA LIBRARIES FROM CARIBBEAN AND EUROPEAN ARCHAEOLOGICAL HUMAN REMAINS

María C Ávila-Arcos[1,2], Hannes Schroeder[1], Marcela Sandoval-Velasco[1], Anna-Sapfo Malaspinas[1], Meredith L Carpenter[2,3], G David Poznik[2], Nathan Wales[1], Jay Haviser[5], Carlos D Bustamante[2], M Thomas P Gilbert[1]

[1]University of Copenhagen, Centre for Geogenetics, Natural History Museum of Denmark, Copenhagen, Denmark, [2]Stanford University, Department of Genetics, Stanford, CA, [3]Leiden University, Faculty of Archaeology, Leiden, Netherlands, [4]Identify Genomics, LLC, Menlo Park, CA, [5]St. Maarten Archaeological Center (SIMARC), Philipsburg, St Maarten/St Martin

Recent methodological developments in the area of ancient genomics have increased our ability to extract ancient DNA (aDNA) from poorly preserved samples. Such developments include Whole Genome Capture (WGC) methods and single-stranded (ssDNA) library protocols. We applied two implementations of WGC methods on standard as well as ssDNA libraries generated from several human archaeological samples with different levels of preservation. The endogenous content of libraries consistently increased after WGC. By contrasting the yield between methods and identifying the differences between protocols, we extracted meaningful information about the dynamics of WGC. We identified potential parameters that can be fine-tuned to produce a better yield from aDNA libraries. Thanks to the increase of endogenous content in post capture libraries; enough genetic data was generated to compare the ancient samples to modern reference panels. This way we were able to identify potential source populations for four samples excavated in the Caribbean, three 17th-Century enslaved Africans and one 400-year-old Taino, as well as for two 2500-year-old Etruscans.

OPTIMIZATION OF RNA SECONDARY STRUCTURE PREDICTION FROM CHEMICAL MAPPING DATA IN *ARABIDOPSIS*

Nathan Shih[1], Yiliang Ding[2], Sharon Aviran[1]

[1]UC Davis, Biomedical Engineering and Genome Center, Davis, CA, [2]John Innes Centre, Cell and Developmental Biology, Norwich, United Kingdom

RNA structure plays an important role in post-transcriptional regulation processes such as translation, RNA processing, and RNA stability. Yet, determining structure from sequence is a challenge that is currently being addressed through experimental and computational approaches. Recent advances in chemical modification strategies integrate both approaches by incorporating the chemical modification information into computational structure prediction, thereby improving its predictive power. In chemical modification experiments, paired RNA residues are modified less frequently than unpaired ones. Subsequently, modifications are detected with reverse transcriptase (RT), which ceases synthesis at these modified sites. A negative control is also introduced to determine the natural rate of RT fall-off in the absence of modification. RT stop events in the control and experiment are then counted via high-throughput sequencing and combined to determine the degree of modification at each residue.

Recently, several groups have demonstrated the power of this high-throughput strategy *in vivo*, *in vitro* and at a transcriptome scale. Here, we present a number of analytical tools and approaches to data analysis and use them to investigate existing datasets of chemical modification experiments. We evaluate various analysis strategies and experimental choices with respect to their potential to maximize the structural information extracted from these large and complex datasets. We also highlight sources of error and uncertainty per the experimental design or constraints. Finally, we propose a refined model that has the potential to enhance the predictive power of these experimental approaches. Through these methods, we can better understand how to interpret data from chemical modification experiments and translate them into applicable structure predictions. Ultimately, improvements in these experimental methods and their analysis will drive a deeper understanding of RNA structure in biological functions and provide better strategies for the design of experimental RNA structure studies.

A FIRST GENERATION SPIDER SILK GENE CATALOG FROM THE GOLDEN ORB-WEAVER (*NEPHILA CLAVIPES*) GENOME

Paul L Babb[1,2], Nicholas F Lahens[1,3], John B Hogenesch[1,3], Ingi Agnarsson[4], Linden Higgins[4], Benjamin F Voight[1,2,3]

[1]Perelman Sch. of Med., Pharmacology, Philadelphia, PA, [2]Perelman Sch. of Med., Genetics, Philadelphia, PA, [3]Perelman Sch. of Med., ITMAT, Philadelphia, PA, [4]Univ. of Vermont, Biology, Burlington, VT

Spider silks are the toughest known biomaterials (>150 MJ/m^3) yet are lightweight and extensible. Owing to these properties, they have direct medical and industrial applications. Many spiders produce multiple types of silk from numerous specialized glands, and these silks are composed of a collection of "spidroin" proteins. The silks vary in their biophysical characteristics and how spiders use them. Spiders responsible for the iconic "orb" webs may possess the most diverse repertoire of silk proteins of all, broadly grouped into seven classes. However, due to the lack of a orb-weaving spider genome, an incomplete catalog of full-length silk protein sequences, and minimal information on expression of genes in silk producing glands, the relationship between silk gene structure and function remains unclear. To this end, we assembled the genome of the Golden Orb-Weaver spider (*Nephila clavipes*) and performed RNA-seq to annotate this genome and reveal patterns of silk gene expression.
From our 2.46 Gb assembly of non-repetitive genome (50x coverage, N50: 62 kb) we identified 24 silk genes, including three that were novel. In several cases, there were multiple spidroin genes on individual scaffolds, supporting a tandem duplication hypothesis as a mechanism responsible for the diversification of spidroin loci. Within and among silk gene classes, we saw substantial diversity in the structure, number, size, and sequences of internal repeat cassette motifs. In one example, we noted a fourfold increase in repeat count for a *N. clavipes* major ampullate spidroin gene (94 repeats) relative to the orthologous locus in a non-orb-weaving spider (23 repeats). These internal repeat cassettes increase the number of structural beta-sheets in spidroins, previously shown to correspond with silk toughness and ostensibly an adaptive phenotype for orb-weaving spiders like *N. clavipes*, which preys on large aerial insects to survive. We found patterns of tissue-specific and ubiquitous expression of silk gene transcripts across 10 silk-gland transcriptomes, highlighting the complexity of silk protein expression. The Golden Orb-Weaver, a model spider for the study of silk biology, is now the first orb-weaving spider with a draft genome sequence, a catalog of full-length silk gene sequences, and expression data for silk producing glands. This will allow a deeper understanding of the interplay between silk genes, protein structure, and biomechanical properties of silks, and will underlie future efforts to mimic the extraordinary properties of spider silks in man-made materials.

MULTIMER FORMATION EXPLAINS ALLELIC SUPPRESSION AT PRDM9 HOTSPOTS

Christopher L Baker[1], Pavlina Petkova[1], Michael Walker[1], Petr Flachs[2], Ondrej Mihola[2], Zdenek Trachtulec[2], Petko M Petkov[1], Kenneth Paigen[1]

[1]The Jackson Laboratory, Center for Genome Dynamics, Bar Harbor, ME, [2]Institute of Molecular Genetics of the Academy of Sciences of the Czech Republic, Division BIOCEV, Prague, Czech Republic

Genetic recombination during meiosis functions to increase genetic diversity, promotes elimination of deleterious alleles, and helps assure proper segregation of chromatids at the first meiotic division. Mammalian recombination events are concentrated at specialized sites, termed hotspots, whose locations are determined through sequence-specific DNA binding of the zinc finger protein PRDM9. Beyond PRDM9's ability to specify hotspot locations, numerous questions remain concerning which hotspots are used in vivo, the relative frequency with which they are activated, and the influences that different *Prdm9* alleles have in heterozygous individuals. To begin to answer some of these questions we designed a genetic strategy to identify trans-acting factors influencing the quantitative activity of hotspots in mice, and identified a single locus that overlaps the location of *Prdm9* itself, that functions in an additive manner. Extending this observation, we found that PRDM9 activity is dosage sensitive both to hotspot activation and fertility, and leads to increased number of aberrant spermatocytes and lowered offspring production. In mice heterozygous for *Prdm9*, alleles compete such that the dominant allele can partially or entirely suppress recombination at hotspots that would otherwise be activated by the recessive allele. The physical basis for this appears to be formation of PRDM9 multimers. PRDM9 protein variants form both homo and hetero multimers that bind at hotspots, suggesting that the allelic dominance seen in heterozygous spermatocytes is a form of allelic dominance occurring at the protein level. In PRDM9 heteromeric complexes, one allele is dominantly used to direct recombination initiation towards its own hotspots, effectively suppressing activation of recombination by the other allele.

MAPPING GENOME SELECTION ONTO EMBRYO DEVELOPMENT IN *DROSOPHILA MELANOGASTER*

David Castellano[1], Irepán Salvador[2], Marta Coronado[1], Isaac Salazar[2,1], Antonio Barbadilla[1]

[1]Universitat Autonoma de Barcelona, Institut Biotecnologia i Biomedicina / Department of Genetics and Microbiology, Cerdanyola (Barcelona), Spain, [2]University of Helsinki, Center of Excellence in Experimental and Computational Developmental Biology. Biotechnology Institute, Helsinki, Finland

Developmental genes that differ in spatiotemporal expression patterns have not been systematically compared regarding their intraspecific variation. Integrating information from population genomics data with recent knowledge on developmental genetics allow getting evidence of the evolution of embryonic development at the microevolutionary level and testing the selective importance of hierarchy in gene networks, pleiotropic interactions or preferential evolution of regulatory vs coding sequences. Here we test the hypothesis of whether different embryonic stages differ in their rates of adaptation within the species *Drosophila melanogaster*. Combining population genomics data from the Drosophila Genome Reference Panel with functional and developmental information from the modENCODE project and the Flyexpress database, we map population genetics parameters measuring natural selection at the DNA level, such as alpha (the fraction of adaptive fixation), in time, space and tissue from each embryo stage. The mapping has allowed knowing how the different parts of the embryo (in space and time) evolve through the evolution of genes expressed in them. Correlations between the rate of adaptive substitution and the embryonic stage of gene expression have been searched. The same analyses were carried out with respect to the area or tissue in the embryo for each gene. Thus, we can find whether there are some areas or tissues of the embryo with higher frequency of genes with evidence of positive selection. We estimate in the same way if the different developmental stages differ in the number of genes with evidence of adaptive evolution. The embryonic stages *completed germ-band extension* and *end of embryogenesis* are the most and less evolutionarily dynamics stages of the Drosophila embryo, respectively. Our results seems to indicate that the rate of adaptive substitution differs among genes and developmental stages in the sense that more pleiotropic genes are show less evidence of positive selection.

MACROPHAGES FROM AFRICAN AND EUROPEAN POPULATIONS RESPOND DIFFERENTLY TO BACTERIAL INFECTION.

Yohann Nedelec[1,2], Ariane Page Sabourin[1], Golsheed Baharian[1], Vania Yotova[1], Anne Dumaine[1], Jean-Christophe Grenier[1], Luis B Barreiro[1,3]

[1]Research Center, CHU Sainte-Justine, Pediatrics, Montreal, Canada, [2]University of Montreal, Biochemistry, Montreal, Canada, [3]University of Montreal, Pediatrics, Montreal, Canada

Infectious diseases have always been a major health problem throughout the world, imposing strong selective pressure on the human genome. Geographically distinct human populations are postulated to have differing histories of pathogen exposure. Indeed, previous studies demonstrate that people of African and European ancestry differ in their susceptibility to certain infectious diseases like tuberculosis, malaria and sepsis. Differences in infection progression between African and European populations suggest inter-population variation in the immune response, possibly caused by the adaptation of Africans and Europeans to the pathogens of their environment. For the first time, we characterize the immune response of people of African and European ancestry to bacterial infections. We infected monocyte-derived macrophages from 84 African Americans and 95 European Americans with the intracellular pathogens *Listeria monocytogenes* and *Salmonella typhimurium* for 4 hours and measured whole genome gene expression of infected and non-infected cells by RNA-sequencing. We also assessed macrophage control of bacterial infection and found that macrophages derived from people of African ancestry presented fewer intracellular bacteria after 24 hours of infection than macrophages from people of European ancestry, suggesting that African Americans are able to better control intracellular bacterial growth. Concordantly with this observation we identified marked inter-population differences in gene expression profiles in response to infection, including a markedly stronger pro-inflammatory response in African Americans. By combining population differences in expression with expression quantitative trait loci (eQTL) analyses we further show that more than 30% of the differences in gene expression observed between the two-studied populations are accounted for by *cis*-eQTL variants that present high levels of population differentiation. Interestingly, several of these highly differentiated *cis*-eQTL appear to have been subject to recent positive selection. Collectively, our findings suggest that the clinical differences in bacterial infectious disease progression observed in populations of African and European ancestry may, at least in part, be the outcome of past natural selection.

RNA-DNA DIFFERENCES IN THE MITOCHONDRIAL 16S RRNA ARE CONSERVED AMONG VERTEBRATES AND AFFECT CELL GROWTH

Dan Bar-Yaacov[1], Idan Frumkin[2], Yonatan Chemla[1,3], Philipp Bieri[4], Nenad Ban[4], Lital Alfonta[1,3], Yitzhak Pilpel[2], Dan Mishmar[1]

[1]Ben Gurion University of the Negev, Life Sciences, Beer Sheva, Israel, [2]Weizmann Institute of Science, Molecular Genetics, Rehovot, Israel, [3]Ilse Katz Institute, Nanoscale Science and Technology, Beer Sheva, Israel, [4]University in Zurich, Biology, Zurich, Switzerland

Recently, we were the first to identify RNA-DNA-Differences (RDDs) in the human mitochondrial DNA (mtDNA). The most prominent RDDs (>30% A-to-U and ~15% A-to-G) occurred in all tested tissues and individuals within the 16S rRNA, at position 2617. Furthermore, structural modelling of the RDDs revealed that both the G and T stabilized the structure of the large ribosomal subunit in contrast to destabilization by an A (the pre-RDDs). We therefore hypothesize that the RDDs are functionally important.

As the first step to test this hypothesis we, analyzed RNA-Seq data from mammals, birds, reptiles, amphibians and fish species, and identified the 2617 RDDs in all, thus strongly implying conservation of the mechanism underlying these RDDs formation. RNA-Seq analysis of purified active mammalian (porcine) mitochondrial ribosome revealed notable increased proportion of the RDDs (~82% A-to-U and ~8% A-to-G, ~22,000x coverage) as compared to RNA extracted from isolated mitochondria and in contrast to isolated mitochondrial DNA-Seq (100% A, ~1,000x coverage). Since currently, there is no available technology to mutate the mtDNA position corresponding to the RDDs we sought for an experimental model. To this end we took advantage of the tremendous structural conservation within the tRNA entry channel between the bacterial and mitochondrial ribosomes (which harbors the RDDs site). Interestingly, the 23S rRNA position in *E. coli* that corresponds to mtDNA position 2617 harbors a guanine – thus the mitochondrial A-to-G RDD recapitulates the bacterial base. We used Multiplex Automated Genome Engineering (MAGE) to replace the G for an A, which represents the mtDNA base. Strikingly, *E. coli* strains harboring a G-to-A mutation grew significantly slower than the wild type strains (harboring a G). Taken together, our findings suggest that the mtDNA RDDs are important for mitochondrial function in vertebrates.

INTEGRATION OF INDEPENDENT HUMAN RNA-SEQ AND PROTEOMICS DATASETS – A FEASIBILITY STUDY

Mitra P Barzine[1], James C Wright[2], Jyoti S Choudhary[2], Alvis Brazma[1]

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Functional Genomic Group, Cambridge, United Kingdom, [2]Wellcome Trust Sanger Institute, Proteomic Mass Spectrometry, Cambridge, United Kingdom

In the last few years many RNA sequencing based gene expression studies have been published. Those assays often present overlaps on the studied conditions, for instance, expression profiling of human tissues. A method that enables integration of gene expression measurements across independent RNA-seq datasets with overlapping conditions would be a useful asset, either as a means to build a baseline expression reference or as a way to find new gene expression correlations. Moreover, with increasing number of large-scale proteomics studies, it would be interesting to compare transcriptomics and proteomics data obtained from similar conditions.

Our study aims to understand to what extent for a given condition, different RNA-seq datasets could be used to infer gene expression information, regardless of library preparation and sequencing platform. This task is challenging, as RNA-seq datasets generated in different labs or at different time, are not directly comparable. We have focused our analyses on healthy human tissue data. Five different datasets, including GTEx, have been used for this study. They present – at least – four common conditions (tissues) which have been compared with different approaches. Among them, correlations between samples of different datasets – as a whole or on particular subsets (e.g. most expressed genes, more variant genes, etc.) – were studied.

Preliminaries results are encouraging as same tissues in different datasets present globally same gene expression profiles. Those findings have also been compared to the literature. Figuring out the genes whose expression is more sensitive to library preparation than to biological conditions is one of the challenges of the study.

Finally, as several human protein datasets have been published recently, comparisons have been carried for protein coding genes on transcriptomic and proteomic levels across all datasets and conditions. While correlations between transcriptomic and proteomic across tissues are not high, same tissues show significantly higher correlations (Fisher test p-value <2.2e-16). This suggests that subsets of genes may improve the correlation by discriminating genes less affected by technological variability.

Keys words/expression: RNA-seq, gene expression, integration, proteome

CONTRASTING PATTERNS IN THE HIGH-RESOLUTION VARIATION OF UNIPARENTAL MARKERS IN EUROPEAN POPULATIONS HIGHLIGHT VERY RECENT MALE-SPECIFIC EXPANSIONS

Chiara Batini[1], Pille Hallast[1], Daniel Zadik[1], Pierpaolo Maisano Delser[1], Andrea Benazzo[2], Silvia Ghirotto[2], Eduardo Arroyo-Pardo[3], Gianpiero L Cavalleri[4], Peter de Knijff[5], Turi E King[1], Adolfo López de Munain[6], Jelena Milasin[7], Andrea Novelletto[8], Horolma Pamjav[9], Antti Sajantila[10], Aslıhan Tolun[11], Bruce Winney[12], Mark A Jobling[1]

[1]University of Leicester, Department of Genetics, Leicester, United Kingdom, [2]University of Ferrara, Department of Life Sciences and Biotechnology, Ferrara, Italy, [3]Complutense University, Department of Toxicology and Health Legislation, Madrid, Spain, [4]The Royal College of Surgeons in Ireland, Molecular and Cellular Therapeutics, Dublin, Ireland, [5]Leiden University Medical Centre, Department of Human Genetics, Leiden, Netherlands, [6]University of the Basque Country, Department of Neurosciences, San Sebastian, Spain, [7]University of Belgrade, Institute of Human Genetics, Belgrade, Serbia, [8]Tor Vergata University, Department of Biology, Rome, Italy, [9]Network of Forensic Science Institutes, Institute of Forensic MedicineBudapest, Hungary, [10]University of Helsinki, Department of Forensic Medicine, Helsinki, Finland, [11]Boğaziçi University, Department of Molecular Biology and Genetics, Istanbul, Turkey, [12]University of Oxford, Department of Oncology, Oxford, United Kingdom

The peopling of Europe has been divided into three main phases encompassing the initial colonization during the Upper Paleolithic, the migrations influenced by the climatic changes of the Last Glacial Maximum and the spread of agriculture during the Neolithic. However, the contribution of events during these different phases to the peopling of the area has been subject of a long-term debate. In particular, special focus has been given to determining the proportions of Europeans descending from Neolithic farmers and Paleolithic hunter-gatherers, colonizing Europe from the Middle East from 10 and 40 thousand years ago (KYA), respectively. Here we analyse a dataset of 340 individuals from 17 European and Middle Eastern populations, resequencing 3.7 Mb of the Y chromosome (MSY) and 16.6 kb of the mitochondrial genome (mtDNA). Comparing these two markers we highlight contrasting patterns in the matrilineal and patrilineal genetic histories. On one hand, the distribution of variation of mtDNA suggests pre-Neolithic expansions throughout Europe, with little influence of subsequent peopling waves. On the other hand, the MSY variation points to a very recent expansion in European populations, dating back only to the Bronze Age. This indicates a widespread male-specific post-Neolithic phenomenon which could be explained by a change in the social structure of human populations at the time.

TRANSCRIPTOME-WIDE REGULATORY NETWORKS REVEAL COORDINATED CONTROL OF SPLICING AND EXPRESSION

Yungil Kim, Ashis Saha, Alexis Battle

Johns Hopkins University, Computer Science, Baltimore, MD

Unraveling the networks describing global regulation of gene expression is important to understanding the cascading impact of genetic variation on the cell and describing complex pathways involved in human disease. The inference of networks from gene expression data has been widely used for identifying regulatory relationships among genes; here, network analyses have predominantly focused on regulation of gene expression by transcription factors. Recently, however, RNA-sequencing has enabled the genome-wide quantification of a diverse range gene expression phenotypes, including alternative splicing, non-coding transcripts, and allele-specific expression. By extending regulatory network inference methodologies to capture these additional facets of the human transcriptome, we are able to capture regulatory relationships that were largely excluded from previous analyses. We have developed methods for reconstructing transcriptome-wide regulatory networks from RNA-seq data, learning sparse probabilistic graphical models representing the interconnected regulation of both transcription and alternative splicing. Our networks integrate expression quantitative trait loci to capture both cis and trans genetic effects on these diverse expression traits. We also integrate allele-specific expression (ASE) along with genotype data, where both can provide evidence of directed, potentially causal regulatory relationships between genes, where correlation among total expression levels alone is ambiguous regarding directionality. We apply our methods to the Depression Genes and Networks study of 922 individuals with RNA-seq from whole blood. The resulting transcriptome-wide network indicates a complex interplay between the regulation of transcription and alternative splicing and describes the global impact of genetic variation on each. We identify genes involved in the regulation of alternative splicing, where our findings include known RNA-binding proteins, and enumerate specific candidate target spliced transcripts. Additionally, we identify cases where alternative splicing of transcription factors are involved in differential regulation of target genes' total expression. Networks inferred from RNA-sequencing offer a broader picture of the complex regulatory control of the transcriptome, allowing us to place genes and variants of interest in a more complete cellular context.

COMPARATIVE GENE EXPRESSION ANALYSIS REVEALS DEEP CONSERVATION OF NON-CODING TRANSCRIPTION IN *DROSOPHILA*

Philippe J Batut, Thomas R Gingeras

Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY

Animal development is orchestrated by the unfolding of regulatory programs that dynamically specify the expression patterns of thousands of protein-coding genes and non-coding transcripts. It has long been thought that changes to these regulatory programs drive the evolution of Metazoans, and yet little is known about the evolutionary dynamics of the genomic elements that regulate gene expression. To experimentally study the evolution of transcriptional promoters, we generated genome-wide profiles of promoter activity[1,2] throughout embryonic development in five *Drosophila* species spanning over 25 million years of divergence. We found that promoter gain and loss have been very active processes throughout the evolutionary history of these species. Developmental gene expression profiles tend to be tightly conserved, and regulatory divergence is shaped by systems-level constraints on gene function and developmental stages. At the sequence level, selective pressures act on both core promoter elements and sequence-specific transcription factor binding sites to maintain developmental gene expression specificity. In addition, we detected extremely prevalent non-coding transcription throughout embryogenesis: 4,050 long non-coding RNA (lncRNA) promoters, most of which have never been described before, are dynamically expressed during that critical period. We identified a *Drosophila* core set of over 1,000 lncRNA promoters conserved over 25 million years, and showed that they are under substantial purifying selection at the levels of promoter sequence and expression specificity. We are currently conducting functional studies on the *FBgn0264479* locus, which harbors a lncRNA that is expressed extremely highly, and in a spatially restricted fashion, during a 3-4 hours period around the onset of gastrulation. Overall, our work demonstrates that functional studies in a phylogenetic context are a powerful approach to probe the biological importance of non-protein-coding genes.

1. Batut, P. et al., High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* (2012).
2. Batut, P. and Gingeras, T. R., RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5'-Complete cDNAs. *Curr Protoc Mol Biol* 104, 25B 11 1 (2013).

READ CLOUDS UNCOVER VARIATION IN COMPLEX REGIONS OF THE HUMAN GENOME

Alex Bishara[1], Yuling Liu[1], Ziming Weng[3], Dorna Kashef-Haghighi[1], Daniel E Newburger[1], Robert West[2], Arend Sidow[2,3], Serafim Batzoglou[1]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Pathology, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA

While an increasing amount of human genetic variation is being identified and recorded, determining variants within repeated sequences of the human genome remains a challenge. Most population and genome wide association studies have therefore been unable to consider variation in these regions. Core to the problem is the lack of a sequencing technology that produces reads with sufficient length and accuracy to enable unique mapping. We present a novel methodology of using read clouds, obtained by accurate short read sequencing of DNA derived from long fragment libraries, to confidently align short reads in repeat regions and enable accurate variant discovery. Our novel algorithm, RFA, captures the relationships among the short reads governed by the long read process via a Markov Random Field. We test our method through extensive simulations and apply it to discover variants in the NA12878 human sample, for which Moleculo sequencing data are available, as well as on an invasive breast carcinoma genome that we sequenced through Moleculo. We demonstrate that our method accurately recovers variation in 155Mb of the human genome, including 94% of 67Mb of segmental duplication sequence and 96% of 11Mb of transcribed sequence that are currently hidden from short read technologies.

REVEEL: LARGE-SCALE POPULATION GENOTYPING USING LOW-COVERAGE SEQUENCING DATA

Lin Huang, Bo Wang, Ruitang Chen, Sivan Bercovici, Serafim Batzoglou

Stanford University, Computer Science, Stanford, CA

Population low-coverage whole-genome sequencing is rapidly emerging as a prominent approach for discovering genomic variation and genotyping a cohort. This approach combines substantially lower cost than full-coverage sequencing with whole-genome discovery of low-allele-frequency variants, to an extent that is not possible with array genotyping or exome sequencing. However, a challenging computational problem arises when attempting to discover variants and genotype the entire cohort. Variant discovery and genotyping are relatively straightforward on a single individual that has been sequenced at high coverage, because the inference decomposes into the independent genotyping of each genomic position for which a sufficient number of confidently mapped reads are available. However, in cases where low-coverage population data are given, the joint inference requires leveraging the complex linkage disequilibrium patterns in the cohort to compensate for sparse and missing data in each individual. The potentially massive computation time for such inference, as well as the missing data that confound low-frequency allele discovery, need to be overcome for this approach to become practical. Here, we present Reveel, a novel method for single nucleotide variant calling and genotyping of large cohorts that have been sequenced at low coverage. Reveel introduces a novel technique for leveraging linkage disequilibrium that deviates from previous Markov-based models. We evaluate Reveel's performance through extensive simulations as well as real data from the 1000 Genomes Project, and show that it achieves higher accuracy in low-frequency allele discovery and substantially lower computation cost than previous state-of-the-art methods.

GENOME ANALYSIS OF THE CORALLIVOROUS STARFISH *ACANTHASTER PLANCI* REVEALS CONSERVATION BETWEEN ECHINODERMS AND CHORDATES

Kennet W Baughman, Kanako Hisata, Eiichi Shoguchi, Nori Satoh

Okinawa Institute of Science and Technology, Marine Genomics Unit, Onna-son, Japan

Acanthaster planci (Common name: Crown-of-thorn Starfish) is a corallivorous starfish, commonly known for its consumption and devastation of hard corals in the Pacific and Indian Oceans. It is currently unclear whether *A. planci* population outbreaks have increased in size or frequency due to human intervention, or whether these population density changes are a natural phenomenon. To date, we have sequenced, assembled, and annotated two separate draft genomes from two individual starfish. One specimen was collected on the Great Barrier Reef of Australia, the second specimen was collected at Motobu, Okinawa, Japan. Based on illumina short read sequencing data, The total lengths of the respective genome assemblies are; Australia: 383,525,304 bp (N50 = 916,880 bp, 3274 total scaffolds) Okinawa: 383,843,944 bp (N50 = 1,521,119 bp, 1765 total scaffolds). The genomes align to each other with 98.7% identity. The *A. planci* Nkx cluster shows microsynteny with chordate Nkx clusters, not previously reported in echinoderms. The Nkx cluster is associated with gill slit development, a morphological feature not found in echinoderms. This genomic organization is consistent with our previously published finding of collinearity within the starfish Hox cluster. 1-MA signaling signaling in oocyte meiotic resumption is unique to starfish among echinoderms, yet downstream components of this activated pathway are conserved across deuterostomes both with regard to signal transduction interactions, as well as at amino acid level of the component proteins. In order to visualize these interactions, the components of the 1-MA pathway were identified from the literature, modeled in Systems Biology Graphical Notation (SBGN) diagram, and candidate transcripts and gene models were identified in the draft genome. The remarkably good starfish genomic assembly may be related to recent population dynamics, an open and pressing question in *A. planci* ecology.

IS SANGER SEQUENCING STILL A GOLD STANDARD?

Tyler F Beck[1], Nancy F Hansen[1,2], James C Mullikin[1,2], Leslie G Biesecker[1]

[1]National Human Genome Research Institute, NIH, Medical Genomics and Metabolic Genetics Branch, Bethesda, MD, [2]NIH Intramural Sequencing Center, Comparative Genomics Analysis Unit, Rockville, MD

The costs associated with next-generation sequencing are rapidly decreasing, and consequently an increasing number of clinical research labs have begun using exome or genome sequencing for gene discovery studies or other genomics studies. While some of these data are generated purely for research, it is often the case that medically relevant variants can be identified that a clinician may wish to return to the participant. Currently, the standard of care is to verify variants using Sanger sequencing before returning the results, which can be costly and time-consuming. However, we question whether Sanger validation is necessary given the current quality of next-generation sequencing. In this study, we will compare and contrast the results found using Sanger sequencing on a large set of genes against those found using exome and genome sequencing from the same cohort of samples, namely the ClinSeq® cohort. We propose that a read quality score minimum can be established above which it is unreasonable that a next-generation sequencing call needs to be validated using Sanger sequencing.

A SITE SPECIFIC MODEL OF THE NEUTRAL MUTATION PROBABILITY FOR WHOLE-GENOME CANCER DATA

Johanna Bertl[1], Qianyun Guo[2], Malene Juul Rasmussen[1], Asger Hobolth[2], Jakob Skou Pedersen[1,2]

[1]Aarhus University, Department of Molecular Medicine, Aarhus, Denmark, [2]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

Understanding the mutational process in cancer cells is crucial to distinguish driver mutations, responsible for the initiation and progress of cancer, from passenger mutations. The high somatic mutation rate in cancer cells and the heterogeneity of the process on different levels make this a challenging question: whole-genome pan cancer analyses have shown that the mutation pattern differs fundamentally not only between different cancer types, but also between patients and along the genome.

With the increasing availability of whole-genome DNA sequence data from cancer cells, typically paired with data obtained from healthy tissue, efficient and scalable analysis methods are called for. Population genetic approaches to detect regions under positive selection are often not directly applicable: the outcome of the evolutionary process that takes place in the cancer tissue is usually only observed in a single biopsy. Therefore, methods have been developed that model the neutral mutation probability in windows or specific genomic elements, based on local genomic characteristics. Alternative approaches study the functional impact of individual mutations and the clustering of mutations along the genome.

Here, we model the somatic mutation probability in cancer cells by considering not only heterogeneity of mutation rates both between patients and tissue types, but also covariates that describe the functional relevance of sites and epigenetic factors like methylation and replication timing. Our statistical framework is flexible enough to include both patient and site specific covariates. Modeling the mutation probabilities at single sites allows us to study the mutational process on multiple resolutions, so we can analyze regions of varying size and different types of genomic elements within our framework.

TISSUE-SPECIFIC PATTERNS OF SOMATIC MUTATION

Francis Blokzijl[1], Myrthe Jager[1], Joep de Ligt[1], Valentina Sasselli[1], Meritxell Huch[2], Luc van der Laan[3], Hans Clevers[1], Edwin Cuppen[1], Ruben van Boxtel[1]

[1]Hubrecht Institute for Developmental Biology and Stem Cell Research, KNAW and University Medical Center Utrecht, Genome Biology, Utrecht, Netherlands, [2]University of Cambridge, Gurdon Institute, Cambridge, United Kingdom, [3]Erasmus MC-University Medical Center, Department of Surgery, Rotterdam, Netherlands

Mutational processes continuously challenge the genetic integrity of cells. Genomic changes occur when damages are not properly repaired, or arise stochastically as a result of DNA replication errors. Adult stem cells are prone to gradually accumulate mutations, since they persist throughout life to maintain tissue homeostasis. Adult stem cells propagate their mutations to daughter cells, which could thereby affect the fitness of the tissue. Identification of patterns of mutation accumulation in various adult stem cell types could help us to unravel tissue-specific susceptibility to age-related pathologies, including cancer.
Here we characterized the frequencies and patterns of somatic mutation accumulation in individual stem cells from the human intestine and liver by whole-genome sequencing. Interestingly, we found that point mutations accumulate linearly with age in both stem cell types. Intestinal stem cells accumulate more point mutations during life than liver stem cells, which could be related to the very high proliferation rate of these cells. In line with this, the distribution of point mutations show a stronger association with replication timing in intestinal stem cells than in liver stem cells. Furthermore, mutational signature analysis indicates that deamination of methylated cytosines is the major mutagenic process in intestinal stem cells. On the structural level, we identified somatic copy number variations of distinct nature in the two stem cell types. In intestinal stem cells, we observed some small deletions, whereas in liver stem cells we identified tandem duplications that affect much larger regions harboring many genes. The tandem duplications show microhomology at the breakpoints, indicating that microhomology-mediated repair was involved. The deletions on the other hand, do not show regions of homology at the breakpoints and suggest involvement of non-homologous end joining.
These results show that somatic mutational patterns differ between adult stem cells from distinct tissues, which may be caused by differences in proliferation rate, active mutagenic processes and operative repair mechanisms.

GENOME-WIDE QUANTITATIVE ASSESSMENT OF ENHANCER ACTIVITIES IN HUMAN CELLS BY STARR-SEQ

Łukasz M Boryń, Muhammad A Zabidi, Cosmas D Arnold, Michaela Pagani, Alexander Stark

Research Institute of Molecular Pathology, Stark Lab, Vienna, Austria

Differential gene expression orchestrates development and is controlled by genomic cis-regulatory elements called enhancers. It is becoming increasingly clear that aberrant gene regulation caused by mutations in enhancer sequences plays an important role in various diseases including cancer. Therefore, the identification of enhancers in human cells and the quantitative characterization of their activities is an important goal.

We adapted STARR-seq, a genome-wide quantitative enhancer screening method originally developed in Drosophila cell lines (Arnold et al., Science 2013) for efficient use in human cell lines. Systematic optimization of several parameters yielded a new mammalian STARR-seq screening setup with high signal-to-noise ratio, improved by more than 8-fold compared to an initial setup. Similarly, an improved library construction protocol allows the efficient cloning of all candidate fragments in large mammalian genomes, yielding libraries that cover about 90% of the human genome.

Coupled to electroporation or viral delivery, this allows efficient STARR-seq screens in mammals.
Importantly, the ability of STARR-seq to measure enhancer activity quantitatively allows identification of active enhancers of different strengths and enhancer activity changes after the stimulation of cellular signaling pathways such as hormone treatment or during cell type transitions such as those occurring in cancer or during cell differentiation.

ALLELIC HETEROGENEITY AND EPISTASIS IN THE GENOMIC ARCHITECTURE OF CANINE BODY SIZE

Jess Hayward[1], Marta Castelhano[2], Liz Corey[2], Nate Sutter[3], Rory Todhunter[2], Adam R Boyko[1]

[1]Cornell University, Clinical Sciences, Ithaca, NY, [2]Cornell University, Biomedical Sciences, Ithaca, NY, [3]La Sierra University, Biology, Riverside, CA

Purebred dogs are characterized by strong artificial selection and the maintenance of hundreds of purebred lines, greatly facilitating genetic mapping of complex traits. Strongly selected traits like body size show a simplified genetic architecture whereby a handful of large-effect QTLs explain most of the phenotypic variation across breeds.

Using body size and genotype data from 180,000 SNPs for 2000 individual dogs from >100 breeds, we confirm that ~17 large-effect QTLs explain nearly 90% of the variation in breed-average body size. However, the explanatory power of these QTLs within breeds or in random-breeding "village dogs" is markedly lower (20-40%), even after accounting for sexual dimorphism and the negative correlation between inbreeding and size.

To improve our ability to predict individual body weights, we built a reference panel of 254 sequenced dogs from diverse populations and imputed >10 million SNPs and indels for our genotyped cohort. For a handful of genes with a known causal size variant, the known variant is the top associated marker within the QTL interval. Surprisingly we identify novel variants in multiple genes harboring previously known size-associated variants, suggesting even species with reduced genetic variation due to historical bottlenecks can commonly exhibit allelic heterogeneity. Higher prediction accuracy is achieved using imputed markers and incorporating multiple markers at heterogeneous loci, especially within populations. Furthermore, for the first time in dogs, we detect epistatic effects between several unlinked markers for body size, as well as evidence of genetic incompatibilities between certain loci. Identifying such interactions not only improves our understanding of the genetic basis of size variation after strong diversifying selection, but can also be used to make informed breeding decisions.

GENETIC DIFFERENTIATION AT LOCI UNDER STRONG BALANCING SELECTION: HLA LOCI IN HUMAN POPULATIONS

Debora Y Brandt[1], Jerome Goudet[2], Diogo Meyer[1]

[1]University of Sao Paulo, Genetics and Evolutionary Biology, Sao Paulo, Brazil, [2]University of Lausanne, Ecology and Evolution, Lausanne, Switzerland

Balancing selection increases the genetic diversity in populations relative to neutral expectations, but its effects on the differentiation among populations are more complex. Although genes showing low levels of population differentiation are commonly considered potential targets of balancing selection, it is also plausible that certain balancing selection regimes increase population differentiation. Here, we investigated population differentiation at the Human Leukocyte Antigen (HLA) genes, which are the most well supported cases of genes under balancing selection in the human genome. To this end, we analyzed data from the 1000 Genomes Project, comparing population differentiation of SNPs in the HLA genes with SNPs in the rest of the genome. The extremely high genetic diversity at the HLA genes makes this region problematic for next generation sequencing (NGS) techniques such as those used by the 1000 Genomes project. We thus filtered out SNPs at the HLA genes with unreliable allele frequencies, identified in a study that used Sanger sequencing as a control for the NGS allele frequencies. We found that SNPs in the HLA genes show overall higher differentiation, measured by Fst. However, when comparing SNPs with similar minor allele frequencies, SNPs in the HLA genes showed lower Fst than those in other genomic regions. This is due to the known effect of balancing selection reducing the number of rare variants, for which Fst is constrained to lower values. Thus, we showed that the assumption that balancing selection reduces population differentiation at individual SNPs holds true, providing that one accounts for MAF. More generally, we show that not accounting for MAF when comparing Fst from different genomic regions can lead to an erroneous interpretation of the data.

ULTRAFAST ACCURATE RNA-SEQ ANALYSIS

Nicolas Bray[1,2], Harold Pimentel[3], Páll Melsted[4,5], Lior Pachter[2,3,6]

[1]UC Berkeley, Innovative Genomics Initiative, Berkeley, CA, [2]UC Berkeley, Department of Molecular & Cell Biology, Berkeley, CA, [3]UC Berkeley, Department of Computer Science, Berkeley, CA, [4]University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, Reykjavík, Iceland, [5]deCODE Genetics/Amgen, Reykjavík, Iceland, [6]UC 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. The pseudoalignment procedure used in kallisto can also reveal information about fusion transcripts in cancer samples. 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.

DISCOVERY OF CROSS TISSUE AND TISSUE SPECIFIC EQTL BY DECONVOLVING RNA-SEQ DATA FROM A MULTI-TISSUE DATASET.

Andrew Brown[1,2], Ana Viñuela[1], Alfonso Buil[1], Richard Durbin[3], Timothy Spector[4], Emmanouil T Dermitzakis[1]

[1]University of Geneva, Dep Genetic Medicine and Development, Geneva, Switzerland, [2]University of Oslo, NORMENT, KG Jebsen Center for Psychosis Research, Oslo, Norway, [3]Wellcome Trust Sanger Institute, Informatics, Cambridge, United Kingdom, [4]Kings College London, Twin research, London, United Kingdom

Much of the recent interest in genetic variants affecting gene expression (eQTL) has focused on understanding their relationship with genetic risk loci for disease. Currently, gene expression studies are reporting eQTLs found in many different tissues. By combining a more complete ascertainment of eQTLs and the particular tissues they act in with information from GWAS, we should have a better concept of mechanisms of disease and the relevant tissues.

With this aim, we have taken RNA-seq expression data from four tissues (skin, fat, whole blood and LCLs) in a cohort of around 800 individuals and applied linear mixed models to deconvolve expression into "tissue specific" and "cross tissue" components for genes with non-zero expression correlation across tissues (6083 out of 11,522 genes). These two new properties (“phenotypes”) were used to identify cis eQTLs, both acting across tissues and in specific tissues, and we compare the results to those of a standard independent cis mapping of expression phenotypes in those tissues (referred to as standard analysis/eQTLs). In detail, we find more than 91% of the genes have a cross tissue eQTL (FDR < 0.05); in comparison the standard analysis found that 84% of genes had an eQTL in at least one tissue. We also find more tissue specific eQTLs than using the standard analysis: between 54% and 94% of genes have an eQTL (depending on tissue), this is on average 20% more than are found with a standard analysis. A larger proportion of cross tissue eQTLs (38%) lie within 20kb of the transcription start site than tissue specific eQTLs (30%), which is consistent with the fact that eQTLs which act in multiple tissues are often located in promoter regions. Replication rates across tissues were higher for tissue-shared eQTLs, with a median replication rate of 93% compared to 73% for standard eQTLs and 69% for tissue specific eQTLs.

Further work will investigate cross tissue and tissue specific eQTL in relation to specificity of ENCODE annotations, and whether these phenotypes can discover eQTL whose effects differ across tissues (a sign of gene-environment interactions). Finally, we are looking to integrate these tissue shared and specific eQTL with GWAS signals, to help prioritise the search for new genetic risk loci, and to infer mechanisms and important tissues in disease pathology.

INFERENCE OF INDIVIDUAL-LEVEL ADMIXTURE DATES

Katarzyna Bryc, Joanna Mountain

23andMe, Inc., Mountain View, CA

Population-level analyses of admixture have demonstrated the feasibility of inferring the timeframes of historical migrations resulting in admixture among continental populations. Here, we explore the ability to infer admixture dates for a single individual under a simple model of admixture. We apply a two-step method that utilizes summary statistics from forward simulations and a tract length based likelihood calculation for inferring genealogical relationships from shared segments. We estimate the ability to resolve the number of generations since a genealogical ancestor, fully from a population, passed down a given ancestry to an admixed descendant. We also explore violations of our model and discuss challenges to admixture dating at the individual level.

EVALUATION OF THE GENETIC REGULATION ACROSS TISSUES IN A TWIN COHORT

Alfonso Buil[1], Ana Viñuela[1,2], Andrew A Brown[1], Kerrin Small[2], Richard Durbin[3], Timothy D Spector[2], Emmanouil T Dermitzakis[1]

[1]University of Geneva, Genetic Medicine and Dev., Geneva, Switzerland, [2]King's College London, Twins Research, London, United Kingdom, [3]Wellcome Trust Sanger Institute, Computational Genomics, Hinxton, United Kingdom

Gene expression can vary from one tissue to another. Here we quantify the effects of genetic regulatory variants that are shared among tissues and the ones that act in a tissue specific manner.
We sequenced the mRNA of ~400 female MZ and DZ twin pairs from the TwinsUK cohort in four tissues: fat, skin, blood and LCLs (2733 samples in total) and used genotypes imputed into the 1000G reference panel. We identified 9166 cis eQTLs in fat, 9551 in LCLs, 8731 in skin and 5313 in blood (1% FDR). To measure the replication rate of different regulatory variants between tissues we calculated pi1 of the association observed in one tissue for the variants discovered in another tissue. We found that about 70% to 80% of the cis eQTLs observed in one tissue are cis eQTLs in another tissue. Using functional annotation from the ENCODE and FANTOM5 projects we observed that genetic mechanisms in proximal transcription factor binding sites (TFBS) are shared among tissues while distal TFBS and enhancers act in a more tissue specific manner.
But known cis eQTL are only a small part of the genetic effects that affect gene expression, we estimate they explain on average 20% of the heritability of expression. To discover to what degree genetic effects are shared among tissues, we used bivariate variance components techniques to decompose the correlation of gene expression between two tissues into genetic and environmental components. We found that 15% of genes that are heritable in both fat and blood and 60% of those genes in fat and LCL show a significant positive genetic correlation between tissues, with an average value of 0.6. We also observed that for most genes we could not reject the hypothesis that the genetic correlation in cis was equal to 1; this supports the idea of shared cis genetic mechanisms between tissues controlling expression. For most of the tissue pairs we found no evidences that trans genetic effects were correlated across tissues, with the exception of fat and skin where we found strong evidence of trans shared genetic effects for 66 genes. We are now in the process of identifying specific examples of trans eQTLs that are active in both fat and skin at the same time.
In conclusion, we observe that both known and unknown cis regulatory variants are largely shared across tissues, while regulatory variants in trans are highly tissue-specific. However we do see evidence for limited sharing of trans effects between fat and skin, which we are investigating further.

CONSTRUCTION OF *ZEA MAYS* HAPLOTYPE MAP

Robert Bukowski[1], Qi Sun[1], Edward Buckler[2]

[1]Cornell University, Institute of Biotechnology, Ithaca, NY, [2]Cornell University, Plant Breeding and Genetics, Ithaca, NY

Significant diversity in maize poses a major challenge in efforts to identify genetic polymorphisms. With only one reference genome (B73) publicly available to date, most variants detected from alignments of short reads are false positives, resulting from misplacement of reads originating from haplotypes absent from the reference genome. Massive structure variations, such as copy number variation or translocations present in the *Zea* population make the interpretation of the alignments even more problematic. In this work, we present a filtering pipeline designed to extract reliable set of polymorphisms from alignment of 37 billion Illumina WGS sequencing reads from 916 diverse maize inbred lines to B73 reference. In an attempt to eliminate false positives, the raw polymorphism set was subject to extensive filtering using segregation test followed by identity by descent (IBD) and linkage disequilibrium (LD) filters based on a set of about 1 million polymorphic sites obtained using the genotyping by sequencing (GBS) method. The GBS markers are concentrated mostly in low-complexity, relatively well conserved genomic regions, making them a reliable anchor for detecting regions of IBD and for LD testing. Using this pipeline, we identified about 60 million polymorphic sites, including roughly 30 million conclusively confirmed to be in local LD with the GBS anchor. The latter subset is most robust with respect to various parameters, such as the quality threshold imposed on alignments used in genotyping or the number of maize lines used, and should be considered most reliable. Interpretation of the alignments - and thus the quality of the obtained variant set - could be improved if the reference were expanded by adding representative haplotypes from lines other than B73.

DISSECTING QUANTITATIVE REGULATION OF ROOT GROWTH USING SYSTEMS GENETICS

Wolfgang Busch, Takehiko Ogura, Santosh Satbhai, Radka Slovak

Austrian Academy of Sciences, Gregor Mendel Institute of Molecular Plant Biology, Vienna, Austria

A fundamental question in biology is how the genome of an organism gives rise to its phenotype. This is an enormous challenge as this question spans multiple scales and numerous layers of complex, often recursive, interactions of biological entities such as chromatin, genes, proteins, cells, tissues, the whole organism and the environment. Consequently, a powerful approach to address the genetic basis of complex phenotypes is to conduct joint analyses on genetic, molecular, and phenotypic data. An excellent model to approach these challenges is the plant *Arabidopsis thaliana* for which unique resources exist, including high resolution genotyping data and full genome sequences of more than a thousand inbred, phenotypically diverse wild strains, and high quality genomic-scale gene network models based on a large number of system-biology-type experimental datasets. We use genetically determined variation of root growth between different strains as a model for how genetic differences affect organ growth in order to determine which genes, networks, and biological pathways lead to differences in root growth and architecture. For this, we use custom phenotyping pipelines that enable us to capture quantitative root phenotypes of a very large number of individuals, genome wide association studies (GWAS) to identify the associated loci in the genome, and systems-biology driven approaches to identify the gene networks and pathways that provide the molecular and cellular context for the underlying genes to quantitatively regulate root growth. Using these approaches, we have recently identified and experimentally verified multiple novel regulators and regulatory modules of epistatically interacting genes that shape root growth. Overall, using this systems-genetics approach enables us to approach the genotype to phenotype question at the level of genetic networks, significantly advancing our comprehension of how complex biological traits are modulated by different genotypes.

MULTI-SAMPLE ISOFORM QUANTIFICATION IN GTEX RNA-SEQ DATA

Andrea E Byrnes[1,2], Francois Aguet[2], David DeLuca[2], Timothy Sullivan[2], Julian B Maller[1,2], Taru Tukiainen[1,2], The GTEx Project Consortium [3], Kristin Ardlie[2], Benjamin M Neale[1,2]

[1]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [2]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA, [3]Broad Institute of Harvard and MIT, The Genotype-Tissue Expression (GTEx) Project Consortium, Cambridge, MA

Alternative splicing is critical for the regulation and diversity of the majority of human genes. The ability of a single gene to give rise to several diverse transcripts, and subsequently proteins, has been implicated in a wide range of processes and disorders from brain function to cancer proliferation. The GTEx project has provided an unprecedented collection of RNA-seq samples from a wide variety of individuals and tissues, ideal for examining the range normal transcriptional diversity. The relatively short read-lengths obtained from current technologies do not provide perfect information as to which isoforms are present, however GTEx provides a unique opportunity for the development of statistical methods for isoform discovery and quantification using data from current-generation RNA-seq technology.

We propose a 2-step, multi-sample method for discovery and quantification of transcript isoforms (both known and novel) from paired-end RNA-seq data, while making use of a reference genome and any available annotation. Our method aims, first, to maximize information about splicing behavior by combining information from all aligned RNA-seq samples in order to construct a graph representing all possible transcripts, similar to approaches taken by PSGInfer and Cufflinks, but on all samples pooled together. In graph-building, we weight each of the possible junctions between exons by the number of junction reads observed across all samples and allow for junctions not present in the annotation. We represent each isoform as a possible path through the graph and the use the weight of each edge as the initial probability in the following step. After constructing all likely isoforms from the data, we use the expectation-maximization algorithm to estimate their relative abundance, similar to the methods applied in RSEM and eXpress. This second step allows us to specifically characterize the isoforms present in any sample (or collection of samples from a common individual or tissue) and quantify their respective transcription for each sample separately. We will discuss the details of this method as well as the relative performance compared to existing methods using additional RNA-seq runs on longer fragments.

LARGE-SCALE GENOTYPING OF POLYMORPHIC INVERSIONS IN THE HUMAN GENOME

Sergi Villatoro[1], Roser Zaurin[1], Magdalena Gayà-Vidal[1], Carla Giner-Delgado[1,2], David Vicente-Salvador[1], David Izquierdo[1], Meritxell Oliva[1], Lorena Pantano[1], Marta Puig[1], Mario Cáceres[1,3]

[1]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra (Barcelona), Spain, [2]Universitat Autònoma de Barcelona, Departament de Genètica i de Microbiologia, Bellaterra (Barcelona), Spain, [3]Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

Despite the initial expectations on the potential effects of the great number of genomic structural variants (SVs) discovered in multiple species, so far the results have been rather disappointing. In particular, inversions are known to have adaptive consequences in different organisms and have been associated to phenotypic and functional differences between individuals. However, their balanced nature and the presence in many cases of highly identical inverted repeats (IRs) at the breakpoints make the study of inversions especially challenging. Here, we present a new method called iMLPA (from inverse multiplex ligation-dependent probe amplification) to genotype simultaneously multiple inversions mediated by IRs in hundreds of individuals, which is based on a combination of inverse PCR and probe hybridization (patent pending). Currently, developed assays are able to genotype more than 30 inversions with as little as 25 ng of DNA per inversion, although it should be easy to add more inversions as they are identified. To test the performance of the technique, we have genotyped 24 of these inversions in 550 individuals of seven diverse human populations from the 1000 Genomes Project with 98.5% genotyping success rate. In addition, by comparing with the results obtained by PCR for a subset of the samples, we have shown that iMLPA is highly accurate and most errors accumulate in specific inversions affected by restriction-site polymorphisms. Finally, by combining the iMLPA results with additional inversion genotyping by regular MLPA and inverse PCR, we have carried out the largest analysis of the impact of inversions in the human genome. This information has allowed us to establish the population distribution and evolutionary history of the inversions, uncover several inversions with functional effects on genes, and show that most inversions with IRs are recurrent and are not linked to SNPs. Therefore, having a high-throughput technique to genotype these inversions is crucial to study their association with complex phenotypes and disease susceptibilities, and could contribute to the unraveling of the hidden heritability of the human genome.

IDENTIFICATION OF GENETIC CHANGES UNDERLYING TAMENESS IN DOMESTIC ANIMALS

Alex Cagan[1], Frank W Albert[1,2], Gabriel Renaud[1], Victor Wiebe[1], Irina Plyusnina[3], Oleg Trapezov[3], Lyudmila Trut[3], Torsten Schöneberg[4], Svante Pääbo[1,5]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]695 Charles E. Young Dr. S, Gonda Research Center, Los Angeles, CA, [3]Siberian Branch of the Russian Academy of Sciences, Institute of Cytology and Genetics, Novosibirsk, Russia, [4]University of Leipzig, Institute for Biochemistry, Leipzig, Germany, [5]Uppsala University, Department of Immunology, Uppsala, Sweden

Domestic animal species share a variety of morphological and behavioral traits, known collectively as 'domestication syndrome'. One such trait is tameness, a reduction in fearful and aggressive behavior towards humans. To identify the genetic changes underlying this trait we study two experimental models of domestication. Two lines of wild-derived brown Norway rat (Rattus norvegicus) and American mink (Neovison vison) have been under continuous divergent artificial selection for 70 and 15 generations respectively. In both species one line is selected for tameness and the other for aggression towards humans. This selection regime has resulted in substantial physiological and behavioural differences between the lines. For both species we generated whole-genome sequence data from 20 individuals from each line. Using this data we identify candidate genes with putatively causative variants segregating between the lines. We detect more sharing of these candidate genes between the rats and mink than expected by chance, suggesting a partially convergent genetic response to selection. To ascertain the relevance of the rat and mink models to the domestication process, and to test whether there is evidence for a convergent genetic basis to domestication in additional animal species, we identified genes with putatively functional alleles segregating at high frequency in several domestic animal species compared to non-domesticated populations descended from the respective wild progenitors. We identify several genes and pathways that may have been targeted by selection for tameness during domestication across species.

A GENOMIC ASSESSMENT OF POPULATION STRUCTURE AND SEX-BASED MIGRATION IN AN ENDANGERED NON-MODEL PRIMATE GENUS, *MICROCEBUS*

Christopher R Campbell[1], Peter A Larsen[1], Jeffrey Rogers[2], Anne D Yoder[1]

[1]Duke University, Biology, Durham, NC, [2]Baylor College of Medicine, Human Genome Seq. Center, Houston, TX

Historically, methods such as F_{ST} have been the primary measures for assessing population structure. Though the basic assumptions on which these methods are based are still relevant, the nature and depth of biological questions expands when posed in the context of genome-scale data. For example, genomic perspectives on gene flow between species promise detailed insights into the frequency and directionality of gene-by-gene introgression. Prior studies of the behavior, parentage, and genetic diversity of wild mouse lemur populations have demonstrated that they are, like many social primates, matrilocal. Female mouse lemurs maintain a home territory and even share nesting sites with their daughters while males migrate, serving as the purveyors of genetic variety. Prior to the application of genome-scale data, field observations as well as DNA paternity tests of the individuals within a study site were necessary to confirm the genetic consequences of this biological system. Only by behaviorally monitoring and DNA testing individuals were patterns of sex-based migration subject to inference. With the availability of a sequenced genome for the gray mouse lemur, *Microcebus murinus*, these hypotheses can be tested at a scale beyond that of several individuals. Questions can be posed such that comparisons of loci across the sex chromosomes reveal the degree to which sex-biased dispersal impacts genome organization within and between populations. The first characterization of prosimian sex chromosome contigs and the generation of RADseq data from field samples spread across the ranges of sister taxa *Microcebus murinus* and *Microcebus griseorufus* reveal the power of this approach. A comparison of F_{ST} between loci on sex chromosomes, autosomes, and within mitochondrial DNA has confirmed previously hypothesized behaviors of *Microcebus* in the wild, at a wider scale than previously possible, and additionally paves the way for future genomic studies of lemurs.

GENOME-WIDE EPIGENETIC REPROGRAMMING DURING NORMAL POSTNATAL DEVELOPMENT OF THE LIVER.

Matthew V Cannon[1], Genay Pilarowski[1], Tenisha Phipps[1], Xiuli Liu[2], David Serre[1]

[1]Cleveland Clinic Lerner Research Institute, Genomic Medicine, Cleveland, OH, [2]Cleveland Clinic, Anatomic Pathology, Cleveland, OH

Most epigenetic programming is thought to take place during early development, with DNA methylation patterns changing less thereafter. However, increasing evidence suggests that epigenetic changes also occur later as part of normal cellular differentiation. The postnatal liver is an excellent model to study epigenetic remodeling during differentiation as the tissue is relatively homogeneous, composed primarily of progenitor cells at birth and fully differentiated cells at weaning.
In the current study, we show extensive epigenetic reprogramming in the liver during postnatal development. 118,877 of 261,317 CpGs measured changed their methylation level by more than 5% from birth to nine weeks of age, with some changing by up to 86% (ie: from 0% to 86% methylation). Interestingly, these changes in DNA methylation occur primarily in intergenic enhancer regions while gene promoters seem less affected. Analysis of 166 CpGs at 8 intermediate time points by locus-specific bisulfite sequencing reveals that this reprogramming occurs between postnatal day 5 and 20, with DNA methylation at specific CpGs changing by up to 57% in this period. This coincides with two essential cellular changes in the liver: the differentiation of hepatocytes and extensive cell division. While cell multiplication leaves a distinct footprint on the DNA methylation patterns, we show that the extensive epigenetic reprogramming occurs concurrently with differentiation of hepatocytes.
The data presented demonstrate extensive epigenetic remodeling in the postnatal liver in mice during differentiation of hepatocytes. Our data suggest that epigenetic remodeling is an important aspect of hepatocyte differentiation, with enhancers potentially playing a particularly important role in controlling the transition from progenitor to differentiated hepatocytes. Our data also clearly show that DNA methylation patterns can change dramatically after early development.

MAPPING THE "DARK MATTER" OF GENOME – LONG REPEATS, COMPLEX STRUCTURAL VARIATIONS AND THEIR BIOLOGICAL RELEVANCE

A Hastie[1], A Pang[1], E Lam[1], T Chan[1], W Andrews[1], T Anantharam[1], X Zhou[1], J Reifenberg[1], M Saghbinia[1], H Sadowski[1], M Austin[1], P Sheth[1], Z Dzakula[1], X Xun[2], T Graves[3], J Sikela[4], P Kwok[5], H Cao[1]

[1]BioNano Genomics, San Diego, CA, [2]BGI, Shenzhen, China, [3]WA University, St. Louis, MI, [4]University of CO School of Medicine, Denver, CO, [5]UCSF, San Francisco, CA

Despite advancements in next-generation sequencing, a portion of the human genome remains unresolved or ambiguously characterized and large genomic structural variations (SV > 1 kb) are found more prevalent than we thought. During assembly, they leave gaps and unknown structural or heterozygous information as the "dark matter" of the genome, often challenging to detect for short read NGS and conventional low-resolution cytogenetic techniques. Rapid comprehensive genome mapping in NanoChannel Arrays represents a platform independent of yet complementary to DNA sequencing for accurate genome assembly and structural variation analysis. *De novo* assembly of these single molecules yields unprecedented long contiguous genome maps, advantageous in spanning over highly repetitive regions and complex structures in their native form.

We present results from complex genomes such as human and cancer. We detected hundreds of large structural variants per genome and haplotype differences in these genomes, 11% of 24,360 large SVs found in the 22 euploid human genome are unique to a specific genome while 23% of those SVs were common to 20/22of the samples. In 1 human genome, we detected over 700 of insertions/deletions and inversions larger than 1 kb. Without considering SVs that overlap with N-base gaps in hg19, 90% of these SVs are supported by orthogonal experimental methods or historical evidence in public databases. A higher portion of the complex genome is composed of previously unknown repeating unit (>2 kb) of large sizes spanning several tens of kilobases to multiple megabases, the exact locations and copy numbers of these repeats often remain elusive with NGS. Without knowing the genomic context of these repeats or the amount of repeats, it is difficult to attach any biological relevance to them. Using BioNano's Irys® platform, repeat regions can be more accurately characterized and put into context. We were able to find repeats and complex SV regions spanning 100-200 kb or more that are clinically associated with cardiovascular disease risk, brain size, obesity and neurobehavior disorders. For the first time, population scale cross-sample genome comparison to identify comprehensive genomic structural variation is feasible on a single platform.

Overall, genome mapping provides highly valuable structural information otherwise hard or impossible to decipher with short read sequencing data alone.

RETROGENES ILLUMINATE DYNAMICS OF NEW GENE STRUCTURE AND REGULATORY EVOLUTION IN MAMMALS

Francesco N Carelli[1,2], Maria Warnefors[1,2], Henrik Kaessmann[1,2]

[1]University of Lausanne, CIG, Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, SIB, Lausanne, Switzerland

New genes are thought to have substantially contributed to phenotypic innovation. However, the mechanisms governing their functional evolution are poorly understood. To illuminate the dynamics of new gene evolution, we investigated so-called retrogenes, which originate as intronless copies of their "parental" source genes through an RNA-mediated mechanism. These copies, usually devoid of the parental promoter, need to gain new regulatory elements and, potentially, new exons/introns to evolve into functional retrogenes, making them ideal models to study new gene origination. Here, we explored retrogene evolution in nine representative mammals and one bird using extensive transcriptome and chromatin modification data. We find that regulatory elements of retrogenes were recruited from other genes, inherited from their parental genes, or, most frequently, obtained from genomic elements in their vicinity (e.g., CpG islands and enhancers). Remarkably, we observed that retrogenes may rapidly evolve new multiexonic structures and may even undergo alternative splicing. Finally, we investigated the functional relevance of mammalian retrogenes, based on evolutionary gene expression analyses, and identified retrogenes that may have contributed to the specific organ biology of different mammalian lineages. Notably, we identified a number of "orphan" retrogenes that functionally replaced their parental genes during evolution. Altogether, our work highlights how intronless gene copies, usually initially devoid of regulatory elements, can evolve into actively transcribed, complex multiexonic genes within a short evolutionary time. It thus provides novel insights into the general mechanisms underlying the origins and functional evolution of new genes.

IDENTIFYING REGIONAL VARIATION AND CONTEXT DEPENDENCE OF HUMAN GERMLINE MUTATION USING RARE VARIANTS

Jedidiah Carlson[1], Jun Li[2], Sebastian Zöllner[1,3]

[1]University of Michigan, Department of Biostatistics, Ann Arbor, MI,
[2]University of Michigan, Department of Human Genetics, Ann Arbor, MI,
[3]University of Michigan, Department of Psychiatry, Ann Arbor, MI

Mutation is the ultimate source of genetic variation and one of the driving forces of evolution. In both germline and somatic tissues, mutation rates vary along the genome, and are affected by local features such as GC content and chromatin structure. Characterizing regional variation of mutation patterns is important for understanding genome evolution and to identify variants causing genetic diseases. However, many aspects of the interplay between genomic features and mutation patterns are poorly understood. Despite its central importance, mutation rate and molecular spectrum are difficult to measure in an unbiased, genomewide fashion. Estimates based on common variants (polymorphisms) and substitutions are confounded by natural selection, population demographic history, and biased gene conversion (BGC). Methods relying on quantifying population incidence rate of single-gene diseases or finding de novo variants by trio sequencing do not provide sufficient data genomewide to assess more than the most basic parameters.

We overcome these limitations by using a collection of > 30 million singleton variants observed in our whole-genome sequencing study of bipolar disorder (n = 4,000 unrelated subjects). Compared to polymorphisms or substitutions, these extremely rare variants (ERVs) arose very recently and are much less affected by the confounding effects of selection, BGC, etc. Compared to trio sequencing studies, the high density of ERVs (>1,000 ERVs per 100kb) provides substantially more power to detect subtle effects of genomic and epigenomic context. With this approach, we assess subtle regional differences in the mutation process across the human genome at a 1-10kb scale. We explore the impact of genomic features, such as GC content, functional annotation, and replication timing on such regional variations. Moreover, we evaluate the impact of local sequence context on mutation rates, thus possibly providing insight into the underlying biological processes creating mutant alleles. By comparing ERVs and common variants across the genome we are able to assess the effect of evolutionary processes. For example, we observe a significant enrichment for AT>GC transitions among common variants but not ERVs, indicating biased gene conversion favoring the derived allele has a strong impact on the human genome. These results will provide a framework for developing a comprehensive atlas of mutation in the human genome and ultimately improve our knowledge of the core process of genetic variation.

THE TIME AND PLACE OF EUROPEAN GENE FLOW INTO ASHKENAZI JEWS

James Xue, Shai Carmi, Itsik Pe'er

Columbia University, Computer Science, New York, NY

Genetic studies have demonstrated that the Ashkenazi Jewish gene pool is a roughly even mixture of Middle-Eastern and European ancestries. However, the time and geographic distribution of the European source(s) are heavily debated, with different hypotheses spanning nearly the entire past two millennia and various locations across the continent. To resolve this puzzle, we first applied local ancestry inference to distinguish between European and Middle-Eastern segments in the Ashkenazi genomes. Then, using a likelihood-based approach and calibration by simulations, we determined the most likely geographic source. To infer the time of admixture, we developed a new method based on analytical results for the distribution of ancestry proportions in admixed genomes. We estimate the European ancestry in the Ashkenazi individuals to be about 70% South-European (the rest being West and East European), and the admixture time to be about 30-45 generations ago.

FUNCTIONAL SCREENING OF lncRNA: TOWARDS THE FANTOM6 PROJECT

Michiel de Hoon[1], Jay W Shin[1], Chung-Chau Hon[1], The FANTOM Consortium[2], Piero Carninci[1]

[1]RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Japan, [2]http://fantom.gsc.riken.jp/, Japan

The FANTOM consortium has been providing transcriptional and regulatory maps to the community over the past 15 years. The recently published FANTOM5 project established the most extensive catalogue of the mammalian transcriptome, covering promoters and enhancers across the broadest collection of human and mouse primary cells and tissues. We subsequently published a novel paradigm describing the general principles governing the dynamic behavior of cells based on an analysis of the temporal interplays between enhancers and promoters in development as well as in response to various external cues across a wide collection of cellular systems.

Here, we will present the FANTOM5 human lncRNA catalogue, built from integrating the FANTOM5 CAGE with various RNASeq sources, including our unpublished data. The current catalogue, which consists of 51,788 lncRNA genes with their promoters precisely mapped, provides a unique portal for understanding the genomic origins of lncRNA. More importantly, it allows us to interrogate potential regulatory lncRNAs across a large collection of primary cells and multiple time courses based on their expression patterns across ~2,000 samples. Our analysis suggests a much broader variety of lncRNA promoters and their dynamic expression in multiple time courses, elucidating the rules that regulate lncRNAs in mammalian cells. We have also produced, for a subset of data, a large collection of small RNA sequencing data.

For the above lncRNAs, we are now embarking on a high throughput functional screening project, building on our unique collection and expression atlas. We are knocking down a broad collection of transcribed lncRNAs in multiple primary human cell types, monitoring their cellular phenotype and obtaining molecular phenotypes by deepCAGE.

JOINT MODELLING OF MULTIPLE TRAITS AND VARIANT SETS INCREASES POWER AND YIELDS NEW INSIGHTS IN THE GENETIC ARCHITECTURE OF COMPLEX TRAITS

Francesco Paolo Casale*[1], Barbara Rakitsch*[1], Christoph Lippert[2], Oliver Stegle[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom, [2]Microsoft Research, eScience, Los Angeles, CA
* These authors contributed equally

Despite the success of Genome-wide Association Studies (GWAS), which have uncovered thousands of individual quantitative trait loci, a substantial portion of the heritability of many complex traits remains to be explained. Recent studies have shown that joint analyses of multiple traits and variant-sets can help recover some of this "missing heritability". In this context, multi-trait models leverage shared genetic and environmental factors affecting multiple traits, whereas variant-set tests aggregate across polygenic signals within genomic regions that harbour causal variants. Despite the effectiveness of such integrative modeling approaches, computational burdens have hindered their widespread use, and in particular it has not been possible to combine both of these concepts.

Here, we present a scalable algorithm for multi-trait set tests (mtSet) that can handle datasets with tens of thousands of samples and dozens of traits, while accounting for population structure and relatedness. We find that mtSet substantially increases power to detect genuine associations when analysing multiple correlated traits.

In a joint analysis of four blood lipid traits from individuals in the Northern Finland Birth Cohort (NFBC), mtSet identified 14 genome-wide significant QTLs, 13 of which are consistent with results from much larger meta-analyses. Power of mtSet was substantially greater than both existing set tests (14 vs 12) and multi-trait models (14 vs 10), demonstrating the merits of integrative modelling. We find similar results in other studies, including a joint analysis of six basal haematology traits in rodents, where mtSet accurately corrected for strong relatedness between samples.

Beyond increasing power in genetic mapping, mtSet can be used to obtain new insights into the genetic architecture of correlated traits. We derive theoretical results that connect the concordance of causal loci between traits to properties of the (genetic) correlation matrix between traits. We develop a rigorous procedure to test for these properties and apply the method to simulated data and blood lipid traits from NFBC. We find instances where naiive analyses would suggest pleiotropic architectures, which is actually due to multiple trait-specific associated loci in the same region.

GLOBAL ANALYSIS OF HUMAN POLYMORPHIC INVERSIONS FROM THE INVFEST DATABASE

Sònia Casillas[1,2], Alexander Martínez-Fundichely[1], Isaac Noguera[1], Mario Cáceres[1,3]

[1]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra, Barcelona, Spain, [2]Universitat Autònoma de Barcelona, Departament de Genètica i de Microbiologia, Bellaterra, Barcelona, Spain, [3]Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

The newest genomic advances have uncovered an unprecedented degree of structural variation throughout genomes, with great amounts of data accumulating rapidly. However, compared to insertions and deletions, inversion prediction presents unique challenges due to the complexity of this type of changes and the underlying high false positive discovery rate. In order to get a reliable estimate of the real number of inversions in the human genome we have integrated multiple sources of information in the InvFEST database (http://invfestdb.uab.cat). InvFEST automatically merges predictions into different inversions taking into account the resolution of each specific study, refines the breakpoint locations, and finds associations with genes and segmental duplications. In addition, it includes data on experimental validation, population frequency, functional effects, and evolutionary history, which is constantly being updated. Recently incorporated features include the implementation of BreakSeq to automatically predict the generation mechanism for each inversion, analyze the DNA properties at the breakpoint sequences, and determine the ancestral orientation. Categorization of the 1092 candidate inversions based on internal scores and manual curation indicates that almost half of them are either unreliable or false. Using the subset of validated inversions from InvFEST, we have performed a meta-analysis with the aim to uncover the main characteristics of inversion polymorphisms in humans. On the one hand, compared to simulated inversions generated between all intrachromosomal inverted segmental duplications (SDs), inversions generated by non-allelic homologous recombination (NAHR) between inverted repeats tend to appear between the most identical, physically-close SDs, are enriched in chromosome X, and tend to break and invert genes less often than expected. On the other hand, inversions not generated by homology-related mechanisms appear in regions that are more flexible and less stable than NAHR inversions, they are genetically longer, and display a larger distance to the closest gene. All in all, while InvFEST aims to represent the most reliable set of human polymorphic inversions to date, the reported results and further analyses underway should contribute to understand their functional and evolutionary impact in the human genome.

NGS-BASED REVERSE GENETIC SCREEN FOR EMBRYONIC LETHAL MUTATIONS COMPROMISING FERTILITY IN LIVESTOCK.

Carole Charlier[1], Wanbo Li[1], Chad Harland[1,2], Mathew Littlejohn[2], Frances Creagh[2], Pierre Faux[1], Mike Keehan[2], Steve Davis[2], Nico Tamma[1], Latifa Karim[1], Naveen Kadri[1], Tom Druet[1], Wouter Coppieters[1], Richard Spelman[2], Michel Georges[1]

[1]Unit of Animal Genomics, GIGA-Genetics, University of Liège, Liège, Belgium, [2]Livestock Improvement Corporation, Research & Development, Hamilton, New Zealand

Here we took advantage of high-throughput sequencing to apply a genotype-to-phenotype reverse genetic strategy to address the decline in fertility suffered by modern cattle populations. It is generally acknowledged that fertility is negatively correlated with production and hence also to have an infinitesimal genetic architecture. Several examples suggested a complementary hypothesis where part of the fertility decline could involve embryonic lethal alleles reaching high frequency in the population as a result of the large-scale use of elite sires.

We first rigorously compared rate and nature of coding variations in conserved exons for six domestic bos taurus breeds (n=59) and six human populations (n=60) and demonstrated that despite the presumably severe domestication bottlenecks and intense artificial selection, present-day domestic cattle are genetically more variable than humans. In contrast, the synonymous/non-synonymous ratio was twice larger in cattle with an equivalent load of loss-of-function (LoF) variants likely due to more effective purging of deleterious recessives in livestock.

Whole genome/exome sequences were generated for healthy animals representative of two specialized bovine populations, the meaty Belgian Blue Cattle breed (n=80) and the New Zealand dairy cattle population (n=500). Sequence data were mined for recent, heterozygous, breed-private mutations of four categories: stop-gain, frame-shift, essential splice-site and predicted harmful missense mutations. A ranked list of thousands of candidates - selected on the basis of (i) the embryonic/fetal expression of the involved gene, (ii) available data from knock-out effects in model organisms, and (iii) minor allelic frequency - was then evaluated for embryonic lethal (EL) outcome. Custom array-based genotyping of large cohorts of healthy individuals revealed a significant depletion in homozygous mutant animals. From permutation testing with neutral SNP matched for minor allele frequency and assuming random mating, we evaluated that < 15% of candidate variations were EL. The effects of top candidates were further investigated prospectively in carrier-carrier mating to validate the absence of mutant offspring. Currently the lethal nature of nine mutations has been confirmed. The resulting information will be useful to avoid at-risk mating, thereby improving fertility.

ANALYSIS OF rRNA SEQUENCES FROM RNA-SEQ DATA FOR TAXONOMIC SURVEY OF MICROBIAL COMMUNITIES

Lei Chen, Lauren Petersen, Blake Hanson, Benjamin Leopald, Erica Weinstock, George M Weinstock

The Jackson Laboratory, Department of Microbial Genomics, Farmington, CT

Genomic sequencing has long been used to investigate complex microbial communities, like those from the human gut. Amplicon sequencing targeting the 16S rRNA gene is most commonly used for a cost effective taxonomic survey. However RNA-Seq analysis of metagenomic samples are being performed more frequently to define the transcriptome of the community. Here we explore the use of rRNA sequences from metagenomic RNA-Seq data for taxonomic survey. Since cellular RNA contains mostly rRNA, rRNA depletion is customarily performed to enrich for mRNA. Omitting the rRNA depletion is an effective way to capture this "taxonomy barcode" in its entirety. Unlike amplicon sequencing, this is effectively shotgun sequencing of the rRNA and requires new informatics approaches for analysis. Sequencing and analysis of an artificial microbial community were performed to develop this approach. These results as well as application to human microbiome samples will be presented.

DEVELOPMENT AND ANALYTICAL VALIDATION OF A PHARMACOGENOMICS ION AMPLISEQ SEQUENCING ASSAY COVERING 138 VARIANTS AND CYP2D6 CNV

Shann-Ching Chen, Manimozhi Manivannan, Guoying Liu, Toinette Hartshorne, Zhoutao Chen, Mark Andersen, Fiona Hyland

Thermo Fisher Scientific, Genetic, Medical and Applied Sciences Division, South San Francisco, CA

Background

Cytochrome P450 enzymes metabolize about 75% of drugs, including oncology drugs, with UGT enzymes metabolizing about another 15%. Variations in gene sequence or in copy number may result in an inactive, defective, unstable, mis-spliced, low expressed, or absent enzyme, an increase in enzyme activity, or an altered affinity for substrates. Pharmacogenomics genes can predict whether an individual is a poor or rapid metabolizer, facilitating dose optimization. Failure to adjust dosage of drugs metabolized by the relevant enzyme can lead to adverse drug reaction, or conversely to too rapid drug metabolism and no drug response.

Materials and Methods

We have designed a pharmacogenomics (PGx) panel to detect 138 targets in 36 genes and CYP2D6 copy number variation (CNV). This panel covers the common targets in genes encoding drug metabolism enzymes and associated transport proteins, including CYP2D6 and CYP3A4. This assay uses Ion AmpliSeq™ technology and contains 132 amplicons in an ultrahigh-multiplex PCR in a single pool, followed by Ion Torrent™ semiconductor sequencing. The assay requires as little as 10 ng of input DNA. This panel can be customized, allowing additional targets to be added.

Results

Analytic validity of the panel was established by sequencing 91 well characterized and annotated cell lines from Coriell, and comparing the Ion AmpliSeq genotypes to the annotated genotypes and to the gold standard TaqMan® OpenArray®. Concordance with TaqMan genotypes was > 99.5%, genotype reproducibility was > 99.7%, and the no-call genotype rate was < 0.5%.

Conclusions

The panel measures PGx gene genotypes with high accuracy. These results demonstrate an assay which can be used to explore potential pharmacogenomic relationships, including the relationship between copy number and genotype of metabolism enzyme targets on drug tolerability and clinical outcomes.

THE ROLE OF GWAS-IMPLICATED TYPE 1 AND TYPE 2 DIABETES LOCI IN THE PATHOGENESIS OF LATENT AUTOIMMUNE DIABETES IN ADULTS (LADA)

Alessandra Chesi[1], Vanessa C Guy[1], Mohammed I Hawa[2], Jonathan P Bradfield[3], Kevin J Basile[1], Hakon Hakonarson[1,3], Charles Thivolet[4], Didac Mauricio[5], Nanette C Schloot[6], Knud B Yderstræde[7], Stanley Schwartz[10], R. David Leslie[2], Bernhard O Boehm[8,9], Struan F Grant[1,3]

[1]Children's Hospital of Philadelphia, Division of Human Genetics, Philadelphia, PA, [2]Queen Mary University of London, Centre for Diabetes, Barts and the London School of Medicine and Dentistry, London, United Kingdom, [3]Children's Hospital of Philadelphia, Center for Applied Genomics, Philadelphia, PA, [4]Lyon 1 University, INSERM, U870, IFR 62, Lyon, France, [5]Hospital Universitari Germans Trias i Pujol, Badalona, Spain, [6]German Diabetes Center, Düsseldorf, Germany, [7]Odense University Hospital, Odense, Denmark, [8]University of Ulm, Ulm, Germany, [9]Nanyang Technological University, Singapore, Singapore, [10]Main Line Health System, Wynnewood, PA

The genetic etiology of adult-onset autoimmune diabetes, and especially 'latent autoimmune diabetes in adults' (LADA), remains unresolved. LADA exhibits features of both type 1 (T1D) and type 2 (T2D) diabetes, earning it a reputation as being "type 1.5 diabetes". LADA cases present initially as T2D (adult-onset but not requiring insulin), but with circulating islet autoantibodies as in T1D. Indeed, 5-10% of 'apparent' T2D patients are typically misdiagnosed LADA cases. Leveraging existing genome-wide SNP genotyping data generated on the Illumina Infinium II OMNI Express platform, we assessed the association of all GWAS-implicated T1D and T2D loci reported to date in 965 LADA subjects and 1134 healthy control subjects recruited from a European multicenter study. Through strict phenotyping criteria, diagnosis was defined by subjects aged 30-70 years, positive for glutamic acid decarboxylase autoantibodies without initial insulin therapy for at least 6 months.

The MHC region was strongly associated with LADA (sentinel SNP rs1063355; odds ratio (OR) = 1.58; $P = 3.24 \times 10^{-10}$). Strong evidence of association was also observed for other T1D-associated loci, including *SH2B3* (rs3184504 and rs1265564, OR = 1.45; $P = 1.18 \times 10^{-6}$), *IL7R* (rs6897932, OR = 1.31; $P = 4.34 \times 10^{-5}$ & rs1445898; OR = 1.23; $P = 2.94 \times 10^{-4}$), *PTPN22* (rs2476601 and rs6679677, OR = 0.71; $P = 2.62 \times 10^{-4}$) and *INS* (rs689, OR = 1.45; $P = 3.46 \times 10^{-4}$); however, there was no association with the strong T1D-associated locus, *CLEC16A*. Contrary to many previous reports, there was no evidence of an association between the key T2D *TCF7L2* locus and LADA; indeed, no T2D loci were associated with LADA.

LADA, the major form of adult-onset autoimmune diabetes as defined here, reveals a genetic etiology very similar to childhood-onset T1D, with the striking exception of *CLEC16A*. Absence of the genetic locus *CLEC16A* could account for a later age at presentation of T1D, notably LADA.

FAST AND SCALABLE STRUCTURAL VARIATION ANALYSIS FOR LARGE-SCALE GENOME SEQUENCING PROJECTS

Colby Chiang[1], Ryan M Layer[2], Gregory G Faust[3], Michael R Lindberg[3], David B Rose[3], Erik P Garrison[4], Gabor T Marth[2], Aaron R Quinlan[2], Ira M Hall[1]

[1]Washington University, The Genome Institute, St. Louis, MO, [2]University of Utah School of Medicine, Human Genetics, Salt Lake City, UT, [3]University of Virginia, Biochemistry, Charlottesville, VA, [4]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Diverse classes of genome variation including single nucleotide variants (SNVs), short insertions and deletions (indels), and structural variants (SVs), are recognized to contribute to common human disease. Thus, comprehensive analysis of all variant types in whole genome sequencing (WGS) studies of tens-to-hundreds of thousands of human samples is of broad academic and clinical interest. However, current methods for efficiently interrogating genetic variation in massive cohorts limit many analyses to SNVs and indels due to the computational burden of SV discovery, which can require over 18 hours to process a single genome, and due to the difficulty of integrating, genotyping and interpreting spatially imprecise SV calls across numerous datasets.

The SpeedSeq pipeline for WGS data analysis and interpretation addresses many of these outstanding issues. On a 50X dataset it can perform alignment, SNV, indel, and SV detection in 13 hours with comparable accuracy to industry standards using a single 16 core server with 128 GB of RAM. In particular, SpeedSeq's SV detection module, composed of LUMPY, SVTyper, and a parallelized implementation of the third-party tool CNVnator, runs in just 1.5 hours per genome, rendering SV analysis computationally feasible for large-scale studies. The output seamlessly integrates into the GEMINI variant interpretation framework, and contains genotype and read-depth information that is crucial for assessing clinical significance.

We also address the "N+1 problem" of joint variant detection, where the addition of a new sample to an existing multi-sample callset necessitates computationally expensive reprocessing from raw alignments. While the N+1 problem has been well studied in the context of SNV detection, it lacks a scalable solution for integrating structural variation from multiple genomes. We introduce a graph-based solution whereby each sample is processed individually but retains LUMPY's variant probability curve in the output. Then, clusters of variants are merged based on their probabilistic overlap and the community profile of the candidate cluster. We show that this SV integration method produces high quality multi-sample callsets that are on par with traditional joint variant calling approaches, is scalable to thousands of samples with reasonable hardware requirements, and retains underlying estimates of uncertainty for each variant call. Taken together, these developments advance the quality, simplicity, and computational feasibility of comprehensive variant detection in massive genome sequencing studies.

THE INTERNATIONAL GENOME SAMPLE RESOURCE: BEYOND THE 1000 GENOMES PROJECT.

Laura Clarke, Holly Zheng-Bradley, Julia Khobova, Avik Datta, Ian Streeter, David Richardson, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Vertebrate Genomics, The Wellcome Trust Genome Campus, Cambridge, United Kingdom

The 1000 Genomes Project provides an essential reference catalog of human variation. It includes more than 80 million variant sites ranging from single nucleotide polymorphisms to structural variant events including inversions and duplications. Global allele frequencies and genotypes for 2504 individuals are also provided; these samples are from 26 different populations across Europe, Africa, East and South Asia and the Americas. This resource has enabled many other projects to better interpret their results. Primary uses for the 1000 Genomes data sets include serving as imputation panels to derive whole genome variant sets from exome or array-based genotypes; as filters of shared variation in rare disease or cancer sequencing projects; and for exploring demography and selection in human populations and evolution.

Although the 1000 Genomes Project has reached its goal and finished successfully, the use of the resulting data sets and analysis results remains high. The International Sample Genome Resource (IGSR), launched in January 2015, will maintain, update and expand this valuable reference data set.

The IGSR will maintain the FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp) and the project website (http://www.1000genomes.org) to ensure the community can access both the raw data and the documentation about the 1000 Genomes Project. We will establish a stable version of the 1000 Genomes Browser and the tools it hosts (http://browser.1000genomes.org) based on the project's final date release. This project specific Ensembl-based browser displays all of the 1000 Genomes variants as soon as possible and will use the GRCh37 assembly of the human reference genome.

IGSR will also extend and expand the 1000 Genomes dataset to ensure it remains useful to the community. These plans fall into three main categories. 1) Remap the existing data to the new human assembly GRCh38. 2) Draw in other data created on the Coriell Cell lines such as the RNA-Seq data created by the Geuvadis project, whole genome sequencing at a deeper level or from new technologies. 3) Add new open population collections sequenced using the same strategy to expand the diversity of the public catalog. We also aim to build new tools to improve how discoverable the data is and allow people to view the data alongside other genomic annotation in Ensembl.

GENE EXPRESSION WITHOUT CANONICAL CHROMATIN MARKING IN DEVELOPMENTALLY REGULATED GENES

Silvia Perez-Lluch[1], Enrique Blanco[1], Joao Curado[1], Hagen Tilgner[2], Roderic Guigo[1], Montserrat Corominas[3]

[1]Center for Genomic Regulation (UPF), Bioinformatics and Genomics, Barcelona, Spain, [2]Stanford University, Genetics, Stanford, CA, [3]Universitat de Barcelona, Genetics, Barcelona, Spain

Chromatin organization plays a key role in the regulation of gene expression, with specific histone modifications assumed to contribute to gene activation and/or repression. In contrast to this view, we found that transcription of genes regulated during development occurs mostly in the absence of histone modifications canonically linked to gene activation. We analyzed modENCODE data in fly, and found that developmentally regulated genes have levels of histone modifications associated to transcription initiation (H3K4me3 and H3K9ac), elongation (H3K36me3), and enhancer activity (H3K4me1 and H3K27ac), which are comparable to those in silent genes, and much lower than in genes stably expressed during development. Conversely, by computing partial correlations, we found that strong chromatin marking is as significantly associated to high expression levels as to high transcriptional stability—a feature that may be general across all metazoans, since we also found it in worm and mouse. These observations do not arise from restricted expression since we found that genes broadly expressed across the larval body at a given developmental time have much lower levels of histone marks than stable tissue specific genes, even though the expression of the former is much higher. We replicated these observations by performing RNASeq and ChipSeq in Wing (WID) and Eye Imaginal Discs (EID), and by carrying out targeted experiments on isolated cells. Our analyses also show that a highly structured chromatin state favors regulated exon inclusion and, consistently, unmarked genes with regulated expression during development exhibit more stochastic splicing than stably expressed genes. We further show that promoters of developmentally regulated genes have a characteristic architecture that globally differentiates them from stably expressed genes. Our results support a model in which chromatin marking is associated to stable, tightly controlled production of RNA, while unmarked chromatin would permit rapid gene activation and de-activation during development. In these genes, Transcription Factors binding to chromatin would play a comparatively more important regulatory role. From the mechanistic standpoint, preliminary experiments show that developmentally regulated genes have strong depletion of ASH2 binding, a key factor for H3K4me3. Consistently we observed no reduction in their expression in ASH2 mutant clones, in contrast to strong reduction for stably expressed genes.

IMMUNE-MEDIATED DISEASE GWAS RISK VARIANTS ARE NOT CONSISTENT WITH eQTL DATA

Alexandra Casparino[1], Chris Cotsapas[2,3]

[1]Boston College, College of Arts and Sciences, Boston, MA, [2]Yale University, Departments of Neurology and Genetics, New Haven, CT, [3]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Boston, MA

Genome-wide association studies have identified hundreds of genetic loci influencing risk of immune-mediated disease. The challenge now is to develop robust approaches to understand how risk variants in each locus alter gene function and thus the molecular processes governing disease risk. We and others have shown that the majority of risk variants are preferentially located on tissue- and cell-type specific *cis*-regulatory elements, strongly suggesting that these risk variants alter gene regulation rather than gene structure. This enrichment is restricted to elements active in immune cell subpopulations, with the strongest signals found in stimulus-activated T cell populations. Independent analyses have shown that risk loci are enriched for eQTL effects, in apparent corroboration of the regulatory nature of risk variants.

These observations hide a contradiction: the *cis*-regulatory region analyses suggest risk variants alter cell-type-specific elements in T cells, but the eQTL data is drawn from EBV-immortalized lymphoblastoid cell lines (LCLs) derived from B cells, which are distinct in function and genome use from T cells. Further, LCLs have undergone dramatic genomic changes since their isolation, affecting their gene regulatory programs. Whilst it is possible that the affected cis-regulatory regions are common to both cell populations, the enrichment analyses do not support a major role for regulatory regions active in LCLs.

To resolve this paradox, we have developed a new approach to compare GWAS and eQTL data directly. Rather than look for overlaps between GWAS and eQTL statistics at individual variants, we compare the patterns of association to find instances where the two traits are *concordant*. We have compared all known GWAS risk loci for multiple sclerosis (MS) and inflammatory bowel disease (IBD) to LCL-derived eQTLs, and find support for GWAS-eQTL consistency in only 9/97 MS loci and 11/162 IBD loci, even though most loci contain multiple strong eQTL signals. We are expanding these analyses to eQTL data from primary immune cells and to other immune-related disease GWAS to establish if risk variants drive tissue-specific eQTLs or if disease risk variant effects on gene expression are much more modest than previously thought.

Overall, our results suggest abundant caution should be exercised when developing hypotheses about the molecular effects of GWAS risk variants and we propose that detailed, tissue-specific resources will be required to understand the gene regulatory effects of these alleles.

MASSIVELY PARALLEL SINGLE CELL PROFILING OF CHROMATIN ACCESSIBILITY BY COMBINATORIAL INDEXING

Darren A Cusanovich[1], Riza Daza[1], Andrew Adey[2], Hannah Pliner[1], Lena Christiansen[3], Choli Lee[1], Michael Morse[1], Joel Berletch[4], Christine Disteche[4], Kevin L Gunderson[3], Frank J Steemers[3], Cole Trapnell[1], Jay Shendure[1]

[1]University of Washington, Department of Genome Sciences, Seattle, WA, [2]Oregon Health and Science University, Department of Molecular & Medical Genetics, Portland, OR, [3]Illumina, Inc., Advanced Research Group, San Diego, CA, [4]University of Washington, UW Medicine Pathology, Seattle, WA

Bulk measurements of genomic information represent averages across the population of cells being assayed, and thus are not ideal for studying heterogeneous or dynamic populations. To address this, recent technical advances have enabled the collection of genomic datasets with single cell resolution, e.g. single cell RNA-Seq. In most such methods, microfluidics are used to isolate single cells to separate chambers followed by biochemical processing of the nucleic acid content of each cell within its own reaction volume. However, it is challenging to work with single cells and low nucleic acid inputs. Furthermore, the number of single cells analyzed with these methods scales linearly with cost and effort. To overcome this, we adapted a combinatorial indexing scheme (Amini et al., *Nature Genetics* 2014) to efficiently generate genomic datasets on thousands of single cells without ever requiring their isolation and individualized processing. As a proof-of-principle, we adapted the 'assay for transposase-accessible chromatin using sequencing' (ATAC-seq; Buenrostro et al., *Nature Methods* 2013) to combinatorial indexing, measuring chromatin accessibility in thousands of single cells in each experiment. We first applied this method to a mixture of human and mouse cells and demonstrated that we recover single cell ATAC-seq profiles that overwhelmingly derive from one or the other species, suggesting minimal cell-to-cell contamination. We next applied this method to heterogeneous populations of cells from the same species and demonstrated that single cells cluster based on cell type-specific chromatin accessibility. To our knowledge, this is the first demonstration of sequencing-based profiling of the chromatin landscape in single cells. More broadly, our strategy of combinatorial cellular indexing is applicable to other genomic assays, potentially enabling the routine, comprehensive molecular profiling of many thousands of single cells.

EXOME-WIDE SEQUENCING SHOWS LOW MUTATION RATES AND IDENTIFIES NOVEL MUTATED GENES IN SEMINOMAS

Ioana Cutcutache[1,2], Yuka Suzuki[1,2], Iain B Tan[3], Subhashini Ramgopal[1,2], Shenli Zhang[2], Kalpana Ramnarayanan[2], Anna Gan[2,4], Heng Hong Lee[2,4], Su Ting Tay[2], Aikseng Ooi[5], Choon Kiat Ong[4], Jonathan T Bolthouse[6], Brian R Lane[6], John G Anema[6], Richard J Kahnoski[6], Patrick Tan[2,7], Bin Tean Teh[2,4,7], Steven G Rozen[1,2]

[1]Duke-NUS Graduate Medical School, Centre for Computational Biology, Singapore, Singapore, [2]Duke-NUS Graduate Medical School, Program in Cancer and Stem Cell Biology, Singapore, Singapore, [3]National Cancer Centre Singapore, Department of Medical Oncology, Singapore, Singapore, [4]National Cancer Centre Singapore, Laboratory of Cancer Epigenome, Singapore, Singapore, [5]Van Andel Research Institute, Laboratory of Interdisciplinary Renal Oncology, Grand Rapids, MI, [6]Spectrum Health Hospital System, Division of Urology, Grand Rapids, MI, [7]National University of Singapore, Cancer Science Institute of Singapore, Singapore, Singapore

Testicular germ cell tumors are the most common cancer diagnosed in young men, and seminomas are the most common type of these cancers. To more thoroughly investigate somatic mutations in seminomas and the genes they affect, we combined exome sequencing and copy-number-alteration analysis of eight tumors and their matched normal DNAs. The rate of nonsynonymous somatic mutations averaged 0.31 per megabase. We detected nonsilent somatic mutations in 96 genes that were not previously known to be mutated in seminomas and of which some may be driver mutations. Many of the mutations appear to have been present in subclonal populations. In addition, two genes, *KIT* and *KRAS*, were affected in two tumors each with mutations that were previously observed in other cancers and are presumably oncogenic. Our results, the first reported based on whole-exome sequencing of seminomas, show that, while rates of somatic mutations are five times higher than previous findings based on very limited data, the frequency of mutations is nevertheless low compared to other common cancers. Thus, seminomas are characterized by a relatively low somatic mutation rate, a finding observed among other cancers that also have excellent rates of disease remission achieved with chemotherapy.

THE NEW ILLUMINA TRUSEQ* EXOME KIT OPTIMIZED FOR LESS OXIDATIVE DAMAGE, HIGHER ENRICHMENT EFFICIENCY AND HIGHER UNIFORMITY OF COVERAGE.

Agata Czyz, David Schlesinger, Lindsay Freeberg, Scott Kuersten, Asako Tan, Victor Ruotti, Dixie Hill, Ramesh Vaidyanathan

Illumina, Inc., Product Development, Madison, WI

The human genome is composed of ~3 billion base-pairs of which only ~1-2% is comprised of coding exons. Methods to enrich human DNA libraries for exons are an effective means to identify and characterize genomic variants linked to disease. The basic principle is to first create a whole genome library and then use computationally designed pools of biotinylated oligomers that can hybridize to human exon sequences followed by capture using streptavidin beads. This approach results in high coverage of each target region and minimizes the needed sequencing depth to accurately determine variants at a reduced cost per sample compared to whole genome sequencing.

To that end, we have developed a new Illumina TruSeq* Exome kit. Starting from only 100ng of genomic DNA our kit delivers comprehensive and highly uniform coverage of exon sequences. The product workflow consist of three general steps: acoustic DNA shearing, a modified TruSeq* Nano library preparation protocol, and a rapid capture exon enrichment step. All steps have been optimized to provide a high percentage of reads/bases on target, high coverage uniformity, and minimized oxidative DNA damage which can be introduced during library preparation.

Costello et al. (2013) demonstrated that mutations can be introduced during acoustic shearing. To counter the observed oxidative damage we added EDTA to our shearing buffer. Furthermore, we optimized the library preparation procedure to obtain a median insert size of 150 bp (+/- 10bp), which provides optimal enrichment metrics. Also, we reduced the number of PCR cycles to limit the amount of duplicates introduced during amplification, but still obtain sufficient yield to enable flexibility in choosing the amount of each sample for enrichment.

The TruSeq* Exome kit is designed for use with Illumina's Coding Exome pool (CEX, 37 Mb), the Expanded Exome pool (EEX, 62 Mb), and custom pools selected using Illumina's DesignStudio. During enrichment optimization we validated a new wash buffer that further protects the DNA from artificially induced mutations. The combination of EDTA in the shearing buffer and the new wash buffer allows us to decrease probability of incorrect variant calls, which is particularly important for researchers interested identifying rare variants.

We have enabled pre-enrichment pooling of up to 12 libraries for higher throughput without compromising standard enrichment metrics: plexing of 3, 6, 9 and 12 libraries all result in >80% reads on-target. Both the high enrichment efficiency and high coverage uniformity makes our new TruSeq* Exome kit an ideal tool for analyzing the exome or other areas of interest.

*For Research Use Only

IDENTIFICATION OF DRIVER MUTATIONS IN NON-CODING REGULATORY ELEMENTS IN BREAST CANCER

Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Florence Coulet, Christopher DeBoever, Angelo Arias, Frauke Drees, Richard Schwab, Kelly Frazer

University of California, San Diego, Institute for Genomic Medicine, San Diego, CA

The presence of hundreds to thousands of somatic mutations in a tumor makes it difficult to identify the driver mutations that confer a selective advantage and allow for uncontrolled proliferation of the cancer cells. To date, efforts to identify driver mutations have almost exclusively focused on the 3% of the genome that encodes for genes, whereas the 97% of the genome that is non-coding remains widely unexplored. In this study, we utilized breast cancer whole-genome and transcriptome sequencing data generated by The Cancer Genome Atlas (TCGA) combined with regulatory regions defined by DNase I hypersensitive sites (DHSs) from ENCODE to perform a genome-wide investigation of regulatory elements for driver mutations.

We developed a method to account for background mutation rates in non-coding sequences and analyzed 47 breast cancer samples in TCGA to identify driver regulatory elements. We focused our analysis on DHSs because they comprise all chromatin areas that are accessible to proteins that bind DNA, such as transcription factors, and are used, in general, as markers for regulatory regions. Eighty-six DHSs were identified that are mutated in two or more breast cancer samples, enriched for mutations above background rates based on their sequence and chromatin properties, and associated with at least one aberrantly expressed target gene. We considered these 86 DHSs as candidate driver regulatory elements and performed a prevalence screen in 185 additional breast cancer samples to further filter out possible false positives. The replication samples included 50 tumors in TCGA with whole-genome sequence data and 135 tumors in which we performed targeted sequencing of the 86 DHSs. In the 185 replication samples we detected mutations in 33 of the 86 candidate DHSs. Finally, we experimentally examined the biological roles of two of the replicated driver DHSs using CRISPR and found that deleting these sequences resulted in expression changes of multiple target genes consistent with them being functional regulatory elements.

This study is one of first large-scale analyses of mutations in non-coding regions in cancer, and the first to experimentally investigate the driver roles of distal regulatory elements in cancer. We demonstrate that in some cases aberrant gene expression is a consequence of mutations in regulatory elements rather than secondary effects of alterations in upstream genes. Our study provides insights into the types of molecular alterations that result in the development of breast cancer and may lead to the identification of new prognostic markers or therapeutic targets.

COMPARATIVE STUDY OF GENE ISOFORM EXPRESSION ESTIMATES USING RNA-SEQ, EXON-ARRAY, AND RT-qPCR PLATFORMS IN GLIOBLASTOMA MULTIFORME

Matthew L Dapas*, Manoj Kandpal*, Yingtao Bi, Ramana V Davuluri

Northwestern University Feinberg School of Medicine, Department of Preventive Medicine, Chicago, IL

Given that genes commonly generate diverse functional products, it is important in gene expression studies to try and evaluate expression at the isoform level. Accurately quantifying isoform-level transcripts remains a challenging enterprise, however, despite the continual introduction of new methods and algorithms, which have subsequently resulted in a great amount of data that don't necessarily correlate. Better understanding how to integrate isoform expression data between platforms and pipelines can therefore improve the reliability and power to detect causal aberrations.

Previous studies have demonstrated strong gene-level correlations between RNA-Seq and microarray expression analyses, but have not studied their concordance at the gene isoform level. To further assess the correlations between platforms at the gene isoform level, we performed RNA-Seq and exon-array analyses on common glioblastoma multiforme (GBM) samples from The Cancer Genome Atlas (TCGA) using a number of different analysis pipelines. The RNA-Seq and exon-array results were also compared against estimates obtained using RT-qPCR on a subset of transcripts in GBM. Poor correlations were found between isoform-level expression estimates obtained by RNA-Seq, exon array, and RT-qPCR, compared to gene-level estimates, with relatively higher correlations for fold change values.

This study highlights the need for improved isoform quantification techniques and suggests that fold change values should be preferred to expression estimates when integrating multiple analysis platforms in meta-analyses of gene isoform quantification studies.

MULTIRESOLUTION NONPARAMETRIC BAYESIAN CLUSTER DETECTION AND ASSOCIATION TESTING FOR WHOLE GENOME SEQUENCING STUDIES WITH APPLICATIONS IN PRIMARY IMMUNE DEFICIENCY STUDY.

Jyotishka Datta[1], Anupama Reddy[2], Sandeep S Dave[2]

[1]Duke University, Department of Statistical Science, Durham, NC, [2]Duke University, Duke Center for Genomic and Computational Biology, Durham, NC

Introduction:
Rare variants play a critical role in explaining the genetic contribution to complex diseases by accounting for disease risk and trait variability, previously unexplained by large GWAS. The association of rare variants with disease has proved to be challenging using existing methods that merely compare their frequencies in different datasets. There is need for powerful new statistical methods that incorporate spatial locations of variants, allow incorporation of previous gene ontology information, scale to massive dimensions, and appropriately characterize uncertainty in inferences. Here, we develop multi-resolution, non-parametric Bayesian methods to identify the spatial clustering of the rare variants; leading to substantial increases in power in large association studies.

Methods:
We developed a multiresolution cluster detection method using binary tree to recursively partition the chromosome and prune 'uninteresting' intervals in a top-down fashion. We then developed a novel scalable Bayesian nonparametric methods to draw inference from the point process model. These methods provided several key advantages including robustness, adaptability to the underlying disease architecture, interpretability of clusters, and biologically relevant segmentation of the genome compared to widely applied methods that assess variant frequency. We applied these methods in 218 cases of patients with primary immune deficiency to identify patterns of genetic variation underlying the disease compared to over 7000 controls.

Results:
Our approach had excellent performance in whole genome and exome sequence data, showing fast, accurate detection of clusters and substantial gains in computing speed relative to the scan-statistic approach. In our immune deficiency patients, we identified novel gene mutations in HRNR that may be related to the regulation of BTK, a gene that is critical in signaling and B-cell development. Our methods are extensible across a large range of disease models and provides a number of advantages including scalability, incorporation of important covariates and adjustment for population stratification.

POST-DOMESTICATION GENOMICS OF CANINE POPULATIONS

Brian W Davis[1], Maud Rimbault[1], Brennan Decker[1], Eric Karlins[1], Cord Drögemüller[4], Vidhya Jagannathan[4], Alexandra M Byers[1], Jason J Corneveaux[3], Adam H Freedman[4], Dayna l Dreger[1], Jeffrey M Trent[3], Danielle M Karyadi[1], Heidi G Parker[1], Matthew J Huentleman[3], Tosso Leeb[2], John Novembre[5], Robert K Wayne[6], Elaine A Ostrander[1]

[1]National Human Genome Research Institute, National Institutes of Health, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, [2]University of Bern, Institute of Genetics, Bern, Switzerland, [3]5The Translational Genomics Research Institute, Phoenix, AZ, [4]Harvard University, Department of Organismic and Evolutionary Biology, Boston, MS, [5]University of Chicago, Department of Human Genetics, Chicago, IL, [6]University of California, Los Angeles, Department of Ecology and Evolutionary Biology, Los Angeles, CA

The transition from ancestral species into contemporary domesticated and companion animal populations has been the subject of increasing study within the context of the initial mechanisms contributing to and resulting from their removal from the wild. However, little investigation has been directed towards post-domestication changes resulting from the intense selective breeding efforts imposed to generate the closed breeding populations observed today. To address the question of novel genomic change ushered in by the formation of breeds and breed-groups, we assembled the largest current catalog of canine variation, originating from 245 whole genomes. We leveraged this dataset against a newly constructed panel of SNP data, consisting of 830 individuals originating in 84 closed-breeding and wild populations genotyped on the IlluminaHD platform. Using a reconstructed phylogeny, we recursively examined genomic signals of selection within each clade, and identified regions of extended and punctuated homozygosity within breeds and breed-groups resulting from artificial selection imposed post-domestication and population bottlenecks associated with breed formation. We identified genomic changes coinciding with alterations in basic biology, including morphological, neurological, and immunological pathways unique to each breed; as well as commonalities within our catalog of over 30 million single nucleotide, small insertion-deletion, and structural variants within breeds with common ancestry. Our evaluation of the novel genomic architecture endemic to each breed and group sheds new light on the recent and rapid change imposed on the domestic dog genome as it has been shaped by artificial selection into both functional and aesthetic niches by human intervention.

REGULATORY VARIATION AND THE GENOMIC CONTEXT OF ALLELE-SPECIFIC EXPRESSION

Joe R Davis[1,2], David A Knowles[3], Yungil Kim[4], Mauro Pala[1,5], The GTEx Project Consortium[6], SardiNIA Project[1,5,7,8], Goncalo Abecasis[7], Carlos D Bustamante[2], Francesco Cucca[5], David Schlessinger[8], Stephen B Montgomery*[1,2], Alexis Battle*[4]

[1]Pathology, Stanford University, Stanford, CA, [2]Genetics, Stanford University, Stanford, CA, [3]Computer Science, Stanford University, Stanford, CA, [4]Computer Science, Johns Hopkins University, Baltimore, MD, [5]Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy, [6]MGH, Analytics and Translational Genetics Unit, Boston, MA, [7]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, [8]Laboratory of Genetics, NIA, Baltimore, MD

The regulatory context of allele-specific expression (ASE) has primarily been explored in conjunction with genetic variants underlying expression quantitative trait loci (eQTLs). Few studies have attempted to identify and characterize regulatory variants via their effects on ASE alone. In this study, we assess the use of ASE to evaluate the impact of non-coding genetic variation and present general guidelines to aid future studies. We present one of the largest aseQTL studies yet, covering four genetics of gene expression datasets and representing multiple cell types and global populations. Specifically, we use RNA-Seq and genotype data from 1) the Depression Genes and Networks Study, 2) the GEUVADIS RNA-Seq Project, 3) the SardiNIA Project and 4) the Genotype-Tissue Expression (GTEx) Project. We investigate the effects of technical artifacts such as sample sequencing depth, batch and gene expression level on the ASE signal. Further, we explore the utility of normalization and latent factor correction, which are commonly applied to gene expression data but not to ASE. We then evaluate aseQTL detection methods, comparing the performance of a generalized linear model, accounting for read depth of each sample, to simple non-parametric tests. Additionally, we identify differences between aseQTLs and cis-eQTLs, comparing the results of aseQTL calling without prior information to calling conditioned on variants previously identified as cis-eQTLs and comparing the replication rates of aseQTL and cis-eQTLs. Ultimately, we lay out general guidelines for analyzing ASE in large RNA-seq datasets, and using aseQTLs to identify and characterize regulatory variants. (* co-senior authors)

LINKAGE AND SEQUENCING IN A BRAZILIAN BIPOLAR FAMILY WITH 111 MOOD DISORDER CASES

Simone de Jong[1], Mateus Diniz[2], Shaza Issam Alsabban[1], Andiara de Saloma[2], Ary Gadelha[2], Jose Paya-Cano[1], Peter McGuffin[1], Camila Guindalini[2], Rodrigo Bressan[2], Gerome Breen[1]

[1]King's College London, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [2]Universidade Federal de Sao Paulo, Interdisciplinary Lab of Clinical Neurosciences (LiNC), and Schizophrenia Program (PROESQ), Department of Psychiatry, Sao Paulo, Brazil

Very large families pose unique opportunities and analytical challenges for psychiatric research. We have identified a family with a high prevalence (~30%) of mood disorders in a rural village in Brazil. The family also features diminishing age of onset over generations and assortative mating, whereby many of the marriages in the family are between individuals with a psychiatric disorder. 333 DNA samples were obtained from a broader pedigree of over 900 subjects in order to generate Affymetrix 10K array genotype data. The pattern of inheritance is complex with 32 Bipolar 1 cases, 11 Bipolar 2 and 59 Depression cases as well as 5 Parkinson's disease cases as well as anxiety disorders. Non-parametric linkage was carried out via MERLIN with exponential calculation of the LOD score and parametric with MERLIN and MCLINKAGE. For the MERLIN analyses the pedigree was broken into sub-families and dummy individuals introduced to break loops. Simulations were carried out to validate thresholds and examine the performance of statistics. We identified four genome wide significant and four suggestive linkage regions on chromosomes 1, 2, 3, 11 and 12 for different phenotype definitions. We exome sequenced a subset of cases (n=27) in order to identify rare variation within the linkage regions shared by affected family members. This revealed deleterious variants in 11p15.4 for MDD and 1q21.1-1q21.3 and 12p23.1-p22.3 for all affecteds, involved in cell signaling, adhesion, translation and neurogenesis. Future plans include generation of Illumina PsychArray genotypes for the 333 family members; which will allow further investigation of rare SNVs and CNVs within the linkage regions. In addition, we plan to examine inheritance of polygenic risk scores and use them to study patterns of anticipation and assortative mating.

GENOME DATA AGGREGATION AND EXCHANGE ACROSS DISTRIBUTED GENOMIC DATA REPOSITORIES

Francisco M De La Vega[1], Ying Wu[1], Tal Shmaya[2], James Wiley[2], Akshay Patel[2], Raja Hayek[2]

[1]Annai Systems, Inc., R&D, Burlingame, CA, [2]Annai Systems, Inc., R&D, Carlsbad, CA

Through advances in high-throughput sequencing we can now survey the prevalence of both rare variants in the human genome and somatic mutations in cancer tumors. However, achieving the statistical power necessary to understand the roles of these rare variants in complex disease and of somatic mutations in the etiology of cancer will require very large sample sizes. Several large-scale sequencing projects with sizes ranging from tens to hundreds of thousands of samples are now being started in proximity to clinical care centers around the world. Realizing the true value of these massive projects will require transparent and highly available access for researchers and clinicians alike, but the scale and distributed nature of these projects will make it impossible to amass all of the data in a single location or cloud for analysis and sharing for both research and clinical applications. Enabling researchers to access and exchange this data will impose difficult demands in data management, metadata indexing, and analysis to which current paradigms of data distribution and access are inadequate. Here we present an informatics architecture to enable the deployment of distributed genomic data repositories allowing the indexing and discovery of raw sequencing data and genetic variants across a network and enable controlled and secure sharing for aggregate analysis and translational applications. The system leverages a standardized metadata schema and provides synchronization and automatic indexing. We deployed this architecture in support of an international effort to identify common patterns of variation among diverse tumor types from data of about 2,500 tumor-normal sample pairs. Here, the data is federated across eight private cloud analysis centers, including: Barcelona Supercomputing Center, BioNimbus, DKFZ, EBI, ETRI, Riken, UCSC, and Annai-ShareSeq. This highly distributable model is scalable and demonstrates the potential of a global genomic data exchange in advancing translational research and ultimately genomic medicine.

BAAL-CHIP: ALLELE-SPECIFIC CHIP-SEQ ANALYSIS FROM CANCER CELL LINES

Ines de Santiago*, Wei Liu*, Ke Yuan, Kerstin B Meyer, Bruce A Ponder, Florian Markowetz

University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom

We describe a statistical approach for the characterisation of allele-specific regulatory SNPs from ChIP-seq data acquired from cancer cell lines. Our approach, termed Baal-ChIP (Bayesian Analysis of Allelic imbalances from ChIP-seq data) uses a Bayesian framework to model the joint effect of read mapping biases and the relative background allele composition on the observed allele read counts obtained from ChIP-seq data. Baal allows the interrogation of multiple ChIP-seq datasets across a singe variant simultaneously and performs well in simulations. We applied this method to 248 ENCODE samples obtained from a panel of 7 cancer and 5 normal cell lines and observed that the majority of the allelic imbalances in cancer cell lines can be explained by imbalances in the background allele frequency due to genomic copy number alterations rather than true motif or sequence regulatory effects. We find that 80% of the identified variants are non-coding, with 60% mapping to cell-type specific enhancers. Baal-ChIP illustrates the value of taking into consideration structural genomic alterations in other to detect putative cis-acting regulatory variants in cancer cell lines.

UNDER THE RADAR: SURVIVAL STRATEGIES OF AN ANCIENT CLONALLY TRANSMISSIBLE CANINE TUMOR

Brennan Decker[1,2], Brian W Davis[1], Maud Rimbault[1], Adrienne H Long[3], Eric Karlins[1], Vidhya Jagannathan[4], Rebecca Reiman[5], Heidi G Parker[1], Cord Drögemüller[4], Jason J Corneveaux[5], Erica S Chapman[1], Jeffery M Trent[5], Tosso Leeb[4], Matthew J Huentelman[5], Robert K Wayne[6], Danielle M Karyadi[1], Elaine A Ostrander[1]

[1]National Human Genome Research Institute, National Institutes of Health, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, [2] School of Clinical Medicine, University of Cambridge, Department of Public Health and Primary Care, Cambridge, United Kingdom, [3]National Cancer Institute, Center for Cancer Research, National Institutes of Health, Pediatric Oncology Branch, Bethesda, MD, [4]University of Bern, Institute of Genetics, Bern, Switzerland, [5]The Translational Genomics Research Institute, TGEN, Phoenix, AZ, [6]University of California, Los Angeles, Department of Ecology and Evolutionary Biology, Los Angeles, CA

Canine transmissible venereal tumor (CTVT) is a parasitic cancer clone that has propagated for thousands of years via direct sexual transfer of malignant cells from one canid to another. Little is understood about the genomic mechanisms that converted an ancient tumor into the world's oldest known continuously propagating somatic cell lineage. We created the largest existing catalog of canine genome-wide variation and compared it against two CTVT genome sequences, thereby separating alleles derived from the founder's genome from somatic mutations, some of which must drive clonal transmissibility. Variant metrics including transition to transversion ratio, nonsynonymous to synonymous ratio, and conservation across species at variant sites all support the hypothesis that novel CTVT variants are dramatically enriched for true somatic mutations. Gene Set Enrichment Analysis of the 1,341 protein-truncating somatic substitutions and indels, as well as the 2,329 protein-disrupting structural variants revealed the greatest enrichment for somatic mutations in the Reactome "Immune System" pathway, with p-values of 1.35E-12 and 1.48E-19, respectively. Upon further investigation of this pathway, we observed overlapping mutations at every step of somatic cell participation immunosurveillance, especially self-antigen presentation and apoptotic pathways. Unbalanced events in CTVT's highly rearranged genomic architecture allowed identification of chronologically early somatic mutations in oncogenesis- and immune-related genes that likely represent key initiators of clonal transmissibility. We also rebuilt surviving elements of the founder canid's germline genome, and the resultant genome-wide maximum likelihood phylogeny supported CTVT origination in an ancient domesticated dog genetically similar to modern Siberian huskies. It is clear that CTVT is exquisitely adapted to its transmissible allograft niche, and we provide the first insights into the specific genomic aberrations that underlie CTVT's dogged perseverance in canids around the world.

GENETIC CONTROL OF CHROMATIN IN A HUMAN POPULATION

Olivier Delaneau[1], Sebastian Waszak[2], Andreas Gschwind[3], Helena Kilpinen[1], Sunil Raghav[2], Robert Witwicki[3], Andrea Orioli Orioli[3], Michael Wiederkehr[3], Maria Gutierrez-Arcelus[1], Nikos Panousis[1], Tuuli Lappalainen[1], David Hacker[2], Nouria Hernandez[3], Alexandre Raymond[3], Bart Deplancke[2], Emmanouil Dermitzakis[1]

[1]University of Geneva, Department of Genetic Medicine and Development, Geneva, Switzerland, [2]EPFL, Laboratory of Systems Biology and Genetics, Lausanne, Switzerland, [3]University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland

Non-coding regulatory DNA variants have been found to be associated with gene expression, yet the precise molecular basis by which they act remains elusive. We hypothesize that an integrated study of chromatin states coupled with personal genome information might enable in-depth characterization of these regulatory variants. We quantified gene expression (mRNA), genome-wide DNA binding of two regulatory proteins (RNA polymerase II, PU.1) and three histone post-translational modifications, that pinpoint promoter, enhancer, and active regions (H3K4me3, H3K4me1, and H3K27ac, respectively) in lymphoblastoid cell lines of 47 unrelated individuals that were whole genome-sequenced as part of the 1000 genomes project. This data allowed us to assess the degree of inter-individual coordinated behavior between distinct layers of gene regulation and to map cis-acting QTLs for each molecular phenotype. We find that various molecular phenotypes show abundant quantitative coordination in activity levels at enhancer and/or promoter elements, forming 'chromatin modules' that can extent over hundreds of kb, regroup hundreds of chromatin sites and correspond to domains of physical interaction as defined via Hi-C. We show that the overall chromatin activity at these modules is tightly correlated with changes in expression levels at nearby genes, highlighting their central role in gene expression variation. We further mapped thousands of cis-acting QTLs for gene expression, histone modification levels and TF binding at 10% FDR. We find that they are widespread across the genome and that they explain a substantial fraction of inter-individual variability in chromatin activity. In addition, we find large overlaps between various QTLs which reflects the genetic signal propagation through multiple phenotypic layers and show that the genetic perturbation of chromatin tend to be causal to changes in gene expression levels. Finally, we also show that chromatin QTLs are strongly enriched in referenced GWAS hits underscoring their relevance in interpretation of complex disease genetics. Overall, this large-scale study that integrates DNA, RNA, regulatory proteins and histone modifications provides novel insights into the mechanisms underlying regulatory variation and their effects on transcription.

VCF.IOBIO : A VISUALLY DRIVEN VARIANT DATA INSPECTOR AND REAL-TIME ANALYSIS WEB APPLICATION

Tonya L Di Sera, Chase A Miller, Yi Qaio, Jon Anthony, Alistair Ward, Gabor Marth

USTAR Center for Genetic Discovery, University of Utah School of Medicine, Eccles Institute of Human Genetics, Salt Lake City, UT

Our new, web-based analysis system, IOBIO (http://iobio.io) is designed to empower biological researchers to easily, interactively, and in a visually driven manner, analyze those portions of vast biomedical datasets that are essential for their research. As part of this effort, we have developed a novel data inspector web application vcf.iobio (http://vcf.iobio.io) that, together with our recently published sequence alignment inspector app (http://bam.iobio.io), is designed to present genomic sequence variation data in a visually compelling way, promoting exploration and real-time interaction. With these data inspector apps, researchers have instant access to global and regional "vitals" by simply launching our app from any modern web-browser and selecting (but not uploading) a variant (VCF) or alignment (BAM), either stored locally on the researcher's hard drive or remotely, e.g. on cloud storage. The vcf.iobio app displays, within seconds, vital statistics of the variant collection in the form of graphical charts like histograms and area charts. These metrics include variant density; Ti/Tv ratio; and the distribution of variant types, insertion and deletion lengths, variant quality, allele frequency, and base changes. The researcher is then able to select a chromosomal sub-region for further analysis and within seconds, the statistics are recomputed. Similarly, our bam.iobio app inspects sequence alignments and displays overall read coverage; the average and distribution of fragment length, mapping and base quality; and ratios of mapped reads, duplicates, and proper pairs. Instead of processing all data points (variant or alignment records) in the file, we collect and process a random sample of these records sufficiently large to generate reliable statistics, representative of the entire dataset, thereby making instantaneous feedback possible.

Because the analysis is performed on the cloud, the computing load on the researcher's computer is minimal, just sufficient to support client-side visualizations. We are currently developing a complementary app, gene.iobio (http://gene.iobio.io), that will permit researchers to examine all details of sequence alignments, genetic variants, and variant annotations in the region of a gene, and to interactively call and re-call variants from this same application. These data inspector apps lay the foundation for further app development and demonstrate that viability of real-time, visually driven analysis in the context of biomedical / genomic big data collections.

GENETIC AND EPIGENETIC SIGNATURES OF GENE REGULATION SPECIFIC TO TYPE 2 DIABETES-RELEVANT TISSUES

John P Didion[1], Stephen C Parker[2], Brooke N Wolford[1], Jeroen R Huyghe[3], Ryan Welch[2], Michael R Erdos[1], Peter S Chines[1], Narisu Narisu[1], Laura J Scott[3], Michael Stitzel[4], Michael Boehnke[3], Francis S Collins[1]

[1]NIH, NHGRI, Bethesda, MD, [2]University of Michigan, Depts. of Computational Medicine & Bioinformatics and Human Genetics, Ann Arbor, MI, [3]University of Michigan, Dept. of Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, [4]The Jackson Laboratory, Genomic Medicine, Farmington, CT

Multiple tissues are involved in glucose metabolism and the etiology of type 2 diabetes (T2D), including skeletal muscle, adipose, and pancreatic islets. Although hundreds of genetic variants have been associated with T2D and related traits, the tissue-specific effects of those variants are not well understood. To identify gene expression signatures in T2D-relevant tissues, the Finland United States Investigation of NIDDM Genetics (FUSION) Study obtained muscle and adipose biopsies from 278 Finnish individuals, and pancreatic islets from 78 unrelated cadaveric donors. We performed dense genotyping and deep mRNA sequencing on all muscle and islet samples; mRNA sequencing of adipose samples and whole-methylome sequencing of a subset of samples is in progress. In addition, we obtained genotype, RNA-seq and methylation data for T2D relevant tissues (including brain, liver, stomach, small intestine) from published studies and the Genotype-Tissue Expression (GTEx) project. We process all samples using a pipeline for imputation, transcriptome assembly, analysis of differential expression (including allele-specific expression) and splicing. We identify transcripts at the exon, gene, isoform and allelic levels that are enriched in each tissue, including hundreds of apparently novel transcripts. We identify tissue-specific expression and splicing quantitative trait loci (e/sQTL), and we use reference chromatin state maps to quantify the proportion of e/sQTLs that occur in tissue-specific chromatin states. We have already found significant enrichment of SNPs associated with T2D and related traits in enhancer regions of skeletal muscle and islets. In addition, we will identify sites of differential methylation that are associated with e/sQTLs in a tissue-specific manner (mQTLs), which will suggest specific links between genetic and epigenetic variation in gene regulation. Because we have roughly equal numbers of male and female samples, we will also be able to identify sex-specific effects on expression and methylation. This resource will form the basis for further understanding the role of functional and regulatory variation in the contribution of each tissue to the etiology of T2D.

SEQUENCING OF FULL-LENGTH RNA TRANSCRIPTS ON THE OXFORD NANOPORE PLATFORM.

Alexander Dobin, Sara Goodwin, Lee-Hoon See, W. Richard McCombie, Thomas R Gingeras

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

The sequencing of transcripts longer than a few hundred nucleotides present multiple challenges that cannot be resolved by the existing high-throughput sequencing technologies. The computational approaches that assemble transcripts from the RNA-seq data often yield inaccurate results owing to the inherent lack of long-range connectivity information in short reads. Other confounding challenges are the very large range of copy number variation for individual RNA transcripts, post-transcriptional RNA processing, and sequence similarity of alternative isoforms and their paralogs. Despite the relatively high base calling error rates and relatively low throughput/productivity, the long-range sequencing approach such as Oxford Nanopore Technology (ONT), combined with novel mapping strategies, holds the promise of providing full-length connectivity information for long transcripts.
Here we report the application of the ONT MinION portable instrument for sequencing cDNAs from human cell lines that have been previously extensively characterized using short-length Illumina reads. Multiple aligners and mapping strategies were assessed to develop the optimum computational pipeline that utilizes short Illumina reads to improve mappability of the long but error-prone ONT reads, and allows for detection of both annotated and novel full-length transcripts. We generated ~40k two-directional (2D) reads per MinION flowcell, of which ~60% can be mapped to human transcriptome with ~1kb median mapped length, and ~22% median error rate. More than 10,000 annotated transcripts were detected, and the ONT reads demonstrated good precision of the transcription start site recovery (deviation ~200 nt) and very high precision of the transcription termination site recovery (deviation <50b). In addition, we identified more than a thousand previously unannotated transcripts, of which ~700 were novel extensions of annotated 5'/3' UTRs, and ~600 were novel splicing isoforms. We also compared the ONT results with those obtained from the Pacific Biosciences sequencer for the same RNA sample, contrasting the strengths and weaknesses of the two long-read technologies. We demonstrate that both technologies can be used in conjunction with short Illumina reads to augment and/or correct the annotations of thousands of human genes.

PRIORITIZING LIKELY CAUSATIVE GENES IN GWAS IDENTIFIED RISK LOCI FOR IMMUNE-MEDIATED INFLAMMATORY DISORDERS USING CELL-TYPE SPECIFIC EQTL INFORMATION.

Elisa Doçampo[1], Julia Dmitrieva[1], Ming Fang[1], Emilie Théâtre[1], Mahmoud Elansary[1], Rob Mariman[1], Ann-Stephan Gori[1], Edouard Louis[2], Michel Georges[1]

[1]GIGA-R & Faculty of Veterinary Medicine, ULg, Unit of Animal Genomics, Liège, Belgium, [2]University Hospital CHU of Liège, Department of Gastroenterology, Liège, Belgium

GWAS have identified hundreds of risk loci for most studied immune-mediated inflammatory disorders (IMIDs). However, typical risk loci span 4-5 genes on average (range: 0 - > 50), and the causative genes identified in only a handful of cases. Formally identifying causative genes remains essential to reap the full benefits of GWAS.
Recent fine-mapping efforts indicate that only a minority of risk variants are coding. This suggests that most risk variants will be regulatory hence affecting disease risk via eQTL effects. To aid in the identification of causative genes for IMID, we therefore generated transcriptome information (HT12 arrays) for six leucocyte populations (CD4, CD8, CD19, CD14, CD15 and platelets) and intestinal biopsies at three anatomical locations (ileum, colon, rectum) for 350 healthy Caucasian individuals. The same individuals were genotyped with SNP arrays interrogating > 700K variants, augmented by imputation from the 1KG project. We identified > 16000 highly significant cis-eQTL. The degree of eQTL sharing between cell types ranged from 38 to 90% highlighting the utility of our multi-tissue panel.
To identify likely causative genes in GWAS identified risk loci, we (i) developed a method that quantifies the correlation between “disease association pattern” (DAP) and “eQTL association pattern” (EAP) and provides an empirical estimate of its significance, and (ii) evaluated the effect of fitting known risk variants as covariates in the eQTL analysis following Nica et al. (2010). We applied both approaches to celiac disease (data provided by Cisca Wijminga) and rheumatoid arthritis (data provided by Soumya Raychaudhuri),and the second one additionally to type one diabetes, multiple sclerosis, systemic lupus erythematosus, ankylosing spondylitis and psoriasis. We obtained a large number of striking DAP-EAP correlations pinpointing likely disease-specific as well as shared causative genes.
Latest results will be presented.

INSIGHTS INTO THE CONSEQUENCES OF SEQUENCE DIVERGENCE USING HIGH-THROUGHPUT POOLED ALLELE REPLACEMENTS

Drew T Doering[1,2], Chris T Hittinger[1,2]

[1]University of Wisconsin-Madison, Laboratory of Genetics, Madison, WI, [2]University of Wisconsin-Madison, Graduate Program in Cellular & Molecular Biology, Madison, WI

The mutation of a single nucleotide in a gene can have wide-ranging effects on the function of its protein product. Understanding the functional consequences of sequence divergence can enable researchers to predict a protein's function in other organisms or study the functional equivalency of protein sequence variation found across various evolutionary distances, including between rare genetic variants present in a population. Using *Saccharomyces* as a model genus, my goal is to understand the functional equivalency of *ATX1*, a gene encoding a cytosolic metal cation chaperone with antioxidant activity, from all seven *Saccharomyces* species as well as its homologs in other yeast and multicellular eukaryotes (including humans). I will use genome editing technologies to conduct pooled swaps of *ATX1* and its homologs, grow the cells in selective conditions, and then sequence the final pool of transformants to evaluate changes in allele frequencies. This change will then be used to derive a fitness value for each homolog, which can be used to compare the effect of a given genetic variant on protein function. This technology can be used to study genes implicated in human disease and serve as a platform for diagnosis of heritable diseases that are due to hypomorphic alleles that are present at low frequencies in the human population. With the emergence of personalized medicine enabled by affordable genome sequencing, physicians could use data generated by this type of study to diagnose a wide variety of genetic diseases using only the patient's genome sequence.

HIGHER MALE THAN FEMALE RECOMBINATION RATE IN CATTLE IS CONTROLLED BY GENETIC VARIANTS EFFECTIVE IN BOTH SEXES

Naveen K Kadri[1], Chad Harland[1,2], Wouter Coppieters[1], Sébastien Fritz[3], Didier Boichard[4], Richard Spelman[2], Chris Schrooten[5], Erik Mullaart[5], Carole Charlier[1], Michel Georges[1], Tom Druet[1]

[1]University of Liège, Unit of Animal Genomics, Liège, Belgium, [2]LIC, Hamilton, New Zealand, [3]UNCEIA, Paris, France, [4]INRA, GABI, Jouy-en-Josas, France, [5]CRV, Arnhem, Netherlands

We herein study genetic recombination in three dairy cattle populations from France, New-Zealand and the Netherlands. We apply a new phasing algorithm extracting familial information suited for large half-sib families to reconstruct haplotypes and detect cross-overs (CO). The software is robust to genotyping and map errors. We identify more than 2,000,000 CO events in sperm cells transmitted by 3008 sires to 94,603 offspring, and more than 500,000 CO events in oocytes transmitted by 11,497 cows to 25,390 offspring. When measured in identical family structures, the average number of CO in males (24.0) was found to be larger than in females (21.8). In males, recombination rates were higher closer to telomeres whereas in females, recombination rates dropped at both centromeres and telomeres (probably as a result of lower informativity). The heritability of the global recombination rate (GRR) was close to 0.20 in males and to 0.08 in females. Genetic correlation ranged from 0.38 to 0.69 depending on the population, indicating that shared variants are influencing GRR in both genders. Haplotype-based genome-wide association studies revealed four genome-wide significant QTLs, including two previously identified ones (involving *REC8* and *RNF212*). For all QTLs, there was a positive correlation between haplotype effects across sexes, ranging from 0.35 to 0.68. We selected two reference panels of respectively 122 and 215 bulls sequenced at cover > 15x to impute variants in the New-Zealand and French populations. All variants identified by next-generating sequencing in 5 Mb windows encompassing the QTL peaks were imputed with Beagle in order to perform a sequence-based association study. For three QTLs, we identified missense mutations in genes known to be involved in meiosis among the most significantly associated variants. These variants were perfectly associated with the haplotypes underlying the QTL effects. The variant identified in *RNF212* had already been reported, whereas missense mutations in *MLH3* (*N408S*) and *HFM1* (*S1189L*) are new findings. Surprisingly, variants previously identified in *REC8* did not capture the QTL effect whereas variants in *RNF212B*, *PPP1R3E*, *BCL2L2*, *HOMEZ* and *PABPN1* had much stronger association with the phenotype. The three missense mutations were significant in both genders with two of them accounting for approximately 10% of the genetic variance in males (the allelic substitution effect being approximately equal to one additional CO per genome). Our results are very different from reports of recombination in other species. For instance, in human, recombination rate is higher in females, distinct variants affect recombination rate in males and females and the genetic correlation is close to 0 whereas in cattle, we observed a higher recombination rate in males controlled by shared variants effective in both sexes.

65,222 WHOLE GENOME HAPLOTYPES FROM THE HAPLOTYPE REFERENCE CONSORTIUM AND EFFICIENT ALGORITHMS TO USE THEM

Richard Durbin[1], on behalf of the Haplotype Reference Consortium[2]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, [2]multiple institutions, in multiple countries, including USA, Netherlands, Finland, United Kingdom

Genotype imputation is central to modern genome wide association studies and other uses of partial or noisy genetic data. By predicting unobserved genotypes, it increases the resolution of the study, facilitates collaborative efforts to find disease genes via meta-analysis and increases statistical power. The basis for imputation is a haplotype reference panel, such as those produced by HapMap and the 1000 Genomes Project (1000GP). The range of allele frequencies that can be accurately imputed depends on panel size: the 1000GP panel is effective down to around 1% or a little lower minor allele frequency (MAF).

The Haplotype Reference Consortium has combined data from over 30,000 individuals from 20 low- to medium- coverage whole genome sequencing studies to construct a new haplotype panel for imputation. Following site and sample QC we are making a final call set on approximately 45 million SNPs with minor allele count at least five, homogeneously recalled and phased from primary data, expecting to release in March 2015. On a test chromosome can now achieve good imputation down to 0.1% MAF.

The resource will be available for imputation via servers at the University of Michigan and the Sanger Institute. Although resolution increases with larger panels, compute time for standard methods also increases. However, imputation is based inherently on haplotype matching, and the Sanger Institute imputation server makes use of new efficient methods based on the Positional Burrows-Wheeler Transform (PBWT). This supports haplotype matching to a highly compressed representation of the reference data in time independent of the panel size, as with BWT or hash-based sequence matching into a sequence database, and makes haplotype based analyses feasible on extremely large sample size data sets. I will describe the panel, its performance for imputation, and the underlying PBWT methods including other potential applications.

IMPROVING PROVIRAL INTEGRATION SITE DETECTION WITH HIGH THROUGHPUT SEQUENCING

Keith Durkin[1], Maria Artesi[1], Nicolas Rosewick[1,2], Michel Georges[1], Anne Van den Broeke[1,2]

[1]Universite de Liege, Unit of Animal Genomics, Liege, Belgium, [2]Universite libre de Bruxelles, Experimental Hematology, Brussels, Belgium

Bovine Leukemia Virus (BLV) and the closely related Human T-cell leukemia virus-1 (HTLV-1) are deltaretrovirus that induce leukemia/lymphoma in about ~5% of infected individuals. The mechanisms responsible for cellular transformation have remained largely enigmatic as both viruses are largely transcriptionally silent in tumors and show apparently random integration in the host genome. The recent application of high throughput sequencing to track proviral insertion sites in the host genome has provided a number of insights into the evolution of deltaretrovirus infections and the progression of tumor clones in deltaretrovirus induced leukemia/lymphoma. However the protocols currently utilised have a number of limitations, including relatively high sequencing costs, the use of custom sequencing primers, no examination of the region upstream of the provirus and limited dynamic range for determining clone abundance. We have developed an alternative high throughput sequencing protocol for identifying proviral integration sites in BLV and HTLV-1 infected individuals that uses off-the-shelf Illumina primers for the addition of adapters and indexes. This greatly simplifies the process of multiplexing libraries, does away with the need for custom sequencing primers and also makes it much easier to multiplex libraries. Additionally our approach assays the region upstream of the provirus in addition to the downstream region, giving additional information on the frequency of 5' deletions in proviruses and increasing the dynamic range of the assay. We have tested the approach on over 100 BLV and HTLV-1 samples, representing both tumors and preleukemic stages. Our approach allowed for a more accurate determination of clone abundance in tumors and by assaying the 5' end of the provirus identifies clones overlooked with previously published methods. Finally, by facilitating greater multiplexing of libraries we have reduced the cost to a level where the technique may be attractive in a clinical setting.

EFFECTS OF TRANS-eQTLs ACROSS MANY HUMAN TISSUES IN THE CONTEXT OF REGULATORY NETWORKS

Barbara E Engelhardt[1], Alexis J Battle[2]

[1]Princeton University, Computer Science, Princeton, NJ, [2]Johns Hopkins University, Computer Science, Baltimore, MD

The genetics of gene regulation is critical to understand because of the mechanistic implications for the genetic regulation of complex traits and disease risk: expression quantitative trait loci, or eQTLs, are enriched for polymorphisms that have been found to be associated with disease risk via genome-wide association studies. However, for a number of reasons, local, or cis-, eQTLs have received the bulk of scientific attention as opposed to distal, or trans-, eQTLs. In particular, cis-eQTL SNPs are in close proximity to the genes that they regulate, allowing for many orders of magnitude fewer statistical tests for association mapping; furthermore, in human studies thus far, cis-eQTLs tend to have larger effects than trans-eQTLs and are more reproducible across studies and tissue types. However, recent work has suggested a greater role for trans-eQTLs as opposed to cis-eQTLs in complex disease, necessitating a more comprehensive understanding of these distal effects. In this work, we characterize trans-eQTLs within the Genotype-Tissue Expression (GTEx) study data, consisting of over 400 individuals with RNA-sequencing samples across 44 tissue types. First, we identify trans-eQTLs using statistical approaches that share strength across multiple tissues and samples. Using these data, we show the spectrum of estimated effects of trans-eQTLs, and compare these with the spectrum of estimated effects for cis-eQTLs. Next, we characterize the tissue specificity of trans-eQTLs relative to cis-eQTLs in these data, and quantify specific discrepancies in trans-eQTL discovery using power estimates of multi-tissue association mapping. Finally, we consider the role of gene co-expression networks in the discovery and characterization of trans-eQTLs, evaluating the simplest hypothesis that the primary mechanism of trans-eQTL SNPs is that they first regulate expression of a proximal gene, which itself participates in regulation of one or more distal genes. We quantify this statement in the context of available gene interaction networks and our collection of trans-eQTLs. These analyses provide a more comprehensive estimate of the effects of trans-eQTLs on gene expression in diverse human tissues, which contributes greater understanding of the tissue-dependent cellular consequences of disease-associated genetic variation.

AVIANBASE: ENABLING COMPARATIVE GENOME ANALYSES OF BIRDS

Lel Eory[1], Bronwen L Aken[2,3], Alan Archibald[1], Paul Flicek[2,3], David W Burt[1]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Genomics and Genetics, Edinburgh, United Kingdom, [2]European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge, United Kingdom, [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

The full genome sequences of 44 birds, recently released by the Avian Phylogenomics Consortium (APC), have opened up the possibility of looking for regions within these genomes which are of functional relevance. On the one hand - at the level of individual species - the sequences, coupled with the initial annotations provided by the APC, can serve as a vehicle for basic research. On the other hand integrating the available data, e.g. by generating multiple sequence alignment for these species, can enable comparative studies which benefits all the species included. Such studies in general can broaden our understanding of genome evolution and the evolution of phenotipic characteristics within birds, or between birds and other taxa, or can help to disentangle the phylogenetic relationship of species.

To enable the integration of bird specific resources, we created Avianbase, an Ensembl based resource, with sequence and annotation data for a total of 48 birds. Our main aim is to analyse the integrated data with a focus on creating a detailed functional map which is relevant to birds, by using methods which detect signatures of selection. Such a map is in turn to be used to drive the identification of novel protein-coding and non-coding genes or binding sites for miRNAs, transcription factors or enhancer and silencer sequences. This level of annotation can play an important role in identifying causative variants which affect putatively functional genomic locations and so can help to make the link between genotypes that underline certain phenotypic differences in individuals.

THE EXPANSION OF HUMAN POPULATIONS OUT OF AFRICA MIGHT HAVE LED TO THE PROGRESSIVE BUILD-UP OF A RECESSIVE MUTATION LOAD

Brenna M Henn[1], Laura R Botigué[1], Stephan Peischl[2,3], Isabelle Dupanloup[2,3], Mikhail Lipatov[1], Brian K Maples[4], Alicia R Martin[4], Muh C Yee[4], Howard Cann[5], Michael Snyder[4], Jeffrey M Kidd[6], Carlos D Bustamante[4], Laurent Excoffier[2,3]

[1]Stony Brook University, Department of Ecology and Evolution, Stony Brook, NY, [2]University of Berne, Institute of Ecology and Evolution, Berne, Switzerland, [3]Swiss Institute of Bioinformatics, Computational Population Genetics, Lausanne, Switzerland, [4]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [5]Fondation Jean Dausset, Centre d'Etude du Polymorphisme Humain, Pais, France, [6]University of Michigan Medical School, Department of Human Genetics and Department of Computational Medicine and Bioinformatics, Ann Arbor, MI

The Out-of-Africa (OOA) dispersal of modern humans ~50,000 years ago is characterized by a series of successive founder events. Population genetics theory predicts that such range expansions should lead to an increase in the proportion of deleterious alleles with distance from the origin, and thus to the build-up of an expansion load. To test this hypothesis, we have sequenced full genomes and high-coverage exomes from over 50 individuals from 7 human populations, establishing a picture of genomic diversity in geographically divergent groups from Namibia, Congo, Algeria, Pakistan, Cambodia, Siberia and Mexico. In line with several recent studies, we find that individual genomes vary modestly in the overall number of predicted deleterious alleles they carry, but we detect clear signatures of purifying selection at the population-level. Specifically, OOA populations show, on average, higher frequencies of deleterious alleles than African populations, and we find that for a majority of deleterious mutations, the population site frequency spectrum shifts increasingly to the right with distance from Africa. We show via spatially explicit simulations that these patterns are consistent with the Out-of-Africa dispersal, particularly under a model where deleterious mutations are recessive. This recessivity is supported by an observed enrichment of deleterious alleles in OMIM genes conferring pathogenicity through a recessive model of inheritance. We conclude that there is a strong signal of purifying selection at conserved positions within Africa, but that most predicted deleterious mutations have evolved as if they were neutral during the expansion out of Africa, where they might have been counter-selected in homozygotes more recently. OOA populations are thus likely to have a higher mutation load due to increased allele frequencies of mildly deleterious variants that are recessive or partially recessive.

HUMAN EPIGENOMIC VARIATION IS DRIVEN BY HISTORICAL AND RECENT CHANGES IN HABITAT AND LIFESTYLE

M Fagny[1,2], E Patin[1], J L MacIsaac[3], M Rotival[1], T Flutre[4], M J Jones[3], H Quach[1], C Harmant[1], L M McEwen[3], A Froment[5], E Heyer[6], A Gessain[7], G H Perry[8], L B Barreiro[9], M S Kobor[3], L Quintana-Murci[1]

[1]Institut Pasteur, Human Evolutionary Genetics, Paris, France, [2]UPMC, Cellule Pasteur, Paris, France, [3]University of British Columbia, Centre for Molecular Medicine and Therapeutics, Vancouver, Canada, [4]INRA, UMR AGAP, Montpellier, France, [5]Sorbonne Universités, IRD-MNHN, UMR208, Paris, France, [6]Université Paris Diderot, MNHN, CNRS UMR7206, Paris, France, [7]Institut Pasteur, Unité d'Epidémiologie et Physiopathologie des Virus Oncogènes, Paris, France, [8]Pennsylvania State University, Depts of Anthropology and Biology, University Park, PA, [9]Université de Montréal, Centre de Recherche CHU Sainte-JustineMontréal, Canada

The epigenomic landscape of the human genome, in particular DNA methylation, is increasingly recognised as an important driver of phenotypic diversity. However, the relative impacts of DNA sequence variation and temporal changes in lifestyle and ecological habitat on the epigenome remain unknown. Here, we generated whole-blood genome-wide DNA methylation and genetic profiles for 352 individuals from 5 populations of African rainforest hunter-gatherers and sedentary farmers differing in their present or historical lifestyles and habitats. We found that historical and current differences in lifestyle and habitat have similarly profound impacts on the global methylome. However, the biological functions affected and the mechanisms underlying DNA methylation variation differed strongly. Methylation variation between populations living in different habitats, but sharing the same historical lifestyle and genetic background, mostly involved genes related to immune system functions. Conversely, methylation variation between populations that share the same current environment, but differ in their historical lifestyle and genetic background, involved genes related to developmental processes. Importantly, methylation variation due to historical differences was strongly enriched in associations with nearby SNPs (meQTLs), contrary to variation due to current differences in environment, which was depleted in such associations. Moreover, meQTLs associated with historical differences in lifestyle explain a larger part of the interindividual variance in methylation levels and have been privileged targets of natural selection, with respect to those associated with recent shifts in environmental exposure. Together, these findings provide new insight into the contribution of epigenetic modifications and DNA sequence variation to the adaptation of human populations to changes in lifestyle and environment over different time scales.

TARGETED HIGH THROUGHPUT SEQUENCING IDENTIFIES NOVEL DISEASE CANDIDATE GENES FOR SYSTEMIC LUPUS ERYTHEMATOSUS IN SWEDISH PATIENTS.

Fabiana H G Farias[1], Maria Wilbe[2], Johanna Dahlqvist[1], Sergey V Kozyrev[1], Dag Leonard[3], Gerli R Pielberg[1], Helene Hansson-Hamlin[4], Göran Andersson[2], Maija-Leena Elorata[3], Lars Rönnblom[3], Kerstin Lindblad-Toh[1,5]

[1]Uppsala University, Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala, Sweden, [2]Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Uppsala, Sweden, [3]Uppsala University, Science for Life Laboratory, Department of Medical Sciences, Section of Rheumatology, Uppsala, Sweden, [4]Swedish University of Agricultural Sciences, Department of Clinical Sciences, Uppsala, Sweden, [5]Broad Institute, Cambridge, MA

Systemic lupus erythematosus (SLE) is an autoimmune disorder with heterogeneous clinical manifestations. The etiology of SLE is complex, involving interplay between genetic and environmental factors. Although many loci have been identified, only a small proportion of heritability of SLE is explained, indicating a demand for different approaches for discovery of new genetic disease variants. The aim of this study was to identify novel rare genetic variants relevant for SLE. Roche NimblegGen capture array was used to target 219 genes selected on basis of their role in immune response, autoimmunity, known association with human SLE and SLE-related disease in dogs. We sequenced a cohort of 144 Swedish SLE patients divided in nine pools and 17 healthy controls that were combined to one pool. The pools were paired-end sequenced using Illumina HiSeq2000 reaching approximately 240X average coverage of target region per individual. Variants were filtered to be present only in cases by excluding variants in controls, 1000 genomes and dbSNP. Variants in genes enriched for SNPs were selected for further characterization in terms of conservation, DNase I hypersensitivity, ENCODE data on histone marks and ChIP-Seq peaks. We selected the three best candidate variants located in the genes *MEF2D*, *TCRA*, and *HAPLN3*. The regulatory potential of the novel variants was investigated with EMSA and luciferase reporter assays, there was indication that our candidate variants may influence gene expression. Genotyping of 742 additional Swedish controls revealed a strikingly higher frequency of the variants in SLE patients. We also observed a correlation between particular disease manifestations and the novel variants. Our study presents a successful strategy to detect novel rare disease variants associated with complex diseases.

RUFUS: REFERENCE FREE VARIANT DETECTION

Andrew Farrell, Gabor T Marth

USTAR Center for Genetic Discovery, Human Genetics, Salt Lake City, UT

K-mer based approaches have been recently developed for rapid species identification in metagenomic analyses, and for identifying genetic variants between pairs of genomes. Because k-mer approaches make it possible to focus directly on sequences that are different between two genomes, these approaches are especially well suited for rapid and unbiased variant identification. Rapid because the analysis involves only the tiny fraction of the data harboring genetic variants, the far larger portion of the data representing sequence shared between the genomes is disregarded. Unbiased, because analysis of the data does not require mapping to an organismal reference sequence but can be carried out in a completely reference-free fashion, avoiding the many mapping biases that prevent the detection of genetic variations in highly diverged genomic regions. Here we report a general approach, RUFUS, for detecting short variants (SNPs, INDELs), larger structural variants, as well as copy number variations (CNVs) using Bayesian comparison of k-mer coverage between the two sequenced genomes.

We have applied RUFUS to both mutational profiling experiments using Toxopasma gondii as well as a variety of human disease family trios to detect novel germline mutations. In our mutational profiling work, RUFUS shows far higher specificity, correctly detecting every single confirmed point mutation detected my mapping based methods, while correctly excluding all false positive calls. RUFUS also detects, can confirmed, several novel point mutations that were not detected with the mapping based approaches, including variants in sequence omitted from the reference, showing improved sensitivity over mapping based methods. Furthermore RUFUS shows increased sensitivity to structural events, correctly identifying two plasmid insertion events that had eluded mapping based detection. Our work with human disease trios suggests a similar increase in specificity and sensitivity for de novo germline mutations as our mutational profiling work. Mapping based methods call thousands of de novo mutations in the child of a trio, far more than the suspected mutation rate of 10-8 suggested by earlier studies. These calls are dominated by mapping errors, which drown out any true signal. RUFUS calls fewer than 200 events per whole genome trio, with only 3 variants in the exome, far closer to the expected mutation rate seen by other studies. RUFUS has even detected a novel 386bp de novo inversion in one sample as well as numerous small insertions and deletions that mapping based methods struggle to call correctly. These results demonstrate the novelty of our approach and its potential for highly sensitive and specific variant identification of somatic variations in tumor genomes, as well as detecting de novo mutations in human trios.

PATIENT-SPECIFIC FACTORS INFLUENCE SOMATIC VARIATION PATTERNS IDENTIFIED BY WHOLE GENOME SEQUENCING OF INDEPENDENT TUMORS FROM VON HIPPEL-LINDAU DISEASE.

Suzanne S Fei[1], Asia D Mitchell[1], Cathy D Vocke[2], Christopher J Ricketts[2], Myron Peto[1], Nicholas J Wang[3], W Marston Linehan[2], Paul T Spellman[1]

[1]Oregon Health & Science University, Molecular & Medical Genetics, Portland, OR, [2]National Cancer Institute, Center for Cancer Research, Bethesda, MD, [3]Oregon Health & Science University, Biomedical Engineering, Portland, OR

Cancer development is presumed to be an evolutionary process that is influenced by the patient's genetic background and environment. In laboratory animals, genetics and environment are variables that can largely be held constant. This obviously cannot be done in humans; however, one can compare independent tumors that have developed in the same patient, effectively constraining genetic and environmental variation for the purpose of studying their contributions. Patients with von Hippel-Lindau (VHL) disease develop multiple, clonally independent, clear cell renal carcinomas in their lifetimes. In this study, we performed whole genome sequencing on 40 tumors from 6 VHL disease patients. We confirmed that the tumors are clonally independent, having distinct somatic single-nucleotide aberrations. Although tumors from the same patient showed many differences, within-patient patterns are discernible. Single-nucleotide substitution type rates were significantly different between patients and showed biases in trinucleotide mutation context. We also observed striking biases in chromosome copy number aberrations. These results support the view that genetic background and environment substantively influence the types of mutations that occur.

INVESTIGATING THE INFLUENCE OF THE GENOMIC CONTEXT ON EXPRESSION AND EVOLUTION OF THE HUMAN miRNAs

Gustavo S Franca[1], Anamaria A Camargo[2], Maria D Vibranovski*[3], Pedro A Galante*[2]

[1]Universidade de Sao Paulo, Departmento de Bioquimica, Sao Paulo, Brazil, [2]Hospital Sirio-Libanes, Molecular Oncology Center - IEP, Sao Paulo, Brazil, [3]Universidade de Sao Paulo, Departamento de Genetica e Biologia Evolutiva, Sao Paulo, Brazil

MicroRNAs (miRNAs) are one of the best functionally characterized classes of non-coding genes. Although many animal miRNAs are deeply conserved, it has becoming clear that miRNA evolution is a quite dynamic process, characterized by high birth-and-death rates and lineage-specific expansions. These hairpin-shaped molecules can emerge in intergenic regions or within coding or other non-coding genes, namely host genes. Curiously, more than half of the human miRNAs are embedded within protein-coding genes in sense orientation, suggesting that this organization have been favorable during evolution. Here, we have assessed the genomic location of human miRNAs regarding their evolutionary origin. We found that the bulk of known human miRNAs originated in primates, with striking elevated birth rates within coding genes particularly after the rodent-primate split. We show that expression patterns of inter- and intragenic miRNAs are related to their age and genomic context. For instance, intergenic miRNAs emerged after the fish-bird split are more highly and broadly expressed when are close to coding genes, whereas young intragenic miRNAs are more broadly expressed than young intergenic ones. By analyzing sequence conservation, we did not find significant differences between inter- and intragenic miRNAs, except for young intragenic miRNAs, which exhibited higher conservation scores than intergenic ones. However, this signal is likely an effect of their surrounding intronic regions, which are slightly more conserved than intergenic counterparts. Thus, although the genomic context seems to not impose strong constraints on the sequence level, based on our partial results, we suggest that the genomic environment exerts relevant influence on the miRNA expression, depending on their evolutionary age. This influence might be both a co-regulation effect, most likely in the case of intragenic miRNAs, or a facilitated transcription due to open chromatin surrounding host or neighboring genes. We aim to provide additional details of these mechanisms and their possible functional consequences in future analyses.

HDACi-INDUCED DIFFERENTIATION OF MYELOGENOUS LEUKEMIA RESULTS IN TARGETED CHROMATIN ACCESSIBILITY CHANGES

Christopher L Frank[1,2], David S Hsu[2], Gregory E Crawford[2,3]

[1]Duke University, Department of Molecular Genetics and Microbiology, Durham, NC, [2]Duke University, Center for Genomic and Computational Biology, Durham, NC, [3]Duke University, Department of Pediatrics - Division of Medical Genetics, Durham, NC

Small molecule inhibitors of histone deacetylases (HDACi) serve as potent anticancer agents for particular malignancies. The exact mechanism by how HDACi work is unclear, but there is evidence these compounds either directly promote apoptosis or sensitize cancer cells by cell cycle arrest and differentiation, facilitating synergistic combinations with other compounds. For over 25 years it has been known the myelogenous leukemia cell line K562 can be differentiated in vitro by exposure to HDACi. Despite this, the precise genomic and epigenomic changes these cells must undergo to assume an erythrocytic state remain poorly defined. To investigate chromatin-based changes responsible for anti-proliferative differentiation, we treated K562 cells for 72 hours with sub-lethal concentrations of the HDACi sodium butyrate or SAHA, and assessed global chromatin accessibility and expression changes by DNase-seq and RNA-seq. Despite the potential for global hyperacetylation, we identified several thousand specific regulatory elements (<10% of total DHS sites) that become significantly more or less accessible following treatment. These regulatory elements are enriched for non-promoter regions of the genome, nearby to genes that change expression with HDACi exposure, and contain abundant motifs for key hematopoietic lineage-defining transcription factors. We verified by ChIP-seq that the pioneer factor PU.1 increases binding at opened hypersensitive sites and likely contributes to active enhancer formation at these sites. Luciferase assays demonstrate HDACi-opened sites containing PU.1 motifs are sufficient for the transcriptional response. Ongoing and future work will examine the necessity of factors like PU.1 for HDACi-induced differentiation of myelogenous leukemia and other cancer cell types.

NATURAL VARIATION IN GENE EXPRESSION AND THE IMPACT ON MUTANT PHENOTYPES

Andrew Fraser, Victoria Vu, Adrian Verster, Tungalag Chuluunbaatar, Mike Schertzberg

University of Toronto, The Donnelly Centre, Toronto, Canada

Many mutations cause genetic disorders. However, two people inheriting the same mutation often have different severity of symptoms and in part this is due to differences in genetic background, that is, to the specific combination of rare and common variants that comprise each individual genome. Predicting how differences in genetic background impact the severity of inherited disorders is critical for personalized medicine yet little is known about the general underlying mechanisms and there have been no systematic studies in any animal to address this. Here we will present the results of the first such systematic analysis of the effects of genetic background on mutant phenotypes in any animal.

We used RNAi to compare loss-of-function phenotypes for ~1400 genes in two natural isolates of *C. elegans* — ~20% of genes have different severity of phenotypes due to genetic background differences. This variation in perturbed phenotypes can be largely predicted from variation in gene expression; this is true whether the perturbation is RNAi, drug treatment, or genetic mutation. Furthermore, we find the same effect in both worms and mammalian cells, suggesting it is a general property of genetic networks. We thus suggest that differences in the severity of mutant phenotypes between individuals are largely the result of natural variation in gene expression. In this way, reading an individual's genome sequence can identify rare disease-causing variants and measuring that individual's gene expression profile can largely predict the severity of the disease symptoms that will result from these inherited mutations.

ASSESSING THE GENETIC IMPACT OF THE INDIAN OCEAN SLAVE TRADE: GENOMIC ANCIENT DNA DATA FROM TWO HISTORICAL CEMETERIES IN MAURITIUS

Rosa Fregel[1], Martin Sikora[1,2], Marcela Sandoval[2], Maria Avila[1], Meredith Carpenter[1], Christopher R Gignoux[1], G David Poznic[1], Krish Seetah[3], Diego Calaon[3], Sasa Caval[3], Carlos D Bustamante[1]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]University of Copenhagen, Centre for GeoGenetics, Copenhagen, Denmark, [3]Stanford University, Department of Anthropology, Stanford, CA

During the time of the slave trade, millions of people were forcibly displaced from Africa. Recent studies have attempted to shed light on the origin and identity of slaves, mostly centered on the trans-Atlantic trade. In contrast, the Indian Ocean slave trade has received considerably less attention, despite the fact that it involved a longer time span. A clear example of the extent of the colonial manpower trade across the Indian Ocean is Mauritius, which had been uninhabited prior to the Age of Exploration. Approximately 100,000 slaves reached Mauritius as part of the Indian Ocean slave trade. By 1800, slaves constituted three quarters of its population. Moreover, after abolition, Mauritius underwent a dramatic population shift when the British replaced African slaves with South Asian indentured workers, who quickly constituted two thirds of the total population. In this study, we used next-generation sequencing of ancient DNA to estimate the genome-wide ancestries of individuals sampled in two archaeological sites in Mauritius, the historical cemeteries of Le Morne (LM; n=4) and Bois Marchand (BM; n=4), which are thought to contain the remains of slaves and indentured workers, respectively.

Damage patterns and length distributions are consistent with ancient DNA molecules. Endogenous DNA accounted for 0.8–45.1% of the total. The increased level of degradation is congruent with the high temperature and humidity of the region, especially in BM. MtDNA genome coverage was almost complete for all samples (97–100%), and sequencing depth ranged from 3.7 to 42.3 fold. MtDNA lineages in LM (L0a, L2a1, L3d and L3f) have a clear sub-Saharan African origin, while results in BM (M49, N5a and U2c) indicate an important South Asian source. Although the genome coverage was low for all the samples (0.14–6.64%), in six of the eight samples, at least 1,500 SNPs intersected with a reference panel. Principal component analysis and admixture estimates indicate that slaves from LM came from Madagascar, as well as from mainland Africa. BM results indicate an important South Asian component. Although the replacement of slave manpower had already begun at that time, we observed a significant African contribution.

This is the first genomic project to successfully recover data from African individuals involved in the Indian Ocean slave trade. Our results elucidate the origin of the slaves and increase our understanding of how African and Malagasy people admixed with the new labor contingents to create the modern Creole population.

THE HUMAN INDUCED PLURIPOTENT STEM CELL INITIATIVE (HIPSCI): MULTI-OMIC CELLULAR GENETICS ON HUNDREDS OF IPS LINES

Daniel Gaffney[1], HipSci Consortium[1,2,3,4]

[1]Wellcome Trust Sanger Institute, Computational Genomics, Cambridge, United Kingdom, [2]King's College London, Centre for Stem Cells and Regenerative Medicine, London, United Kingdom, [3]University of Dundee, Centre for Gene Regulation & Expression, Dundee, United Kingdom, [4]University of Cambridge, Department of Haematology, Cambridge, United Kingdom, [5]University of Cambridge, Anne McLaren Laboratory for Regenerative Medicine, Cambridge, United Kingdom, [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom

Induced pluripotent stem cells (IPSCs) are a powerful model system for studying the cellular basis of human health and disease. The Human Induced Pluripotent Stem Cell Initiative (www.hipsci.org) is a collaborative project generating IPSC lines with genotype and multi-omic phenotype information from 500 healthy individuals and 500 individuals with rare disorders, which will be made available to the research community and industry. To date, we have generated and phenotyped 640 high quality IPS lines from 221 individuals enabling a comprehensive survey of the sources of variation and heterogeneity in IPSCs. Here, we describe the initial analysis of the genomes, epigenomes, transcriptomes and proteomes of IPS lines from healthy individuals generated by HipSci. Despite an extensive literature on genetic and epigenetic heterogeneity in IPSCs, the lines we have generated are remarkably genetically and epigenetically stable. Our initial results illustrate that genetic effects on multiple IPSC cellular phenotypes are detectable and often explain substantially greater variation than many other technical and biological factors, including cell culture conditions, donor gender, age and tissue of origin. We present the first methylation, histone modification, expression and protein QTL maps in IPSCs, revealing how the effects of genetic change manifest across multiple cellular layers, and how these associations overlap with QTLs from adult tissues and with results from disease association studies. Finally, we examine how genetic and non-genetic factors influence key IPS cell properties such as cell growth rates and differentiation ability. We anticipate that both the lines and data generated by HIPSCI will form an invaluable resource for the future study of human disease and development

NEGATIVE FEEDBACK BUFFERS EFFECTS OF REGULATORY VARIANTS

Julien Gagneur[1], Daniel M Bader[1], Stefan Wilkening[2], Gen Lin[2], Manu Tekkedil[2], Kim Dietrich[1], Lars Steinmetz[2]

[1]Gene Center, LMU, Biochemistry, Munich, Germany, [2]EMBL, Genome Biology, Heidelberg, Germany

Mechanisms conferring robustness against regulatory variants have been controversial. Previous studies suggested widespread buffering of RNA misexpression on protein levels during translation. We do not find evidence that translational buffering is common. Instead, we find extensive buffering at the level of RNA expression, exerted through negative feedback regulation acting in trans, which reduces the effect of regulatory variants on gene expression. Our approach is based on a novel experimental design in which allelic differential expression in a yeast hybrid strain is compared to allelic differential expression in a pool of its spores. Allelic differential expression in the hybrid is due to cis-regulatory differences only. Instead, in the pool of spores allelic differential expression is not only due to cis-regulatory differences but also due to local trans effects that include negative feedback. We found that buffering through such local trans regulation is widespread, typically compensating for about 15% of cis-regulatory effects on individual genes. Negative feedback is stronger not only for essential genes, indicating its functional relevance, but also for genes with low to middle levels of expression, for which tight regulation matters most. We suggest that negative feedback is one mechanism of Waddington's canalization, facilitating the accumulation of genetic variants that might give selective advantage in different environments.

A PANEL OF INDUCED PLURIPOTENT STEM CELLS FROM CHIMPANZEES: A RESOURCE FOR COMPARATIVE FUNCTIONAL GENOMICS

Irene Gallego Romero*[1], Bryan J Pavlovic*[1], Irene Hernando-Herraez[2], Tomas Marques-Bonet[2], Louise C Laurent[3], Jeanne F Loring[4], Yoav Gilad[1]

[1]University of Chicago, Department of Human Genetics, Chicago, IL, [2]UPF-CISC, Institute for Evolutionary Biology, Barcelona, Spain, [3]University of California San Diego, Department of Reproductive Medicine, San Diego, CA, [4]The Scripps Research Institute, Center for Regenerative Medicine, San Diego, CA

Comparative genomics studies in primates are extremely restricted because we only have access to a few types of cell lines from non-human apes and to a limited collection of frozen tissues. In order to gain better insight into regulatory processes that underlie variation in complex phenotypes, we must have access to faithful model systems for a wide range of tissues and cell types. To facilitate this, we have generated a panel of 7 fully characterized chimpanzee (*Pan troglodytes*) induced pluripotent stem cell (iPSC) lines derived from fibroblasts of healthy donors. All lines appear free of integration from exogenous reprogramming vectors, can be maintained using standard iPSC culture techniques, and have proliferative and differentiation potential similar to human and mouse lines. To begin demonstrating the utility of comparative iPSC panels, we collected RNA sequencing data and DNA methylation profiles from the chimpanzee iPSCs and the corresponding fibroblast source cell lines, as well as from 7 human iPSCs and their source cell lines, which encompassed multiple cell types and population origins. Overall, we observe much less within-species regulatory variation in the iPSCs than in the somatic cells, indicating that the reprogramming process erases many of the differences among different somatic cell types, including some due to inter-individual variation. We identify 4,609 differentially expressed genes and 3,529 differentially methylated regions between the iPSCs of the two species, many of which are novel inter-species differences not observed between the original somatic cells. Amongst these is the well-known pluripotency-associated transcription factor REX1, which appears dispensable for pluripotency in chimpanzee iPSCs. Our results indicate that this panel of chimpanzee iPSCs will help realise the potential of iPSCs in primate studies, and in combination with genomic technologies, transform studies of comparative evolution.

* Authors contributed equally

SEXUAL DIMORPHISM IN GENE CO-EXPRESSION NETWORKS

Chuan Gao[1], Shiwen Zhao[2], Ian C Mcdowell[2], Christopher D Brown[3], Barbara Engelhardt[4]

[1]Duke University, Department of Statistical Science, Durham, NC, [2]Duke University, Department of Computational Biology and Bioinformatics, Durham, NC, [3]University of Pennsylvania, Department of Genetics, Philadelphia, PA, [4]Princeton University, Department of Computer Science, Princeton, NJ

Many complex phenotypes and common diseases in humans exhibit substantial sexual dimorphism. While sexual dimorphism remains grossly understudied, evidence suggests that much of this dimorphism can be traced to sex-specific transcriptional and post-transcriptional regulation. In this work, we develop statistical models that leverage data available from the Genotype-Tissue Expression (GTEx) project to characterize the sex-specificity of transcriptional regulatory networks and distal genetic regulation of gene transcription. Using these models, we simultaneously recover gene co-variation that is shared across sexes, differential across sexes, and unique to each of the two sexes. Technical and biological confounders are naturally modeled in this framework. In the GTEx project pilot data, we are able to recover sex-specific co-expression networks, and we are able to use these networks to identify sexually dimorphic expression quantitative trait loci (eQTLs). We replicate a subset of the co-expression network edges and eQTLs in other publicly available eQTL study data sets. When these results are analyzed in conjunction with tissue-specific networks from the same GTEx study data, we find suggestive tissue specific and tissue-ubiquitous explanatory mechanisms of sexually dimorphic complex traits. Furthermore, an analysis of the sex-specific co-expression networks reveals interactions that may be useful in explaining specific sexually dimorphic human disease traits, including autoimmune disorders and cardiovascular health.

THE MOBILE ELEMENT LOCATOR TOOL (MELT)

Eugene J Gardner[1,2], Nelson T Chuang[1,2], Vincent Lam[1,3], Ashiq Masood[1,3], 1000 Genomes Project Consortium[1], Ryan E Mills[4], Scott E Devine[1,2,3]

[1]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [2]University of Maryland School of Medicine, Molecular Medicine Graduate Program, Baltimore, MD, [3]University of Maryland School of Medicine, Greenebaum Cancer Center, Baltimore, MD, [4]University of Michigan Medical School, Computational Medicine and Bioinformatics, Ann Arbor, MI

Mobile Element Insertions (MEIs) are frequently found in human genomes, with each diploid genome harboring approximately1,200 non-reference Alu, L1, and SVA MEI insertions. This is equivalent to approximately 700 kbps of uncharacterized structural variation per human genome. We developed the Mobile Element Locator Tool (MELT) in collaboration with the 1000 Genomes Project (1KGP) to discover MEIs using Illumina whole genome sequencing data. We identified 16,631 (12,748 Alu, 3,048 L1, and 835 SVA) non-reference MEIs from the 26 diverse phase III 1KGP human populations. Built to be highly scalable, portable, and simple to deploy, MELT can be used to analyze a single genome or thousands of genomes. It can be used on low coverage and high coverage genomes and on libraries of varying trace lengths. MELT detects a wide range of features associated with MEIs, including: MEI insertion site, MEI type (Alu, L1, SVA), MEI length and orientation, target site duplication sequence, insertion site deletion, 3' transduction, and twin priming. MELT also reports the genotype for each sample. Internal sequences are assembled and used to determine the MEI subfamily status and to identify internal mutations, if present. We are currently conducting a combined PCR and sequencing based validation study to measure the accuracy of all of these predictions.

HYPOTHESIS-FREE DETECTION OF GENETIC NOVELTY ARISING FROM *DE NOVO* MUTATIONS AND RECOMBINATION REVEALS THE STRUCTURAL PLASTICITY OF THE MALARIA GENOME

Kiran V Garimella[1], Susana Campino[2], Samuel Oyola[2], Mihir Kekre[2], Eleanor Drury[2], Michael Krause[1,4], Zamin Iqbal[1], Alistair Miles[1,3], Rick Fairhurst[4], Dominic Kwiatkowski[1,2,3], Gil McVean[1,3]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [2]Pathogen Genetics Programme, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [3]MRC Centre for Genomics and Global Health, University of Oxford, Oxford, United Kingdom, [4]National Institute of Allergy and Infectious Disease, Laboratory of Malaria and Vector Research, Bethesda, MD

In the malaria pathogen, *Plasmodium falciparum*, point substitutions, structural mutations and non-allelic recombination are all known to be important in generating diversity at key loci involved in diverse phenotypes including virulence, drug resistance and antigenic variation. However, to date we know little about the general rate and profile of mutation and recombination in sexual and asexual stages of the pathogen life-cycle, nor how these relate to genomic features. Moreover, the low complexity and high structural diversity of its genome make studying such processes by direct observation of events within pedigrees challenging due to poor read alignment and missing sequence within the reference.

To address this challenge, and to characterize the spectrum of events that generate novelty within *P. falciparum*, we have developed a new 'hypothesis-free' approach to detecting genetic novelty introduced through mutation and recombination events. Our method is based on *de novo* genome assembly of high-coverage data from the parents and offspring and the detection of novel sequence within the offspring. We validate the method using simulations and comparison to independent data and calculate its power to detect variants generated by a range of mechanisms. We then apply the method to high coverage sequencing data (70-120X) from ~120 isolates obtained from four crosses involving lab-adapted strains. We identify a wide range of cross-overs, gene-conversions and *de novo* mutations, from SNPs and short indels to complex non-allelic recombination events. Importantly, we find that point mutations represent a small minority of the total events or novel sequence generated, suggesting that the unusual landscape of diversity in the malarial genome reflects, at least in part, underlying mutational pressures.

BRAINSPAN ATLAS OF THE DEVELOPING AND ADULT HUMAN BRAIN TRANSCRIPTOME

Mark Gerstein[1,2], Yuka I Kawasawa[3], Robert R Kitchen[2], Nenad Sestan[3], on behalf of the Brainspan Consortium[1,2,3]

[1]Yale University, Computational Biology and Bioinformatics, New Haven, CT, [2]Yale University, Molecular Biophysics & Biochemistry, New Haven, CT, [3]Yale University, Department of Neurobiology and Kavli Institute for Neuroscience, New Haven, CT

Detailed knowledge of transcriptional and post-transcriptional events in the human brain is essential to improve our understanding of development, function and disease. Here we present the BrainSpan Project, an ongoing large-scale effort to generate a multimodal atlas of the human brain, including an integrated survey of transcriptome, epigenome and genome across development and adulthood complemented by cellular gene expression mapping and anatomical reference atlases. To illustrate the utility of this resource, we provide a detailed catalog, generated through deep sequencing of RNAs from sixteen regions of adult male and female brains, of transcribed elements including known and novel protein-coding and non-coding transcripts, exons, transcriptionally active regions, and splicing and RNA editing events. Importantly, we show that transcriptional and post-transcriptional features exhibit extensive spatial variation. Furthermore, we find evidence for co-expression of coding and non-coding RNAs in region- and cell type- associated patterns, with strong anti-correlations between groups of mRNAs and microRNAs implying negative regulation of region-specific transcription. This freely accessible and mineable resource provides valuable opportunities for integrated investigations of the human brain transcriptome and facilitates translational research into neurological and psychiatric disorders.

SPATIAL SINGLE-CELL TRANSCRIPTOMICS REVEALS GENE EXPRESSION REGULATION IN THE DEVELOPMENT OF ANGIOSPERM AND GYMNOSPERM LEAF PRIMORDIA

Stefania Giacomello[1], Barbara Terebieniec[2], Fredrik Salmén[1], Nicolas Delhomme[2], Nathaniel Street[2], Joakim Lundeberg[1]

[1]SciLifeLab, Gene Technology, Stockholm, Sweden, [2]Umeå Plant Science Centre, Plant Physiology, Umeå, Sweden

Single-cell sequencing has provided insights into the importance of considering cell level dynamics and variance of gene expression. However, current methods do not provide a spatial context. Such spatial information is a key aspect in several biological and medical experiments. We developed a method, termed Spatial Transcriptomics, which generates single-cell resolution expression profiles while maintaining the spatial context of those cells within tissue structures. We applied this method to generate single-cell gene expression atlases of *Populus tremula* and *Picea abies* leaf primordia.

The method consists of placing tissue micro-sections onto arrays coated with barcoded probes specific for capturing mRNA. Barcodes are used to spatially locate an mRNA within the tissue micro-section and are at a density high enough to provide single cell resolution. Captured mRNA is then reverse transcribed, cDNA is removed from the array and used to prepare *ad hoc* Illumina sequencing libraries.

We applied the method to assay spatial gene expression patterns of ~21.000 genes in order to provide new insight into the role of transcriptional control during the evolution of leaf development between the model angiosperm and gymnosperm species, both of which we have produced genome assemblies for. The availability of spatial, single-cell information represents an unparalleled resource for identifying development transcriptional patterns controlling leaf morphological development. We are utlizing the unparallelled data resource represented by these spatial expression atlases to undercover contrasting developmental expression patterns among genotypes within a species and between species to advance understanding of the evolution and control of natural variation in leaf shape and form.

THE ROLE OF H3K27 IN IFNg-MEDIATED GENE EXPRESSION

Yu Qiao[1], Eugenia G Giannopoulou[1,2], Celeste Fang[1], Lionel B Ivashkiv[1,3]

[1]Hospital for Special Surgery, Arthritis and Tissue Degeneration Program and David C. Rosensweig Center for Genomics Research, New York, NY, [2]New York City College of Technology, City University of New York, Biological Sciences Department, Brooklyn, NY, [3]Weill Cornell Medical College, Department of Medicine, New York, NY

Interferon gamma (IFNg) is a major cytokine responsible for lymphocyte activation in enhanced anti-microbial and anti-tumor responses. It is the major pro-inflammatory cytokine to drive monocyte and macrophage classical differentiation (pro-inflammatory phenotype), which is critical for intracellular pathogen clearance. How IFNg induces gene expression in macrophages has been well studied: IFNg activates JAK-STAT1 signaling pathway, and the transcriptional factor STAT1 binds and facilitates the transcription of genes with GAS (IFNg activation sequence) motif. Recently, we found that IFNg also remodels chromatin structure through the regulation of histone acetylation as another layer of gene transcriptional regulation. However, IFNg not only induces hundreds of genes, but also represses hundreds of others that promote alterative macrophage differentiation and function. The mechanism of IFNg-mediated gene repression remains elusive. The IFNg repressive effect is mainly thought to be indirect, and the general understanding has been limited to the balance and competition between IFNg-STAT1 and IL10-STAT3. In this study, we investigate epigenetic mechanisms in IFNg-mediated gene repression and focus on the role of Polycomb Repressive Complex 2 (PRC2)-mediated H3K27me3 in comparison with H3K27ac. We first discovered that H3K27 marks (ac and me3) are good indicators of gene activity state in human primary monocytes. H3K27me3 occupies silenced genomic regions and H3K27ac active regions, which correlates with the expression levels of the associated genes. Genes in bivalent domains, with both H3K27ac and H3K27me3 marks, have intermediate expression levels. Upon IFNg stimulation, the change of these two histone marks correlates well with gene expression: increase/decrease of the H3K27ac histone mark agrees with increased/decreased gene expression respectively, while an opposite pattern is observed for H3K27me3. We also found that unlike H3K27ac, which undergoes dramatic genome-wide shift during IFNg response, H3K27me3 mark is much less subject to modulation. In fact, genome-wide analysis shows regulation of this mark at only a few specific loci, indicating high levels of specificity of the PRC2-H3K27me3 mechanism in IFNg response. Further functional/pathway analysis revealed that the targets of H3K27me3 regulation are genes that play important roles in monocyte differentiation and function. These findings indicate that PRC2-mediated transcriptional regulation plays an important role in macrophage inflammatory response.

HUMAN-SPECIFIC GENE EVOLUTION AND DIVERSITY OF THE CHROMOSOME 16P11.2 AUTISM CNV

Giuliana Giannuzzi[1], Xander Nuttle[2], Michael H Duyzend[2], Peter H Sudmant[2], Osnat Penn[2], Giorgia Chiatante[3], Maika Malig[2], John Huddleston[2,4], Laura Denman[2], Lana Harshman[2], Jacqueline Chrast[1], Carl Baker[2], Archana Raja[2,4], Kelsi Penewit[2], Francesca Antonacci[3], Alexandre Reymond[1], Evan E Eichler[2,4]

[1]University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland, [2]University of Washington, Department of Genome Sciences, Seattle, WA, [3]University of Bari, Department of Biology, Bari, Italy, [4]University of Washington, Howard Hughes Medical Institute, Seattle, WA

Recurrent deletions and duplications at 16p11.2 are a major contributor to autism. They are associated with schizophrenia and extremes of body mass index and head circumference. These rearrangements occur via nonallelic homologous recombination (NAHR) between directly oriented segmental duplications at BP4 (breakpoint 4) and BP5, ~600 kbp apart.
Using whole genome sequencing data from 2,551 humans, 86 great apes, a Neanderthal, and a Denisovan, we observed extensive copy number variation at BP4 and BP5 in humans. Through Illumina and PacBio sequencing of large-insert clones from chimpanzee and orangutan, we generated complete sequence over the 16p11.2 locus and reconstructed its evolutionary history. Three inversions occurred in the human lineage after divergence from orangutan, affecting >1 Mbp of sequence and 45 genes, together with the addition of ~1 Mbp via segmental duplication. The latter includes a ~102 kbp segment containing *BOLA2* that duplicated ~183 kya, i.e. the time when *Homo sapiens* emerged as a species. Modern humans carry at least one additional copy of *BOLA2* (ranging from 3 to 10 diploid copies) in contrast to apes and archaic hominins, where the gene exists as two diploid copies. We fully sequenced four distinct human haplotypes at 16p11.2 and discovered approximately the same ~102 kbp segment tandemly duplicated in BP4 and BP5, likely leading to different predisposition to NAHR. We are currently assaying *BOLA2* copy number and refining breakpoints in >125 patients with a 16p11.2 BP4-BP5 deletion or duplication.
BOLA2 is ubiquitously expressed, present in all eukaryotes, and involved in the regulation of iron metabolism. Expression levels in human lymphoblastoid cell lines correlate with copy number ($r = 0.29$), with expression of genes on chromosomes 16p13 and 19p13, and with expression of genes encoding for mitochondrial and ribosomal proteins. These findings raise the exciting possibility that an evolutionary advantage linked to the emergence of duplicated genes in the last 200,000 years of human evolution, underlies the predisposition to recurrent rearrangements at 16p11.2 associated with autism.

CLAN GENOMICS: RARE VARIANTS IN COMPLEX DISEASE REVEALED FROM WHOLE EXOME SEQUENCING

Richard A Gibbs[1], Eric Boerwinkle[2], James R Lupski[1,3,4]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]University of Texas, Human Genetics Center, Houston, TX, [3]Baylor College of Medicine, Department of Human and Molecular Genetics, Houston, TX, [4]Baylor College of Medicine, Department of Pediatrics, Houston, TX

The Clan Genomics hypothesis recognizes that if it is rare alleles that have the most significant impact on disease risk and pathology, then alleles contained and arising in recent ancestors will have the most profound impact on disease (PMID: 21962505). Tests of the Clan Genomics Hypothesis are based upon the population frequency of alleles that are shown to contribute to disease phenotypes: The model is supported if combinations of rare alleles are shown to determine phenotypes related to common disorders, while a clear role for common alleles supports other models. We have obtained experimental evidence in support of the Clan Genomics hypothesis from two qualitatively different approaches. First, family-based whole exome sequencing (WES) in both our CLIA/CAP certified clinical sequencing laboratory and in the research arena, for discovery in Mendelian disease, has now aggregated data from more than 10,000 cases of pediatric disease. Second, we have sequenced more than 15,000 individuals by WES and WGS in deeply phenotyped cohorts, to identify rare alleles that contribute to endophenotypes, including elevated metabolic intermediates (PMID: 25575548) and drug sensitivities (PMID: 25390462) or resistance (PMID: 25587968). Each supports the Clan Genomics Hypothesis – the family studies reveal complex phenotypes arising from compounded mutations at multiple loci; contributions from rare CNVs; an under-appreciated burden of de novo mutations; and an excessive carrier burden that explains some syndromic cases without simple Mendelizing disease. The data from the large cohorts show clearly how rare variants can have a strong effect on intermediate phenotypes, that lead to complex overall disease patterns. In aggregate these studies further establish the role of rare variation as a contributing factor in complex phenotypes related to disease. In contrast with the otherwise striking paucity of direct evidence of functional effects of common variants, all experimental test of our hypothesis continue to lend support to the concepts formulated in the Clan Genomics Hypothesis.

DESIGN AND IMPLEMENTATION OF THE NEXT GENERATION OF GENOME-WIDE ASSOCIATION STUDIES WITH THE MULTI-ETHNIC GENOTYPING ARRAY

Christopher R Gignoux[1], Genevieve L Wojcik[1], Henry R Johnston[2], Christian Fuchsberger[3], Suyash Shringarpure[1], Alicia R Martin[1], Stephanie Rosse[4], Daniel Taliun[3], Ryan Welch[3], Carsten Rosenow[5], Hyun M Kang[3], Gonçalo Abecasis[3], Michael Boehnke[3], Zhaohui Qin[2], Christopher Carlson[4], Carlos D Bustamante[1], Kathleen C Barnes[6], Eimear E Kenny[7]

[1]Stanford University, Genetics, Stanford, CA, [2]Rollins School of Public Health, Emory University, Biostatistics and Bioinformatics, Atlanta, GA, [3]University of Michigan School of Public Health, Biostatistics, Ann Arbor, MI, [4]Fred Hutchinson Cancer Research Center, Public Health Sciences, Seattle, WA, [5]Illumina, Inc, San Diego, CA, [6]Johns Hopkins University, Medicine, Baltimore, MD, [7]Icahn School of Medicine at Mount Sinai, Genetics and Genomic Sciences, New York, NY

In the past decade, genome-wide association studies (GWAS) have been extraordinarily successful. However much of this work has only been performed in populations of European descent. To address this disparity, a collaboration between Illumina, PAGE (Population Architecture Using Genomics and Epidemiology), CAAPA (Consortium on Asthma among African-ancestry Populations in the Americas), and T2D Genes Consortium developed the 1.7M SNP Multi-Ethnic Genotyping Array (MEGA). This array is designed to interrogate diverse variation from across the frequency spectrum and screen for relevant prior genetic discoveries.

The GWAS backbone is informed by whole genome sequences from the 1000 Genomes Project and CAAPA, with additional compatibility from the Illumina HumanCore array. We developed a novel cross-population tag SNP strategy to maximize imputation accuracy across six continental populations, with improved performance from previous generations of arrays. We chose rare, functional candidates from >36,000 multi-ethnic exomes. We also curated variants with domain experts, including boosted coverage in regions of interest (e.g. the MHC), over 5,000 ancestry informative markers, uniparental markers, and over 25,000 variants of clinical, prior GWAS, pharmacogenetic, and eQTL importance.

This array is being genotyped on a large number of diverse studies beginning with >60,000 individuals from PAGE and CAAPA. To aid global research we are also genotyping several thousand individuals, including HGDP and a large panel of Native Americans, to aid in rare variant calling, ancestry characterization, and admixture analyses. We are also evaluating imputation strategies for MEGA given the steadily increasing pool of available genomes. We intend MEGA to be a platform and an analytical resource for researchers interested in large-scale studies of diverse populations.

EVOLUTIONARY HISTORY AND SELECTIVE PRESSURES ACTING ON HUMAN POLYMORPHIC INVERSIONS

Carla Giner-Delgado[1,2], David Castellano[1], Magdalena Gayà-Vidal[1], Sergi Villatoro[1], David Izquierdo[1], Isaac Noguera[1], Marta Puig[1], Mario Cáceres[1,3]

[1]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain, [2]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain, [3]Institució Catalana de Recerca i Estudis Avançats, (ICREA), Barcelona, Spain

Chromosomal inversions have been a paradigm for evolutionary biology for decades. An important effect of inversions is that they suppress recombination in heterozygotes. This suppression may mediate processes such as local adaptation, evolution of sex chromosomes, and speciation. However, very little is known about the evolutionary dynamics of these mutations, especially in humans. Here, we took advantage of the large-scale genotyping effort of the INVFEST project, where 45 inversions have been interrogated in 550 individuals from seven populations of the 1000 Genomes Project, to investigate the leading evolutionary forces acting on them and point out candidates to have a functional role. By analyzing the nucleotide variation patterns within the inverted region we have determined that inversions mediated by inverted repeats (IRs) tend to be present in multiple and distant haplotypes, revealing a high degree of recurrence, while those without IRs seem to have a unique origin. Next, we used different strategies to detect selection pressures acting on this type of structural variants. First, we compared the observed site frequency spectrum and geographical distributions of the inversions against simulated neutral mutations taking into account the ascertainment bias in our calling panel and the human demography. Interestingly, this showed that inversions mediated by IRs are significantly more frequent and ubiquitously distributed than expected, in agreement with the recurrence hypothesis. In contrast, inversions without IRs follow the expected frequency and global distribution. Second, we estimated the age of unique inversions from nucleotide variation levels, which allowed us to narrow the expected frequency distribution and detect deviations from neutrality. Overall, these analyses detected different selective forces acting on specific inversions, including purifying selection, positive selection, and balancing selection. Our study therefore contributes to the understanding of this type of structural variants, and reports inversions that deserve further molecular and phenotypic characterization.

UTILIZING GENE EXPRESSION TO UNCOVER GENOTYPE-DEPENDENT EFFECTS OF BMI IN MULTIPLE TISSUES.

C A Glastonbury[1], A Viñuela[1,2], A Buil[2], P C Tsai[1], R Durbin[3], E Dermitzakis[2], T Spector[1], K Small[1]

[1]King's College London, Department of Twin Research and Genetic Epidemiology, London, United Kingdom, [2]University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva, Switzerland, [3]Wellcome Trust Genome Campus, Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Most complex traits are the result of environmental and genetic factors that may interact in complex ways. In order to determine if BMI-associated changes in gene expression are genotype-dependent we sought to identify gene x environment interactions (GxE) on the regulation of gene expression using BMI as the environmental exposure (GxE_{BMI}) and gene expression as the outcome. We performed a genome-wide cis scan (±1MB from TSS, MAF >5%) for GxE_{BMI} regulatory effects in RNA-sequencing data from three primary tissues and one cell line (Adipose, Skin, Whole Blood & LCLs) obtained from 856 healthy female twins from the TwinsUK cohort. At an FDR of 5% we identified four GxE_{BMI} regulatory effects in Adipose and one in Skin (Adipose: HACL-rs1464171, $p=4.77 \times 10^{-9}$, ALG9-rs3851570, $p=2.35 \times 10^{-8}$, GAA-rs34041461, $p=4.84 \times 10^{-7}$, SMG6-rs79963031, $p=6.64 \times 10^{-7}$; Skin: RHPN2-rs71351149, $p=4.52 \times 10^{-8}$). As trans-eQTLs are enriched for cis effects, we assessed whether any of the five GxE_{BMI} cis variants also show GxE_{BMI} effects in trans (±5MB from TSS). rs3851570, associated in cis to ALG9, was found to interact with BMI to regulate adipose expression of 257 genes in trans at $p < 1 \times 10^{-6}$ (Top hit: ZNF423 $p=8.23 \times 10^{-13}$). The genes associated in trans to rs3851570 were significantly enriched for oxidative phosphorylation (B-H corrected $p = 7.6 \times 10^{-6}$) and respiratory chain processes (B-H, p=0.0013) indicating rs3851570 interacts with BMI to regulate cellular metabolism. Further analysis of adipose cis-acting interactions at a relaxed threshold of $p < 1 \times 10^{-6}$ shows they are highly tissue specific and the lead SNPs are not directly associated to either BMI ($p > 0.05$; GIANT GWAS dataset, N= ~250,000) or gene expression (TwinsUK N=856). The genes influenced by these GxE_{BMI} regulatory effects are enriched for metabolic processes, but are not enriched for inflammatory processes associated with increased BMI in adipose tissue. Given this enrichment, we tested the identified SNPs for GxE_{BMI} interactions on metabolite levels, measured in an independent sample of twins (N = 4,300). We found one bonferroni significant interaction influencing levels of the metabolite Cholesterol (PI4KB-rs3002292, $p=7.59 \times 10^{-7}$). These findings suggest intermediate traits can act as powerful measures to detect non-additive genetic effects and demonstrate the replication of GxE effects across multiple layers of biology.

GENOME AND DEVELOPMENT EVOLUTION IN AMOEBOZOA

Gernot Glöckner[1], Thomas Winckler[2], Falk Hillmann[3], Angelika A Noegel[1], Pauline Schaap[4]

[1]University of Cologne, Biochemistry I, Medical Faculty, Cologne, Germany, [2]Friedrich Schiller University Jena, Pharmaceutical Biology, Jena, Germany, [3]Hans Knöll Institute Jena, Molecular and Applied Microbiology, Jena, Germany, [4]University of Dundee, Cell and Developmental Biology, Dundee, United Kingdom

Background
Amoebozoa are one of the 6 major branches of eukaryote evolution. They are more closely related to animals and fungi than to other eukaryote groups. Some species within the Amoebozoa, the Dictyosteliida, have a complex life cycle with a multicellular life stage. *D. discoideum* is a valuable model system within this group, which we use to address questions of development, social interaction, and phenotypic variation. We have analysed the genomes of several species of this group covering more than 600 million years of evolution to have a basis for further studies.
Results
Global transcriptomic changes of different members of the social amoebae were monitored throughout the life cycle. This revealed that around half of the genes in each organism are differentially expressed compared to the vegetative state. Of these, however, only a few hundred are conserved throughout the social amoebae as orthologs with similar expression during the life cycle. These conserved genes are enriched in functions related to signal transduction, development, and cellulose metabolism and likely constitute the building blocks for the establishment of the social life cycle of the Dictyosteliida. We further investigated, which of these genes were present in the most closely related species without a social life. To this end we sequenced the genome of *Protostelium mycophaga*, a species able to form single celled fruiting bodies and started to compile transcriptional profiles for this species.
Conclusion
Social amoebae represent a simplified model system to examine the evolution of complex multicellular organisms. Our analysis shows, which genes and gene classes are required to achieve a basic multicellular life stage. We were thus able to identify major building blocks of the life cycle and trace the occurrence of some of them to a simpler, unicellular organism. Obviously, additional, species specific functions modulate the life cycle, but non-conserved differential expression might be largely due to transcriptional hitchhiking effects.

DISTINCT CLASSES OF ENDOGENOUS RETROVIRAL ELEMENTS MARK THE CELL POPULATIONS IN HUMAN PREIMPLANTATION EMBRYOS

Jonathan Goeke[1], Xiinyi Lu[2], Yun Shen Chan[2], Huck-Hui Ng[2], Lam-Ha Ly[2], Friedrich Sachs[2], Iwona Szczerbinska[2]

[1]Genome Institute of Singapore, Computational and Systems Biology, Singapore, Singapore, [2]Genome Institute of Singapore, Stem Cell and Developmental Biology, Singapore, Singapore

About half of the human genome consists of highly repetitive elements, most of which are considered dispensable for human life. Here, we report that repetitive elements originating from endogenous retroviruses (ERVs) are systematically transcribed during human early embryogenesis in a stage-specific manner. Our analysis highlights that the long terminal repeats (LTRs) of ERVs provide the template for stage-specific transcription initiation, thereby generating hundreds of co-expressed, ERV-derived RNAs. Conversion of human embryonic stem cells (hESCs) to an epiblast-like state activates blastocyst-specific ERV elements, indicating that their activity dynamically reacts to changes in regulatory networks. In addition to initiating stage-specific transcription, many ERV families contain preserved splice sites that join the ERV segment with non-ERV exons in their genomic vicinity. In summary, we find that ERV expression is a hallmark of cellular identity and cell potency that characterizes the cell populations in early human embryos.

THE EVOLUTION AND FUNCTIONAL IMPACT OF HUMAN STUCTURAL VARIANTS SHARED WITH ARCHAIC HOMININ GENOMES

Yen-Lung Lin[1], Pavlos Pavlidis[2], Emre Karakoc[3], Jerry Ajay[4], Omer Gokcumen[1]

[1]University at Buffalo, Biological Sciences, Buffalo, NY, [2]Foundation of Research and Technology – Hellas, Institute of Molecular Biology and Biotechnology (IMBB), Heraklion, Greece, [3]Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plon, Germany, [4]University at Buffalo, Computer Science, Buffalo, NY

Allele sharing between modern and archaic hominin genomes has been variously interpreted to have originated from ancestral genetic structure or through non-African introgression from archaic hominins. However, evolution of variable human deletions that are shared with archaic hominin genomes have yet to be studied.

We identified 427 variable human deletions that are shared with archaic hominin genomes, ~87% of which originated before the Human-Neandertal divergence (ancient) and only ~9% of which have been introgressed from Neandertals (introgressed). Recurrence, incomplete lineage sorting between human and chimp lineages, and hominid-specific insertions constitute the remaining ~4% of allele sharing between humans and archaic hominins.

We observed that ancient deletions correspond to more than 13% of all common (>5% allele frequency) deletion variation among modern humans. Our analyses indicate that the genomic landscapes of both ancient and introgressed deletion variants were primarily shaped by purifying selection, eliminating large and exonic variants.

We found 17 exonic deletions that are shared with archaic hominin genomes, including those leading to 3 fusion transcripts. The affected genes are involved in metabolism of external and internal compounds, growth and sperm formation, as well as susceptibility to psoriasis and Crohn's disease. Our analyses suggest that these exonic deletion variants have evolved through different adaptive forces, including balancing and population specific positive selection.

Our findings reveal that genomic structural variants that are shared between humans and archaic hominin genomes are common among modern humans and can influence biomedically and evolutionarily important phenotypes.

A SIMPLE AND POWERFUL NEW APPROACH FOR GENERATING AND IMPROVING GENOME ASSEMBLIES

Nicholas H Putnam[1], Jonathan C Stites[1], Brendan L O'Connell[1,2], Brandon J Rice[1], Jarrod A Chapman[3], Charles W Sugnet[1], Tomas Marques-Bonet[4], Wesley C Warren[5], Andrew Fields[1], Paul D Hartley[1], David Haussler[2], Daniel S Rokhsar[3], Richard E Green[1,2]

[1]Dovetail Genomics, LLC, Santa Cruz, CA, [2]University of California, Santa Cruz, Biomolecular Engineering, Santa Cruz, CA, [3]DOE Joint Genome Institute, Walnut Creek, CA, [4]Universitat Pompeu Fabra/Consejo Superior de Investigaciones Cientificas, Institut de Biologia Evolutiva, Barcelona, Spain, [5]Washington University School of Medicine, The Genome Institute, St. Louis, MO

Since the onset of high-throughput sequencing, the field has struggled to devise fast and reliable ways to use these data for high-quality genome assembly. We present a new *in vitro* method for generating long-range connectivity information. This method, which involves *in vitro* chromatin assembly to condense naked DNA, requires only a few micrograms of material, no expensive equipment or exotic reagents, and can be done in a few days.

Using this long-range connectivity information, we present assembly results from several vertebrate genomes including human (NA12878), the American alligator, the chimpanzee, and the Prairie chicken. In each case, we are able to generate highly accurate assemblies with scaffold N50s of greater than 10 Mb from a simple combination of short-insert shotgun sequence data and one Illumina HiSeq lane of our new library. This two-library approach enables high-quality genome assembly in under six weeks from less than 10 ug of total DNA.

In addition to the 100-fold increase in genome contiguity, these data can also be used for haplotype phasing and structural variant detection. We are currently deploying this approach for *de novo* assembly and assembly improvement for a wide range of plant and animal genomes, including human disease genomes.

IDENTIFYING SIGNATURES OF PATERNAL TRANSGENERATIONAL GENETIC EFFECTS ON MOUSE TRANSCRIPTOMES

Rodrigo Gularte Mérida[1], Audrey Tromme[2], Fabien Ectors[1,2], Benoit Hennuy[3], Wouter Coppieters[3], Carole Charlier[1,4], Michel Georges[1,4]

[1]Unit of Animal Genomics, GIGA -- Research, Liège, Belgium, [2]Mouse Transgenics, GIGA -- Technology Platforms, Liège, Belgium, [3]Genomics, GIGA -- Technology Platform, Liège, Belgium, [4]Faculty of Veterinary Medicine, Université de Liège, Liège, Belgium

Evidence from C. elegans and D. melanogaster support the existence of paternal transgenerational genetic effects, i.e. the effect of untransmitted paternal alleles on the offsprings' phenotype. To test whether such effects might also occur in mammals, we performed two sequential backcrosses (F1, N2) of chromosome substitution males harbouring an A/J chrormosome (MMU15, 17, 19 or X) in an otherwise C57BL/6J (B6) background, to B6 females. We selected 684 out of 2,077 N2 offspring having a uniform B6 genotype using 236 SNP markers spanning the four chromosomes of interest. We extracted RNA from five tissue types (pituitary, heart, and liver from 60 day-old males; embryo and placenta from 13 dpc male embryos). For each tissue type, we generated two pools with RNA from eight individuals for the four BC2 populations, as well as for matched pure-bred B6 animals. We performed RNA-Seq on all samples and searched for strain-specific effects on gene expression by comparing each strain to B6 using the DESeq R/Bioconductor package. We identified 53 instances of differential expression with false discovery rate (FDR) ≤ 0.05. 24 of these were detected in B6.A-15 B6/B6, seven in B6.A-17 B6/B6, 14 in B6.A-19 B6/B6 and eight in B6.A-X B6/B6. Nine instances were detected in the embryo, three in the placenta, eight in heart, 17 in liver, and 16 in pituitary. The tissue that was predominantly affected differed significantly between lines. The 53 instances of differential expression involved 47 distinct genes. Mid1 was differentially expressed in four tissues (embryo, placenta, heart and pituitary) of one line (19-B6/B6), Serpina3k was differentially expressed in one tissue (pituitary) of three lines (15-B6/B6, 19-B6/B6 and X-B6/B6), and Coq10b was differentially expressed in one tissue (pituitary) of two lines (15-B6/B6 and X-B6/B6). The 50 other genes were differentially expressed in one tissue and one line. Differentially expressed genes mapped uniformly across the genome. There was no evidence for preferential mapping of differentially expressed genes to the chromosome tested for TGE (15, 17, 19 and X). The paternal TGE effects ranged from a 20.1-fold up-regulation to a 12.8-fold down-regulation of the target genes. Thirty-five genes were up-regulated, while 18 genes were down-regulated. Validation using independent methodology and samples is in progress. Latest results will be presented.

ERROR CORRECTION AND DE NOVO ASSEMBLY OF OXFORD NANOPORE SEQUENCING

James Gurtowski[1], Sara Goodwin[2], Scott Ethe-Sayers[2], Panchu Deshpande[2], Michael C Schatz[1], W. Richard McCombie[2]

[1]Cold Spring Harbor Laboratory, Computational Biology, Cold Spring Harbor, NY, [2]Cold Spring Harbor Laboratory, Genetics, Cold Spring Harbor, NY

Through its ongoing MinION access program, Oxford Nanopore has introduced the first commercially available nanopore based sequencer. This thumbdrive size device has the potential to revolutionize the field of genomic sequencing with long reads (maximum length currently approaching 100kbp), high throughput (~10bp / second per pore) and relatively simple sample preparation. These devices and data characteristics promise exciting advances in a number of genomic assays including de novo genome assembly, structural variation discovery as well as transcript analysis.

Here we present our experiences working with Oxford Nanopore sequencing of E. coli K12 and the S. cerevisiae (yeast) strain W303. We discuss the salient features of the technology including the read length distribution, error model and overall throughput. We also discuss the tools and algorithms we use for downstream analysis and discovery; including our alignment method and parameters. Our focus, however, is on the current and future applications of this technology, especially for de novo genome assembly. Using a hybrid error correction approach called Nanocorr, we corrected errors in the Oxford Nanopore reads using high identity MiSeq reads collected from the same strains. The process improved the raw nanopore reads with a mean error rate of 35% to an average per base identity of 97%. The corrected reads were then assembled using the Celera Assembler which now supports reads up to 500kbp. Our E. coli assembly was nearly perfect, yielding a single contig matching the reference genome with a 99.99% per base accuracy. The yeast W303 assembly had an average identity of 99.85% after polishing and a N50 value of 585kb; nine times greater than the pure MiSeq assembly and resolving all but the most complex repeats. By comparing the nanopore assemblies to those produced with other technologies we have gained insight into the current and future uses of this technology.

THE INTERPLAY OF GENOMES AND EPIGENOMES IN HEMATOPOIETIC DEVELOPMENT AND CARDIOVASCULAR DISEASE

Nicole Soranzo[1,2]

[1]Wellcome Trust Sanger Institute, Human Genetics Department, Cambridge, United Kingdom, [2]University of Cambridge, Department of Haematology, Cambridge, United Kingdom

Blood cell formation is tightly regulated within physiologic ranges by a complex interplay of genetic and non-genetic factors. Our group uses large scale genomic explorations based on sequencing technologies to study the contribution of human genetic variation to hematopoietic development. I will describe ongoing efforts to discover genetic determinants of haematological variation through genome-wide association studies. I will present integrative analyses of genetic data, gene expression and epigenetic marks in three immune cell types for mapping the functional consequences of genetic variants. Finally, I will show how this information contributes to our understanding of genetic predisposition to cardiovascular disease risk.

POPULATION-SCALE AND SINGLE-CELL RNA SEQUENCING PROVIDE INSIGHT INTO THE PATTERN OF X CHROMOSOME INACTIVATION ACROSS HUMAN TISSUES

Taru Tukiainen[1,2], Alexandra-Chloe Villani[2,3], Andrew Kirby[1,2], David DeLuca[2], Rahul Satija[2,4], Andrea Byrnes[1,2], Julian Maller[1,2], Tuuli Lappalainen[4,5], The GTEx Project Consortium[2], Aviv Regev[2], Nir Hacohen[2,3], Kristin Ardlie[2], Daniel MacArthur[1,2]

[1]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [2]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA, [3]Massachusetts General Hospital, Center for Immunology and Inflammatory Diseases, Charlestown, MA, [4]New York Genome Center, New York, NY, [5]Columbia University, New York, NY

Incompleteness and skewing of X chromosome inactivation (XCI) can result in biases in disease susceptibility and presentation between sexes and across individuals, but the full extent and heterogeneity of XCI remains unclear. We have deployed several complementary approaches based on high-throughput RNA sequencing to comprehensively profile the landscape, regulation and variability of escape from XCI.

Using detailed gene expression data from the GTEx consortium, including more than 30 tissue types and over 350 individuals, we show that a large majority of previously reported escape genes demonstrate male/female expression differences detectable at population-level. For many of these genes sex-biased expression is present and directionally similar across the various tissues studied, a pattern distinct from autosomal sex-biased expression, suggesting XCI is tightly and uniformly regulated across human tissues. Notably, however, escape genes close to an edge of an escape domain (e.g. KAL1) show more tissue heterogeneity and subtle sex-bias.

To complement these observations and assess individual-level variability in escape we have analyzed single cell RNA-seq data across two tissue types, and assessed the allelic imbalance across the X chromosome from deep sequencing of 17 tissues from a female presenting with completely skewed XCI. These analyses highlight well-known escape genes (e.g. USP9X), replicate novel candidates from the population-scale analyses (e.g. ZRSR2) and confirm variable escape genes (e.g. TIMP1) and elaborate the underlying dynamics. While finding little evidence for tissue-specific escape the analyses demonstrate tissue heterogeneity in expression from the inactive X (e.g. GYG2).

Together these analyses provide a comprehensive view of the landscape of escape from XCI in adult tissues, essential for understanding the impact of this process on sex differences and inter-individual variability.

THE MITOCHONDRIAL RESPONSE TO STRESS

Na Cai[1], Simon Chang[2], Yihan Li[1], Warren Kretzschmar[1], Jingchu Hu[3], Jonathan Marchini[4], Richard Mott[1], Jun Wang[3], Kenneth Kendler[5], Jonathan Flint[1]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]College of Medicine, Chang Gung University, Department and Graduate Institute of Biomedical Sciences, Tao-Yuan, Taiwan, [3]Beijing Genomics Institute, Shenzhen, Shenzhen, China, [4]University of Oxford, Department of Statistics, Oxford, United Kingdom, [5]Virginia Commonwealth University, Department of Psychiatry MCV, Richmond, VA

Our genetic analysis on Major Depressive Disorder (MDD) using whole-genome sequence data from saliva samples of 10,640 Han Chinese women (5303 cases of Major Depression, 5337 controls) yielded three surprising findings. Firstly, a genome-wide association study (GWAS) found associations between SIRT1 and SLC25A37 genes and MDD. Secondly, we found an enrichment of rare deleterious variants in genes with mitochondrial functions in cases of MDD. Lastly, we saw significantly higher mitochondrial DNA (mtDNA) levels in cases than controls. Given these observations, we hypothesize the mitochondria is central to responses to stressful life events or pathological depressive states. Using mouse models of chronic stress we found a dose-dependent increase in mtDNA levels as period of stress increased, and it was accompanied by a decrease in oxygen consumption. Using mitochondrial sequences from both human and mouse whole-genome sequencing, we found cases of MDD and stressed mice alike showed increased levels of loss-of-function heteroplasmic mutations in genes encoding key components of the electron transport chain predicted to contribute the most to reactive oxygen species (ROS) production. We hence propose a metabolic switch mediated by preferential expansion of mtDNA containing specific sequence mutations, the sensitivity of which to environmental stress may confer susceptibility to MDD.

MECHANISTIC BASIS AND CAUSALITY ANALYSIS OF SINGLE-NUCLEOTIDE VARIANT UNDERLYING THE FTO OBESITY LOCUS REVEALS NEW PATHWAY FOR TISSUE-MITOCHONDRIAL THERMOGENESIS REGULATION IN ADIPOCYTES

Melina Claussnitzer[1,2,3,4], Simon Dankel Nitter[5], Gerald Quon[1,2], Kyoung-Han Kim[6], Gunnar Mellgren[5], Chi-Chung Hui[6], Hans Hauner[4], Manolis Kellis[1,2]

[1]MIT, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Broad Institute, of MIT and Harvard, Cambridge, MA, [3]Harvard Medical School, Boston, MA, [4]Technische Universität München, Munich, Germany, [5]University of Bergen, Department of Clinical Science, Bergen, Norway, [6]University of Toronto, Department of Molecular Genetics, Toronto, Canada

Illustrating both the power and the limitations of genome-wide association studies, the FTO locus harbors the strongest genetic association with obesity, but its mechanistic basis remains elusive. Here, we provide evidence that the rs1421085 T-to-C single-nucleotide variant underlies the observed association, by disrupting a conserved motif for the ARID5B repressor, resulting in activation of a potent super-enhancer in adipocyte precursor cells, over-expression of its targets IRX3 and IRX5, and a cell-autonomous shift from energy dissipation to lipid accumulation in risk allele carriers. We establish variant, regulator, and downstream target gene causality by bi-directional CRISPR/Cas9 genome editing of the rs1421085 single-nucleotide variant, coupled with knock-down and over-expression, using adipose-derived progenitor cells from homozygous risk and non-risk allele carriers. At the cellular level, risk allele carriers show decreased mitochondrial energy dissipation, increased lipid storage, and a shift from browning to whitening adipocyte gene expression programs. At the organismal level, adipose-specific Irx3 inhibition in mouse results in 60% reduced fat mass ratio, high-fat-diet resistance, lipid store reduc-tion, and increased energy expenditure, with unchanged activity or appetite. Single-nucleotide C-to-T genome editing of rs1421085 in patient samples rescues the phenotypic signatures of obesity, with white adipocyte browning, including increased mitochondrial activity, thermogenesis, and reduced lipid anabolism. By elucidating the regulatory circuitry of the FTO locus, our results implicate mitochondrial activity and thermogenesis in adipose lineages, and specifically ARID5B, rs1421085, IRX3, and IRX5, as potential tissue-autonomous therapeutic targets in obesity, which can have societal implications in the current obesity epidemic.

SPARSE WHOLE GENOME SEQUENCING IDENTIFIES SUSCEPTIBILITY LOCI FOR MAJOR DEPRESSIVE DISORDER IN HAN CHINESE WOMEN

Jonathan Flint, on behalf of the CONVERGE consortium

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

Major depressive disorder (MDD) is a leading cause of disability worldwide and poses a major challenge to genetic analysis. Using low coverage genome sequence of 5,303 Chinese women with recurrent MDD, selected to reduce phenotypic heterogeneity, and 5,337 controls screened to exclude MDD, we identified and replicated two genome-wide significant loci contributing to risk of MDD on chromosome 10: one near the SIRT1 gene the other in an intron of the LHPP gene. Analysis of a subset of 4,509 cases with melancholia yielded an increased genetic signal at SIRT1, but not LHPP. Exclusion of 668 cases reporting childhood sexual abuse, a strong environmental risk factor for MDD, led to increased odds ratios at both loci and detection of another genome wide significant locus near SLC25A37, a mitochondrial ion transporter on chromosome 8. Genome-wide analysis of rare exonic variants showed that cases have a small but significant enrichment of deleterious coding variants. This was due in part to mutations in genes involved in mitochondrial biology, consistent with the involvement of SIRT1 and SLC25A37 in mitochondrial function. These results demonstrate the complexity of genetic effects contributing to MDD, attributable to the disease's etiologic heterogeneity, and suggest that the pathogenesis of MDD includes a mitochondrial origin.

NEW INSIGHTS INTO SCHIZOPHRENIA RISK FROM A GENOME-WIDE STUDY OF CNV IN 41,321 SUBJECTS

Daniel P Howrigan[1], Christian R Marshall[2], Daniele Merico[2], Bhooma Thiruvahindrapuram[2], Wenting Wu[4], Michael C O'Donovan[3], Stephen Scherer[2], Benjamin M Neale[1], Jonathan Sebat[4]

[1]Massachusetts General Hospital, Molecular Biology, Boston, MA, [2]The Hospital for Sick Children, The Centre for Applied Genomics, Toronto, Canada, [3]Cardiff University, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff, United Kingdom, [4]University of San Diego, Department of Psychiatry, La Jolla, CA

Copy number variants (CNVs) across the genome have been implicated as risk factors for schizophrenia (SCZ). Identification of associated loci and pathways has been challenging due to the low frequencies of variants, requiring large-scale collaborative effort and meta-analytic methods across genotyping platforms.

We developed a centralized CNV calling pipeline and applied it to an international cohort comprised of 41,321 subjects (21,094 SCZ cases and 20,227 controls) from the Psychiatric Genomics Consortium study of SCZ. Following raw data processing, samples were filtered based on array QC metrics and a consensus CNV call set was generated from the intersection of multiple callers. CNVs were filtered within each dataset based on frequency (MAF < 1%), probe density, size, and overlap with segmental duplications or regions prone to VDJ recombination.

Genetic associations were investigated by case-control tests of CNV burden at four levels: (1) genome, (2) pathways, (3) genes, and (4) probes. Analyses controlled for SNP-derived principal components, genotyping platform, and individual-level probe intensity. Multiple-testing thresholds for genome-wide significance were estimated from false-discovery rates drawn from permutation (for SNPs and genes) or by the Benjamini Hochberg procedure (pathways).

Overall CNV burden is enriched among SCZ cases, even after excluding CNVs implicated in previous studies. Deletions were significantly enriched among gene sets related to synaptic function and nervous system development, with GO synaptic genes and activity-regulated cytoskeleton-associated protein (ARC) complex remaining significant after removing previously implicated SCZ loci.

CNVs at multiple loci surpass genome wide correction, including deletions at 15q13.3 and 22q11.2 and duplications at 16p11.2. Gene-based tests identified additional novel loci, including multiple loci on the X chromosome. While X chromosome CNVs had no prior reports of association to schizophrenia, we find evidence that Xq28 is a CNV “hotspot” harboring both protective and risk CNV. We show that CNV analysis in large GWAS datasets can advance our understanding of rare structural variation in SCZ, and through large-scale collaboration, allow for more accurate estimation of reported CNV risk factors.

WHOLE GENOME SEQUENCING OF DIVERSE HUMAN POPULATIONS RESOLVES CAUSAL REGULATORY VARIANTS

Marianne K DeGorter[1,2], Tracy Nance[1,2], Rachel Agoglia[2], Adam Auton[3], Stephen B Montgomery[1,2]

[1]Stanford University, Department of Pathology, Stanford, CA, [2]Stanford University, Department of Genetics, Stanford, CA, [3]Albert Einstein College of Medicine, Department of Genetics, Bronx, NY

A major challenge to identifying causal regulatory variation is distinguishing the causal variant from multiple tightly-linked alternatives. By leveraging differing patterns of linkage disequilibrium across populations, it is possible to localize the causal variant. Now, with the availability of whole genomes from multiple populations from Phase 3 of the 1000 Genomes Project, coupled with gene expression data in 520 individuals from seven populations representing five continental groups, we leveraged genetic variability to identify causal regulatory variants shared among human populations. We found that nearly half of all expression quantitative loci (eQTLs) discovered (FDR < 0.05) in each of the seven populations represent blocks of tied variants, some with up to several hundred variants in perfect linkage disequilibrium in our sample. By meta-analysis of gene expression in the six other populations, we can break ties and assign a single candidate causal variant in the majority of loci with tied variants. We further demonstrate that these candidate causal variants are more likely to overlap transcription factor binding sites and H3K27ac marks from ENCODE, compared to a SNP chosen naively from the haplotype block ($p < 0.001$). When considering the relative utility of combinations of populations, we detect that inclusion of African populations provides the largest improvement in our ability to detect functional variants due to increased genetic diversity in these populations. Applying this approach, we further refine and report the properties of causal variants underlying several GWA studies. Overall, we demonstrate that a multiple population approach will be an important and efficient design for follow-up investigations which aim to localize causal variants from eQTL and GWA studies.

DETECTING GENE-BY-ENVIRONMENT INTERACTIONS USING ALLELE SPECIFIC EXPRESSION

David A Knowles[1], Joe R Davis[2], Stephen B Montgomery*[2], Alexis Battle*[3]

[1]Stanford University, Radiology, Stanford, CA, [2]Stanford University, Pathology, Stanford, CA, [3]Johns Hopkins University, Computer Science, Baltimore, MD

The impact of environment and lifestyle on human health is dramatic, with major risk factors including substance use, pollution, diet and exercise. However, the interaction between environment and individual genetic background is poorly understood; detecting gene-by-environment interactions is both statistically and computationally challenging. Using RNA sequencing of primary tissue (whole blood) from 922 individuals in the DGN cohort, combined with extensive annotation of environmental factors including drug use and behavioral factors, we evaluate GxE effects at a cellular level. Even in this study, standard association methods detect only a handful of gene-by-environment interactions at a lenient FDR of 0.1 across 30 environment variables. RNA-seq offers an alternative strategy however: we look for genes whose allelic expression is associated with an environment variable of interest. Measuring two different alleles within a single individual and time point offers a highly controlled comparison of environmental modulation of genetic effects, robust to the many biological and technical variables that differ between samples, and thus provides additional power and specificity to identify causal factors. To test for such associations, we have developed a hierarchical Bayesian model of allele-specific read counts that takes into account local cis-regulation, read depth, and technical influences on overdispersion of RNA-seq data. This methodology significantly increases power by leveraging the controlled, within individual, nature of allelic expression and by integrating over potential causal cis-variants. Applying this method across 30 environment variables, we find dozens of GxE interactions at FDR 0.1, a substantial increase over the standard approach. Individually significant genes include the glycolysis enzyme PFKL for exercise, NPRL3, involved in the in homeostasis of body fluid volume, for blood pressure medication, and CDH23 for decongestant medication, which is expressed in the neurosensory epithelium. For many of the associations we discover using the allele specific signal it is possible to find candidate causal variants using total expression. In conclusion, we show it is possible to leverage the novel information provided by RNA-seq beyond total expression to unravel the influence of disease-associated environmental factors on gene regulation.

(*co-senior authors)

COMPUTATIONAL CHALLENGES IN SINGLE-CELL BIOLOGY AND APPLICATIONS IN MAMMALIAN DEVELOPMENT

John Marioni[1,2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom, [2]Wellcome Trust Sanger Institute, Cambridge, United Kingdom

Cell identity and function can be characterised at the molecular level by unique transcriptomic signatures. At the organismal level, different tissues possess distinct gene expression profiles and individual cells in early-stage embryos display highly divergent transcriptomic landscapes.

Until recently, molecular fingerprints were generated by profiling of gene expression levels from bulk populations of millions of input cells. These ensemble-based approaches meant that the expression value for each gene was an average of its expression across a population of input cells. However, there exist many biological questions where bulk measures of gene expression are insufficient. For instance, during early development there are only a small number of cells, each of which can have a distinct function and role. In these, and other settings, assaying gene expression at the single-cell level represents a powerful tool for biological discovery.

Critically, recent experimental advances, in particular single-cell RNA-sequencing (scRNA-seq), have greatly improved the high-throughput generation of cDNA libraries from the poly-adenylated fraction of mRNA molecules within a single cell. scRNA-seq can be applied to assay the individual transcriptomes of large numbers of cells isolated via microfluidics or other microwell-plate-based techniques. The combination of a large number of cells and high-throughput profiling of gene expression (and other omics measurements) at the single-cell level is crucial for answering many biologically relevant questions and provides an opportunity for new discoveries in important areas of biology. However, to fully exploit these opportunities it is critical that computational methods are developed in parallel to experimental developments.

In this presentation I will discuss computational strategies we have developed to model single-cell transcriptomics data and illustrate how they can be used to provide insight into a variety of questions relating to early mammalian development.

EXPLORATION OF GENETIC VARIATION AND GENOTYPES AMONG MILLIONS OF GENOMES

Ryan M Layer[1], Konrad J Karczewski[2], Exome Aggregation Consortium (ExAC) [3], Aaron R Quinlan[1]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [3]Exome Aggregation Consortium, Cambridge, MA

Genome studies of thousands of individuals are becoming the norm as successive projects consider larger cohorts in search of genetic contributions to disease. Between public research projects (e.g., UK100K), private ventures (e.g., 23andMe), and clinical efforts (e.g., drug trials), we will soon have sequenced millions of genomes and cataloged hundreds of millions of variants. The problem is that current tools do not scale to such datasets. The Variant Call Format (VCF), the standard for representing genetic data, is effectively a matrix where rows correspond to sites of genetic variation, columns to individuals, and cells to the genotype of one individual at one locus. VCF has been crucial for consistency, data sharing, and interoperability. The issue is that storing large cohorts in VCF requires substantial resources. Binary VCF (BCF) reduces file size via compression, but the inflation step adds computational overhead.

These challenges motivated Genome Query Tools (GQT), an open source tool and API for efficiently querying population-scale genotype datasets. Through efficient data compression, GQT creates an index that is a fraction of the size of the VCF (e.g., 100X smaller than the VCF for 1000 Genomes), and executes most queries in seconds by directly analyzing the compressed data (e.g., 50X faster than BCFTOOLS). Importantly, speedup only increases as datasets grow; which we will demonstrate with the Exome Aggregation Consortium data set with more than 60,000 human exomes. GQT also provides a rich query interface includes sample phenotypes, genotypes, and inheritance patterns into searches. The utility of GQT is not limited to exploratory queries. The data model is general enough to enable traditional operations, fast enough to complete in a fraction of the previous time required, and can run on your laptop. As an example, PCA of the 1000 Genomes Phase 3 data set (2405 individuals, 90 million variants) required 15 hours on one CPU. This is compared to two hours on 480 cores using Google's latest genomics API.

GQT could also play a role in the Global Alliance for Genomics and Health, whose sharing model shifts a significant amount of computational burden to data providers. Therefore, its success depends on the efficiency of the supporting infrastructure. GQT can support many operations while also reducing the storage and compute burden. With a simple interface and interactive-speed results to the largest data sets, GQT provides researchers and clinicians with a powerful tool to explore population-scale variation datasets.

https://github.com/ryanlayer/gqt

A DNA CODE GOVERNS CHROMATIN ACCESSIBILITY

Tatsunori Hashimoto*[1], Richard Sherwood*[2], Daniel Kang*[1], Amira Barkal[1,2], Haoyang Zeng[1], Bart Emons[2], Sharanya Srinivasan[1,2], Nisha Rajagopal[1], Tommi Jaakkola[1], David Gifford[1,3]

[1]MIT, CSAIL, Cambridge, MA, [2]Harvard Medical School, Dept. Genetics, Brigham and Women's Hospital, Boston, MA, [3]Harvard University and Medical School, Dept. Stem Cell and Regenerative Biology, Cambridge, MA

We find that genome wide chromatin accessibility in a range of human and mouse cell types can be accurately predicted solely from DNA sequence. The code reveals that local, non-specific cooperation, largely among pioneer TFs, is sufficient to predict chromatin accessibility.

To discover the chromatin accessibility code we developed a new machine learning method, the Cooperative Chromatin Model (CCM), that models a base-pair resolution genome accessibility signal with the combined effect of thousands of k-mers with invariant spatial effects. To test the accuracy of a CCM at predicting genomic chromatin accessibility, we trained a CCM on DNase-seq data from chromosomes 1-13 of human K562 cells. We then predicted DNase-seq data on a held-out chromosome (chromosome 14). The CCM predictions are remarkably similar to actual DNase-seq reads, producing a chromosome-wide Pearson's correlation value of 0.801 between predicted and actual reads with ~40% of the peaks being predicted. A CCM trained on DNase-seq of purified DNA stripped of proteins fails to predict held-out chromatin accessibility with Pearson's correlation of 0.469 showing that the CCM is not merely reading out DNase or sequencing bias. We also find we can also predict DNase-seq peaks with a model trained on ATAC-seq data and vice versa, suggesting that the code we discover is not mainly reproducing protocol artifacts.

Using Single Locus Oligonucleotide Transfer (SLOT), a novel high-throughput in vivo testing platform based on highly efficient site-specific CrispR-based homologous recombination, we show that a CCM is capable of predicting the chromatin accessibility of a wide range of synthetic DNA sequences. We designed a library of 12,000 175 bp DNA sequences to test the CCM's ability to predict chromatin accessibility of any DNA sequence in a controlled chromatin context. We performed SLOT to integrate our library of DNA sequences with diverse chromatin opening properties into a genomic locus that resides in natively inaccessible chromatin. Barcode sequencing of DNase-hypersensitive phrases reveals strong linear concordance with phrases predicted by the CCM to promote open chromatin.
Collectively, our results suggest that a cooperative model of cis-acting DNA sequence features explains the majority of cellular chromatin accessibility.

GLOBAL SHIFTS IN ISOFORM USAGE IN RESPONSE TO INFECTION SUGGEST CONCERTED REGULATION BY TRANSCRIPTIONAL AND RNA PROCESSING MECHANISMS

Athma A Pai[1], Yohann Nedelec[2], Golsheed Baharian[2], Jean-Christophe Grenier[2], Vania Yotova[2], Christopher B Burge[1,3], Luis B Barreiro[2,4]

[1]MIT, Biology, Cambridge, MA, [2]CHU Sainte-Justine Research Center, Genetics, Montreal, Canada, [3]MIT, Biological Engineering, Cambridge, MA, [4]University of Montreal, Pediatrics, Montreal, Canada

Changes in gene regulation have long been known to play an important role in both innate and adaptive immune responses in human cells, with many previous studies extensively characterizing transcriptional responses to infection. However, post-transcriptional mechanisms, especially those involved in mRNA processing and alternative splicing, are now gaining prominence as crucial regulators of immune defenses. We sought to investigate the role of mRNA processing in the cellular responses of human macrophages to live bacterial infections. To do so, we infected primary macrophages of 60 individuals with Listeria monocytogenes and Salmonella typhimeurium for 4 hours and conducted mRNA-sequencing on matched non-infected, Listeria- and Salmonella-infected samples. Using these data, we quantified gene expression levels and isoform abundances in response to bacterial infection. Consistently with previous reports, we observe extensive changes in expression profiles between non-infected and infected samples. Notably, we also see an widespread increase in isoform diversity, concomitant with many significant changes in individual mRNA processing events, most of which are shared responses to both bacteria. In response to both bacteria, we see global shifts towards (1) the inclusion of cassette exons and (2) the expression of shorter 3'UTRs. We confirmed these patterns in cells profiled after 24hrs of infection, indicating that these are largely non-reversible signatures of an immune response. Interestingly, genes with cassette exon inclusion are more likely to show an up-regulation of gene expression levels following infection, providing evidence for possible concerted regulation of alternative splicing and transcriptional effects to achieve an optimal gene expression response. We were particularly interested in the 3' UTR shortening observed in these non-proliferating macrophages, since this is an alternative polyadenylation signature commonly observed in proliferating or cancerous cells. We show that changes in 3'UTR shortening across individuals often correlates with variation in associated gene expression level. Interestingly, the 3'UTR fragments lost in response to infection are significantly enriched for miRNA binding sites for a large number of miRNAs that we show to be differentially expressed following infection. Overall, our results suggest widespread regulation of immune response by mechanisms of mRNA processing, including possible co-opting of a signature of proliferation to regulate inter-individual variation in immune responses.

VISUALIZING HUMAN TRANSCRIPTION AT NUCLEOTIDE RESOLUTION USING NATIVE ELONGATING TRANSCRIPT SEQUENCING

Stirling Churchman

Harvard Medical School, Department of Genetics, Boston, MA

Major features of transcription by human RNA Polymerase II (Pol II) remain poorly defined due to a lack of quantitative approaches for visualizing Pol II progress at nucleotide resolution. We developed a simple and powerful approach for performing native elongating transcript sequencing (NET-seq) in human cells that globally maps strand-specific Pol II density at nucleotide resolution. NET-seq exposes a mode of antisense transcription that originates downstream and converges on transcription from the canonical promoter. Convergent transcription is associated with a distinctive chromatin configuration and is characteristic of lower-expressed genes. Integration of NET-seq with genomic footprinting data reveals stereotypic Pol II pausing coincident with transcription factor occupancy. Finally, exons retained in mature transcripts display Pol II pausing signatures that differ markedly from skipped exons, indicating an intrinsic capacity for Pol II to recognize exons with different processing fates. Together, human NET-seq exposes the topography and regulatory complexity of human gene expression.

CONSTRAINTS IN GENE EXPRESSION ACROSS TISSUES AND SPECIES

Alessandra Breschi[1], Dmitri D Pervouchine[1], Sarah Djebali[1], Carrie A Davis[2], Alex Dobin[2], Julien Lagarde[1], Roderic Guigó[1], Thomas R Gingeras[2]

[1]Centre for Genomic Regulation and UPF, Bioinformatics and Genomics, Barcelona, Spain, [2]Cold Spring Harbor Laboratory, Functional Genomics, Cold Spring Harbor, NY

Mouse has been a long-standing model for human biology and disease. Here we characterize by RNA-seq the transcriptional profiles of a large and heterogeneous collection of mouse tissues, augmenting the mouse transcriptome with thousands of novel transcript candidates. Comparison with human cell lines reveals substantial conservation of transcriptional programs, and uncovers a distinct class of genes with levels of expression that have been constrained early in vertebrate evolution. This core set of genes captures a substantial fraction of the transcriptional output of mammalian cells, and participates in basic housekeeping processes. Perturbation of these constrained genes is associated with significant phenotypes including embryonic lethality and cancer. Evolutionary constraint in gene expression levels is not reflected in the conservation of the genomic sequences, but is associated with conserved epigenetic marking, as well as with post-transcriptional regulatory program, including subcellular localization and alternative splicing. Genes with unconstrained gene expression, both across tissues and species, are likely to be involved in tissue and organism specificity. However, it is still unclear given a fixed evolutionary distance, which factor plays a more important role in transcriptome definition. In fact, controversy exists on whether transcriptome clustering of homologous tissues in human and mouse is dominated either by species or by tissue. Here we investigated transcriptomic data from six organs in seven amniotes (including mouse and human) to study the patterns of transcriptome variation across tissues and species. We decomposed the variance of each gene expression in the fraction contributed by species and the fraction contributed by tissue, and we found a continuum in the relative contribution of these two factors. Genes whose expression varies in different tissues but remains relatively constant across species drive the evolutionary constraint, while genes with a high variability across species, but constant in different tissues, contribute to tissue definition. Some normalization methods act on these relative variances and alter the contribution of species and tissues on the total expression variability, and are the basis of the contradictory results presented so far. Decomposing variation in gene expression is essential in order to identify those genes—that vary across tissues, but not across species—in which mouse is a good model for human biology.

IDENTIFYING DISEASE-ASSOCIATED GENETIC VARIANTS AFFECTING VITAMIN D RECEPTOR BINDING: A CHIP-EXO STUDY

Giuseppe Gallone[1], Antonio J Berlanga-Taylor[1], Wilfried Haerty[1], Giulio Disanto[1], Sreeram Ramagopalan[2], Chris P Ponting[1]

[1]University of Oxford, MRC Functional Genomics Unit, Oxford, United Kingdom, [2]AstraZeneca, Global Medicines Development, Observational Research Centre, Shanghai, China

We asked whether genetic risk for complex traits and diseases can arise from loss or gain of DNA-binding by VDR, the nuclear receptor for vitamin D, a modifiable environmental factor proposed to alter susceptibility to developing multiple complex diseases. 15,509 VDR-binding peaks were defined in at least 3 of 27 calcitriol–stimulated HapMap lymphoblastoid cell samples. These peaks occur preferentially within autoimmune disease-associated GWAS intervals, and also within enhancers that frequently contain a VDR:RXR heterodimeric binding motif. Pan-vertebrate conservation at nucleotide sites within this motif mirrors the base-loading of the motif's position weight matrix. In 88% of observations involving variants strongly affecting the VDR:RXR binding motif, the change in the strength of the VDR binding motif correctly predicted the direction of VDR binding affinity change. 1,695 high-confidence variants were significantly associated with variable VDR-binding using an analysis of allele-specific binding imbalance and/or QTL-association testing with Bayesian regression modelling.

Most of these 1,695 variants lie within annotated transcription factor binding peaks and/or enhancers, and 137 intersect DNase I sensitivity QTLs or LCL-specific gene expression QTLs (eQTLs). 13 of these variants directly overlap with GWAS lead SNPs from GRASP. Compared against a background of all VDR-binding sites, these 1,695 variants are strikingly enriched in: (i) strong enhancers, (ii) regions of strong selective constraint, (iii) DNAse I sensitivity or expression QTLs, and (iv) VDR:RXR binding motifs. 7 SNPs (2.0-fold enrichment; $p = 0.02$) overlap with the list of BROAD Probabilistic Identification of Causal SNPs (PICS) variants that are considered likely causal for altered susceptibility of one or more of 21 autoimmune diseases.

Our results support the hypothesis that variation at enhancers contributes to complex disease. They further suggest altered VDR-binding as a candidate mechanism for at least some complex traits.

PRIVACY AND INFORMED CONSENT IN AN ERA OF COMPUTATIONAL GENOMICS: A COMPARATIVE ANALYSIS OF ICELAND AND THE UNITED STATES

Donna M Gitter

Baruch College, City University of New York, Department of Law, New York, NY

The proposed research project is a comparative analysis of the implementation in Iceland and the United States of bioinformatics and computational genomics. Researchers can use these technologies to calculate the probability that an individual carries a particular genetic variant, without sequencing that person's DNA, thereby developing estimated data for inclusion in research databases. The increasing use of these technologies, coupled with the creation of genetic and medical databases, threaten to undermine the concepts of informed consent and privacy with respect to individuals' health data.

This research project will compare the use of bioinformatics and computational genomics in Iceland to develop estimated genetic data, and the associated privacy and informed consent concerns, with the application of these technologies to the Utah Population Database (UPDB) at the University of Utah. The UPDB is the only database of its kind in the United States and one of few such resources in the world. What makes the database unique is the extensive set of family genealogies, maintained by the Church of Jesus Christ of Latter-Day Saints, in which family members are linked to demographic and medical information, analogous to the genealogical records in Iceland. Moreover, while not as consanguine as the Icelandic population, due to immigration patterns in the U.S., most Utahans are descended from a common set of European ancestors. Researchers have identified the UPDB as one of the world's richest sources of detailed information useful for research on genetics, epidemiology, demography, and public health. As with Iceland, advances in the fields of bioinformatics and computational genomics will permit researchers to develop estimated data about Utah residents who did not agree to contribute DNA to a research study. This research will inform the creation of appropriate privacy and informed consent policies for the use of estimated data.
.

A NETWORK ENSEMBLE OF microRNA AND GENE EXPRESSION IN OVARIAN CANCER

Andrew Quitadamo, Benika Hall, Xinghua Shi

University of North Carolina at Charlotte, Bioinformatics and Genomics, Charlotte, NC

Ovarian cancer is a deadly female reproductive cancer. Understanding the biological mechanisms underlying ovarian cancer could help lead to quicker and more accurate diagnosis and more effective treatments. Both changes in microRNA(miRNA) expression and miRNA/mRNA dysregulation have been associated with ovarian cancer. With the availability of whole-genome miRNA and mRNA sequencing we now have new potentials to study these associations. In this study, we performed a comprehensive analysis of miRNA and mRNA expression in ovarian cancer using an integrative network approach combined with eQTL mapping. Our method is composed of expanding networks from eQTL associations, building network associations in eQTL mapping, incorporating miRNA target predictions, and then combining the networks into an integrated network. This integrated network includes various types of relationship among miRNAs and genes including miRNA eQTL associations, miRNAs and their targets, protein-protein interactions, co-expressions among miRNAs and genes respectively. Applied to the ovarian cancer data set from The Cancer Genome Atlas (TCGA), we created an integrated network that provided a more inclusive view of miRNA and gene expression in ovarian cancer. Simply analyzing each interaction component in isolation, such as the eQTL associations, the miRNA-target interactions or the protein-protein interactions, would create a much more limited network than the integrated one. In summary, we developed an integrative approach to construct an integrative network that illustrates the complex interplay among miRNA and gene expression from a systems perspective. Such an integrative network can further our understanding of the underlying mechanisms in studying ovarian cancer.

GREAT APE Y CHROMOSOME DIVERSITY REFLECTS SOCIAL STRUCTURE AND SEX-BIASED BEHAVIOURS

Pille Hallast[1], Pierpaolo Maisano Delser[1,2], Chiara Batini[1], Daniel Zadik[1], Werner Schempp[3], Mariano Rocchi[4], Chris Tyler-Smith[5], Mark A Jobling[1]

[1]University of Leicester, Department of Genetics, Leicester, United Kingdom, [2]Museum National d'Histoire Naturelle, Ecole Pratique des Hautes Etudes, Paris, France, [3]University of Freiburg, Institute of Human Genetics, Freiburg, Germany, [4]University of Bari, Department of Biology, Bari, Italy, [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

The diversity of the male-specific region of the Y chromosome (MSY) in humans has been widely exploited to shed light on population history, sex-biased processes, and social selection. For humans a robust MSY phylogeny of haplogroups exists based on thousands of slow-mutating SNPs defining the relationships between different lineages. Analysis of the same locus in great apes, however, has been scanty and has mostly been based on MSY-specific STRs rather than SNPs. This is partially due to the fact that among great apes the Y chromosome has been sequenced only for human and chimpanzee.

Here, we used a custom enrichment approach based on the human reference and sequenced ~3.9 Mb of DNA from the genomes of 19 male great apes to high coverage. Combining our data with MSY sequences extracted from the whole genomes of 20 additional individuals (Prado-Martinez et al. 2013) yields a total sample of four bonobos, 19 chimpanzees, 10 gorillas, and two Bornean and four Sumatran orangutans. Analysis of these great-ape sequences retained between 2 and 3.6 Mb of human-orthologous MSY material per species, identified species-specific deletions and duplications and thousands of novel Y-chromosome SNPs defining highly-resolved MSY phylogenies. Comparison with mtDNA phylogenies from the same individuals, as well as autosomal and X-chromosomal SNP data provides insights into the social structure and sex-specific behaviors of each species.

LARGE MULTIALLELIC COPY NUMBER VARIATION IN HUMANS

Robert E Handsaker[1,2,3], Vanessa Van Doren[1,2,3], Jennifer R Berman[4], Giulio Genovese[1,2,3], Seva Kashin[1,2,3], Linda M Boettger[1], Steven A McCarroll[1,2,3]

[1]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA, [2]Stanley Center for Psychiatric Research, Cambridge, MA, [3]Harvard Medical School, Department of Genetics, Boston, MA, [4]Bio-Rad Laboratories, Inc., Digital Biology Center, Pleasanton, CA

Copy number variation (CNV) is widespread in the human population and affects thousands of human genes, exons and functional elements. Rare and de novo deletions and duplications, often involving the deletion or duplication of hundreds of kilobases of DNA, can be substantial risk factors in many human diseases. Progress has been made in recent years in understanding simple, di-allelic CNV that arises by one-step deletion or duplication. But perhaps the most intriguing form of CNV is that one that is today least characterized. A substantial fraction of all inherited CNV in humans is contributed by variants that appear in different genomes at widely different numbers of copies (from zero up to 12 or more) and cannot result from just two segregating alleles. The difficulty of molecularly typing these multiallelic CNV (mCNVs) have excluded them from most studies of human genetic and phenotypic variation. As a result, the structural alleles and haplotypes that segregate at such loci and their relationship to phenotype remain largely unknown.

We developed ways to use increasingly abundant whole genome sequencing data to identify the alleles and haplotypes at these loci. We applied these methods to 849 genomes sequenced in Phase 1 of the 1000 Genomes Project at low coverage (4-8x) to identify 8,659 CNV loci, of which 1,356 appear to arise from three or more segregating alleles. The resulting copy-number inferences are strongly concordant (at the level of integer copy-number genotypes) with data from molecular-biological assays that we developed at the same time.

Strikingly, we find that mCNVs contribute most of the gene dosage variation in humans, seven times the contribution of simpler di-allelic deletions and duplications. Moreover, this variation in gene dosage appears to generate concomitant variation in gene expression, with strong correlation at almost all of the loci evaluated.

We describe a phenomenon of "runway" duplication haplotypes, on which specific genes, including *HPR* and *ORM1* have mutated to high copy number in specific human populations during the past 50,000 years, while remaining relatively constant on other haplotypes and in other populations.

We describe our initial strategies for analyzing multiallelic CNVs in large cohorts via imputation and provide a community resource to support imputation of these mCNVs, thus enabling these variants to be analyzed systematically for their relationships to phenotypes.

UNCOVERING SINGLE NUCLEOTIDE POLYMORPHISMS AFFECTING SLEEP DURATION IN *DROSOPHILA* USING ARTIFICIAL SELECTION

Susan T Harbison[1], Yazmin L Serrano Negron[1], Nancy F Hansen[2]

[1]Laboratory of Systems Genetics, National Heart Lung and Blood Institute, Bethesda, MD, [2]Comparative Genomics Analysis Unit, National Human Genome Research Institute, Bethesda, MD

Recent work suggests that both sleep disorders and voluntary sleep loss are risk factors for human metabolic and cardiovascular disease; however, the role of sleep in human health remains elusive. *Drosophila* have the same behavioral characteristics of sleep as mammals, enabling the use of this powerful model organism to decipher the genetic architecture of sleep. We constructed a highly mixed variable population from long- and short-sleeping lines of the *Drosophila* Genetic Reference Panel (DGRP). We split the highly variable mixed population into replicate populations and subjected them to 13 generations of artificial selection for long and short night sleep. We also maintained two unselected populations contemporaneously with the selected populations in order to account for inbreeding and genetic drift. The response to selection was very rapid; after 13 generations, night sleep in the long and short sleep populations diverged by an average of 10 hours. Estimates of heritability (± SE of the regression coefficient) for long-sleeping populations were 0.310 ± 0.02 and 0.238 ± 0.03 (replicates 1 and 2, respectively), and -0.179 ± .03 and -0.215 ± 0.02 for replicates 1 and 2 of the short-sleeping populations; all heritability estimates were significantly different from zero (P <0.0001). The slope of the regression for the control populations was -0.108 ± 0.31 and -0.271 ± 0.21 for replicates 1 and 2, respectively; neither of these slope estimates was statistically significant. Day sleep duration, night average bout length, and day bout number had significant correlated responses to selection for night sleep. We examined changes in the underlying genotypes of these flies in response to selection using two measures. First, we used a PCR-based assay to genotype 4,371 flies at 96 markers previously associated with sleep; we measured both sleep and genotype in single flies in three separate generations of selection. Second, we sequenced the genome of pooled flies from each selection population in seven different generations. We assessed genotypes and calculated allele frequencies in the sequence data at 2,154,277 polymorphic locations known to segregate among the parental DGRP lines. Preliminary results show dynamic shifts in allele frequencies in both the selected and control populations. Additional analyses will address the role of these shifts in the genetic architecture of sleep.

TRANSCRIPTOME DYNAMICS DURING MOUSE EMBRYONIC BRAIN DEVELOPMENT

Manoj Hariharan[1], Yupeng He[1], Rosa Castanon[1], Joseph R Nery[1], Len Pennacchio[2], Axel Visel[2], Joseph R Ecker[1]

[1]The Salk Institute for Biological Studies, Genomic Analysis Laboratory, La Jolla, CA, [2]Lawrence Berkeley National Laboratory, Berkeley, CA, [3]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA

Human and mouse share a very close resemblance during early developmental stages – morphologically and at the molecular level. Several severe birth defects in humans (like orofacial clefts (which includes clefts of the lip and/or palate) and congenital heart defects) manifest at these early stages of development and have a concurrent timeline in mouse. Since it is difficult to obtain tissue samples from human fetuses, mouse tissues serve as surrogates to study the various molecular trajectories involved during development.

We produced and analyzed transcriptome profiles of various tissues across three developmental time points – E11.5, E14.5, E16.5 and P0 from mouse which correspond to 33 days, 52 days, 59 days and neonatal in human. The tissues include three brain regions and liver. We performed RNA-seq (with an average of >60M reads) in order to quantify the presence of all transcripts. We also compare their abundances across the various tissues and across developmental stages. We observed subsets of genes that are differentially expressed across development and tissue of expression. We also identified key master regulators by analyzing the spatio-temporal expression of transcription factors (TFs), which drive the expression of downstream genes. Some of these are TFs with known functions in the tissue of expression while others are TFs with no obvious roles. These could be co-factors or competitive in their association with other TFs. We built gene regulatory networks (GRN) based on the spatio-temporal expression overlaid with the cellular role of the master regulators and their downstream genes. We integrate this information with known protein-protein interaction maps to ascertain the co-regulatory role of TFs. Using these approaches we identify enriched network modules, which can be used to assess the functional relevance of the co-regulated components during development. Apart from this, the GRN model also allows for perturbation of key nodes, which enables us to delve deeper into the global effects of aberrant gene expression and regulation, including clinical manifestations.

TISSUE-SPECIFIC IDENTIFICATION OF LNCRNAS IN MAMMALIAN GENOMES USING TARGETED RACESEQ AND CAPTURE SEQ.

Jennifer Harrow[1], Julien Lagarde[2], Javier Santoyo-Lopez[3], Barbara Uszczynska[2], Electra Tapanari[1], Laurens Wilming[1], Sarah Djebali[2], Anne-Maud Ferreira[4], Rory Johnson [2], Alexandre Reymond [4], Roderic Guigo[2]

[1]Wellcome Trust Sanger Institute, Computational Biology, Hinxton, United Kingdom, [2]Center for Genomic Regulation, Informatics, Barcelona, Spain, [3]Genomics and Bioinformatics Platform, Andalusia, Seville, Spain, [4]University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland

Many groups are generating and data-mining a wealth of illumina RNAseq data available in the public domain to identify “tens of thousands” of novel long non-coding RNAs from a host of different organisms. The reliability of these models is variable and can depend on length and quality of input data and algorithms used. As part of the GENCODE consortium we are combining different next generation sequencing resources to produce a reference non-coding gene catalogue in human and mouse publicly available in UCSC and Ensembl browsers. Currently we have identified around 15 000 human loci which have potential of being spliced long non-coding (lnc) genes.. Nomenclature and classification of these entities is inconsistent across species and usually based on proximity to other coding genes rather based on their function. As part of a pilot project, we have analysed 400 lncRNAs identified as partial transcripts because of lack of CAGE data or polyadenylation signals. We compared data from nested and un-nested RACE-seq in 7 human tissues. The reads revealed a wealth of previously unknown isoforms, and we investigated the quality of the extensions using standard TSS and TTS support evidence. The majority of lncRNAs sequences appear to be poorly conserved on the sequence level, yet annotating both mouse and human regions in parallel helps identify syntenically equivalent transcripts. We are using capture-seq technology followed by PacBio sequencing on thousands of non-coding RNA targets to compare expression of lncRNAs in human and mouse. In summary we highlight how this mix of next generation data may double the number of genes in GENCODE and produce a new challenge in cataloguing functional lncRNAs for mammalian genomes.

DE NOVO ASSEMBLY AND STRUCTURAL VARIATION DISCOVERY IN IN HUMAN DISEASE AND NON-DISEASE STATE GENOMES USING EXTREMELY LONG SINGLE-MOLECULE IMAGING

Alex Hastie, Ernest Lam, Tiffany Liang, Andy Pang, Saki Chan, Han Cao

BioNano Genomics, Research, San Diego, CA

Structural variation analysis (SVA) of human genomes is usually a reference based process and therefor biased and incomplete. In order to have a comprehensive analysis of structural variation, a *de novo* approach is needed. *De novo* genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in the human genome. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses. As a result of the remaining limitations of DNA sequencing and analysis technologies, it is not feasible to create high quality assemblies of individuals to detect and interpret the many types of structural variation that are refractory to high throughput or short-read technologies.

We present a single molecule genome analysis system (Irys®) based on NanoChannel Array technology that linearizes extremely long DNA molecules for direct observation. This high-throughput platform automates the imaging of single molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness for unambiguous assembly of complex genomes. High resolution genome maps assembled de novo preserve long-range structural information necessary for structural variation detection and assembly applications. Dozens of human genomes have been de novo assembled by Irys to date, including cancer genomes. Structural variation analysis reveals insertions, deletions, inversions and translocations. We have generated genome maps for two trios, CEPH trio (NA12878, NA12891 and NA12892) and an Ashkenazi jewish trio (AJ; NA24385, NA24143 and NA24149). From these genome maps, we detect hundreds of structural variants, including large deletions that delete genes in the mother and son from the AJ trio. We have also investigated the amylase locus in both trios as well as ~20 other individuals and have found at least 15 different structural variants. Human amylase genes have variable copy number and this variation is believed to have been evolved to adapt to increase starch intake, we are able to identify multiple copy neutral variants, i.e. inversions, for each for the same copy number variants. Each genome shows additional dramatic structural variation, including many megabases of variation within genomic regions not included in the public reference genome assembly, underscoring the need for more de novo approaches to genome analysis.

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

Jim Havrilla[1], Aaron Quinlan[2]

[1]University of Virginia, Biochemistry and Molecular Genetics, Charlottesville, VA, [2]University of Utah, Department of Human Genetics, Salt Lake City, UT

Numerous methods exist to predict the impact of a genetic variant on protein function. For example, the RVIS (Residual Variation Tolerance Score) study from Petrovski et al., provides a gene-wide score by regressing the number of common missense variants vs. the total number of variants. In contrast, the CADD (Combined Annotation Dependent Depletion) 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 with a Support Vector Machine. A more recent approach from Gulko et al. integrates DNase-seq, RNA-seq and histone modification data to create an evolution-based measure of phenotypic function (the "fitCons" score).

None of these methods, however, directly utilize protein domain information – they only look at genes in the broader sense, not the numerous small and large functional portions of a protein for which they actually code. Accordingly, by comparing the Exome Aggregation Consortium's catalog of protein-coding genetic variation from more than 60,000 human exomes with the Pfam protein domain database, we have comprehensively measured the landscape of genetic variation among all characterized protein domains. Computing the non-synonymous to synonymous (dN/dS) variant ratios as well as the distribution of those ratios for each domain per protein has allowed 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 fundamental rationale of the model is that variants overlapping protein domains that are tolerant of 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 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.

DYNAMIC DNA METHYLATION LANDSCAPE DURING MOUSE EMBRYONIC BRAIN DEVELOPMENT

Yupeng He[1,2], Manoj Hariharan[2], Chongyuan Luo[2], Joseph R Nery[2], Rosa Castanon[2], Mark A Urich[2], Huaming Chen[2], Yin Shen[3], Bin Li[4], Wei Wang[5,6], Axel Visel[7], Len A Pennacchio[7], Bing Ren[4], Joseph R Ecker[2,8]

[1]Bioinformatics Program, University of California, San Diego, La Jolla, CA, [2]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, [3]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, [4], Ludwig Institute for Cancer Research, La Jolla, CA, [5]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, [6]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, [7]Genomics Division, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA, [8] Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA

Cytosine methylation is essential for mammalian brain development. Previous studies revealed that methylation can occur at both CG and non-CG context in brain. However, little is known about the dynamics of the DNA methylation landscape in the brain over the trajectory of mammalian embryogenesis. In this study, we generated deep (60x coverage) whole-genome bisulfite sequencing data for three brain regions, forebrain, midbrain and hindbrain, from E11.5, E14.5, E16.5 and P0 mouse embryo as well as several other tissues from the same stages.
Comparing methylation in the CG context of three brain regions and other tissues across developmental stages, we identified differentially methylated regions (DMRs) that readily distinguish brain regions from each other and from other tissue samples. Majority of these CG DMRs were near corresponding tissue-related genes (i.e. heart DMRs identify genes with heart specific functions) and contain enhancer-like chromatin modifications, implicating these as distal regulatory elements. By tracking temporal CG methylation changes with a tissue, hundreds of thousand of regions showing epigenomic dynamics (developmental CG DMRs) were pinpointed. Strikingly, as tissues mature, the vast majority of DMRs show a loss of DNA methylation, which is accompanied by an increase of active chromatin marks, indicating that CG demethylation and regulatory element activation are general trends during embryonic tissue development.
In contrast, during brain maturation, accumulation of methylation that occurred at non-CG site (mCH) was observed to occur in the bodies of genes involved in the early development of brain. Interestingly, the accumulation of mCH in various brain regions matched the developmental timing of brain maturation, with mCH first observed in in hindbrain, then midbrain and finally forebrain. mCH accumulated in the bodies of many neuronal transcription factors and genes involved in the early development of brain, suggesting that mCH is an important marker of brain development.
Our study provides the first whole genome base-resolution maps of temporal DNA methylation across a comprehensive set of mouse tissues during embryogenesis. These deep data sets greatly extends our view of the dynamic epigenome occurring during early development which may provides new insight into the regulatory elements that guiding tissue differentiation during fetal development.

FUNCTIONAL ANALYSIS OF THE ETV6/RUNX1 FUSED GENE IN ALL

Jason Wray[1], Dapeng Wang[1], Sladjana Gagrica[1], Shamit Soneji[2], Amit Mandoli[3], Joost H Martens[3], Henk G Stunnenberg[3], Javier Herrero[1], Tariq Enver[1]

[1]Universty College London, Cancer Institute, London, United Kingdom, [2]Lund University, Division of Molecular Medicine and Gene Therapy, Lund, Sweden, [3]Radboud University Nijmegen, Centre for Molecular Life Sciences, Nijmegen, Netherlands

Childhood Acute Lymphoblastic Leukaemia (ALL) is the most common form of childhood leukaemia. 25% of these cancers are characterised by the pre-malignant fusion of two transcription factors, ETV6 and RUNX1. The translocation (t(12;21)) associated with this fusion is frequently associated with the loss of PAX5 and the second ETV6 allele. In normal conditions, RUNX1 requires CBFβ to bind its target. RUNX1 can repress transcription when it associates with SIN3A and recruits HDAC, de-acetylating H3. RUNX1 can also promote transcription through recruitment of p300 HAT. ETV6/RUNX1 retains the ability to bind RUNX1 targets through the Runt domain but recruitment of corepressors by the ETV6 moiety is thought to convert its function to a dominant repressor. The fusion also inherits the Pointed Domain (PD) of ETV6 which may mediate homo- or hetero- dimerization with other PD-containing proteins.

We identify direct targets of the ETV6/RUNX1 in three patients as well as in the Reh cell line. In order to disentangle the competition between the wild-type RUNX1 and the ETV6/RUNX1, we have developed new systems in the human NALM6 and the mouse BA/F3 cell lines with ETV6/RUNX1 transgenes.

We use a BA/F3 murine pro B cell line with an inducible ETV6/RUNX1 transgene. We performed ChIP-seq for both RUNX1 and ETV6/RUNX1 as well as for the following histone marks: H3K27ac, H3K27me3, H3K36me3, H3K4me1 and H3K4me3 to understand how the fused gene affects the chromatin. We find that the main effect of ETV6/RUNX1 is a decrease in H3K27ac.

In the human NALM6 leukaemic cell line, we also observe how ETV6/RUNX1 can induce histone deacetylation at RUNX1 binding sites. In a similar experiment, we express an ETV6/RUNX1 transgene (ETV6/RUNX1-R139G) where the runt domain from the RUNX1 moiety has been mutated rendering it unable to bind DNA. This results in an even more pronounced decrease of H3K27ac levels around RUNX1 target regions, presumably the result of competition for CBFβ hampering RUNX1 binding. ETV6/RUNX1 lacking the PD retains HDAC activity and binds to the same sites as ETV6/RUNX1 but also to novel sites indicating that the PD alters the DNA-binding behaviour of RUNX1.

In summary, we propose two alternative mechanisms by which ETV6/RUNX1 can decrease histone acetylation at RUNX1 binding sites. 1. Through direct binding and recruitment of HDACs. 2. Indirectly, through competition for CBFβ and subsequent loss of RUNX1 binding and HAT activity

CHROMOSOME-SCALE SCAFFOLDING OF THE MAP-BASED REFERENCE ASSEMBLY OF BARLEY BY CHROMATIN INTERACTIONS

Axel Himmelbach[1], Mascher Martin[2], Beier Sebastian[2], Scholz Uwe[2], Stein Nils[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Department of Genebank, 06466 Stadt Seeland, Germany, [2]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Department of Breeding Research, 06466 Stadt Seeland, Germany

Barley is a cereal grass that is both an important crop species and a classical genetic model organism. The recent progress in high-throughput sequencing and mapping technology has boosted the development of advanced sequence resources for the huge (5 Gbp) and highly repetitive barley genome. The International Barley Genome Sequencing Consortium (http://www.barleygenome.org) is currently constructing a high-quality reference assembly of the barley genome by sequencing a minimum tiling path of overlapping bacterial artificial chromosomes (BACs). Individual BACs are assembled from shotgun sequencing reads of multiplexed short-insert paired-end and long-distance mate-pairs libraries. Recently chromosome conformation capture sequencing (HiC and TCC) was presented as a powerful tool for scaffolding whole genome assemblies in human (Burton et al. 2013, doi:10.1038/nbt.2727). We are testing this approach now for scaffolding of the map-based reference sequence assembly of the 5 Gbp barley genome and will present initial results.

THE ANNOTATION INTEGRATOR: A NEW WAY TO COMBINE DATA SOURCES UNDERLYING THE UCSC GENOME BROWSER

Angie S Hinrichs, Kate R Rosenbloom, Matthew L Speir, Donna Karolchik, Ann S Zweig, Robert M Kuhn, W J Kent

University of California Santa Cruz, UC Santa Cruz Genomics Institute, Santa Cruz, CA

Many users of the UCSC Genome Browser have asked for a way to join multiple annotation sources by their genomic positions -- for example to find genes that overlap items in a user's custom track -- while including all annotation columns from all sources in the output. In response, we have developed a new tool, the Annotation Integrator (AI), which can combine data from up to five annotation sources based on the overlap of feature positions along the reference genome and output the entire annotation data set or a user-selected subset of annotation columns. The output is formatted as tab-separated text that can be viewed in the web browser window or downloaded as a file, optionally compressed with gzip. Like the Genome Browser, the AI can integrate data from the browser database, user-provided custom tracks, and track hubs. Internally, the AI uses new data access libraries with a streaming architecture in order to start sending results as soon as possible. In the coming year, we plan to extend the AI's functionality to include popular features of the UCSC Table Browser, for example filters on column values and more fine-grained control over how overlap is defined.

EXPLORING BREAST CANCER HETEROGENEITY THROUGH LOW-INPUT RNA-SEQ DATA IN DUCTAL CARCINOMA IN SITU (DCIS)

Yu-Jui Ho, Molly Hammell

Cold Spring Harbor Laboratory, Watson School of Biological Science, Cold Spring Harbor, NY

Ductal Carcinoma In Situ (DCIS) is currently believed to be the precursor of Invasive Ductal Carcinoma (IDC), the most common form of breast cancer, accounting for 80% of all breast cancers. It has been estimated that more than 50% of all patients diagnosed with DCIS would never progress to invasive cancer in the patient's lifetime, yet nearly all patients opt for treatment to avoid that risk. This reflects an urgent need for understanding the pathway for progression from relatively benign DCIS to the malignant IDC. Adequately characterizing DCIS lesions, however, requires new experimental and computational tools that can handle the extremely small cell counts typically seen – on the order of 10-100 cells per biopsy section.

Here we present an approach for characterizing DCIS tissue samples by a supervised classification method. Our rationale is to find common gene expression patterns that can be used to classify unknown samples, and compare with existing molecular subtypes. By comparing early non-invasive DCIS lesions to IDC in patients with co-occurring IDC or later development of IDC, we aim to study heterogeneity and cancer progression in DCIS.

To this end, we focus on improving the computational analysis of RNA-Seq data. We show that a combination of improved statistical approaches and better normalization procedures can help identify significant changes in low input samples. We present a comparison to existing RNA-Seq analysis software and preliminary analysis of DCIS transcriptomes.

MULTIPLE HAPLOTYPE-RESOLVED GENOMES REVEAL POPULATION LEVEL GENE AND PROTEIN DIPLOTYPE PATTERNS.

Margret R Hoehe[1], George M Church[2], Hans Lehrach[1], Eun-Kyung Suk[1], Thomas Huebsch[1]

[1]Max Planck Institute for Molecular Genetics, Vertebrate Genomics, Berlin, Germany, [2]Harvard Medical School, Department of Genetics, Boston, MA

To fully understand human biology and link genotype to phenotype, the phase of DNA variants must be known. Here we present a first comprehensive analysis of multiple haplotype-resolved genomes to assess the nature and variation of haplotypes and their pairs, diplotypes, in European population samples. We use a set of 14 haplotype-resolved genomes generated by fosmid clone-based sequencing, complemented and expanded by up to 372 statistically haplotype-resolved genomes from the 1000 Genomes Project. We find immense diversity of both haploid and diploid gene forms, up to 4.1 and 3.9 million corresponding to 249 and 235 per gene on average. Less than 15% of autosomal genes have a predominant form. This diversity converges upon a 'common diplotypic proteome', a set of 4,269 genes encoding two different proteins in over 30% of genomes. Moreover, we find that mutations predicted to alter protein function exist, in each of the 386 genomes, significantly more frequently in cis than in trans configurations, with an average cis/trans ratio of 60:40. In addition, we observe different classes of cis and trans-abundant genes. This work identifies key features characterizing the diplotypic nature of human genomes and provides a conceptual and analytical framework, rich resources and novel hypotheses on the functional importance of diploidy. Ultimately, our work contributes novel insights into the 'true nature of genetic variation', which cannot be understood without knowing the distribution of variants on each of the two parental sets of chromosomes.

TRACKING DATA PROVENANCE AT THE ENCODE DCC

Eurie L Hong[1], Venkat S Malladi[1], Benjamin C Hitz[1], Esther T Chan[1], Jean M Davidson[1], Timothy R Dreszer[1], Marcus Ho[1], Brian T Lee[2], Nikhil R Podduturi[1], Laurence D Rowe[1], Cricket A Sloan[1], J. Seth Strattan[1], Forrest Tanaka[1], W. James Kent[2], J. Michael Cherry[1]

[1]Stanford University, Genetics, Stanford, CA, [2]University of California, Santa Cruz, Center for Biomolecular Science and Engineering, Santa Cruz, CA

The provenance of experimental reagents and transparency of computational analyses are essential to compare, reproduce, and interpret experimental data. The task of tracking this information consistently across diverse sequencing assays can be especially challenging in large projects like the ENCODE (ENcyclopedia Of DNA Elements) Consortium that perform 40+ genomic assays using 400+ cell and tissue types. The identification of a transcription factor binding site or the quantification of a transcript's expression level is dependent on the software versions, the parameters used when running that software version, which files were used, the library preparation methods, and how the biological samples were selected or obtained. To capture the provenance of experimental methods and computational results, the ENCODE DCC (Data Coordination Center) has created a rich data model that represents how experiments were performed, what software and pipelines were used, and which files were analyzed. These details of the experimental and computational methods, known as metadata, can then be used to identify related data for further analysis, interpret the results of the assays, and allow reproducibility of pipelines that are run to generate the data. All metadata and data generated by the ENCODE Consortium are freely available at at the ENCODE Portal (https://www.encodeproject.org/).

POPULATION GENOMICS OF A GLOBAL SAMPLE OF 200 *PLASMODIUM VIVAX* MALARIA PARASITES

Daniel N Hupalo[1], Zunping Luo[1], Patrick L Sutton[1], Eli Moss[2], Daniel E Neafsy[2], Jane M Carlton[1]

[1]New York University, Center for Genomics and Systems Biology, New York City, NY, [2]The Broad Institute of MIT and Harvard, Malaria Genome Sequencing and Analysis Group, Cambridge, MA

The parasite *Plasmodium vivax*, implicated in less virulent but longer lasting cases of malaria, is a major public health concern, causing the majority of malaria infections each year outside Africa. Here we describe the collection and sequencing of more than 170 clinical isolates of *P. vivax* sampled from eight regions around the world, including New World isolates from Central and South America as well as Old World isolates from East Asia, Southeast Asia and Melanesia. We performed population genomic analysis of these genomes combined with previously sequenced isolates, to form a global data set of ~200 *P. vivax* genomes. We addressed two issues that complicate *P. vivax* population genomics. First, owing to the capacity of *P. vivax* parasites to relapse from a dormant liver stage, patient samples are often multi-clonal. We describe both molecular and bioinformatic methods to assess the complexity of infections within clinical isolates. Second, *P. vivax* infections present with low parasite densities, so that DNA extracted from patient samples is dominated by human host DNA. We have used hybrid selection to enrich parasite DNA from clinical samples displaying a diversity of parasite densities. We used the resulting high coverage genome sequences to identify over 150,000 SNPs and generate the first global genome wide polymorphism profile for *P. vivax*. Using this panel of variation we identified strong population structure within *P. vivax*, including a genetically distinct population from Papua New Guinea, and the likely European origin of new world *P. vivax* populations. We also evaluated diversity across all *P. vivax* genes, identifying divergent regions in antigen genes between and within populations. Lastly, we investigated several known genes of interest, including the human host Duffy antigen genotype, using in silico and molecular biology techniques. Our analyses highlight the utility of generating a global dataset of genomes for an infectious disease, which can help clarify demographic history and adaptation within the species, in addition to providing new paths towards answering long standing biological questions.

SPATIOTEMPORAL EXPRESSION OF ALTERNATIVELY SPLICED ISOFORMS IN THE DEVELOPING HUMAN BRAIN

Lilia M Iakoucheva, Guan N Lin, Roser Corominas, Jonathan Sebat, William Yang

University of California San Diego, Psychiatry, La Jolla, CA

Alternative splicing plays an important role during brain development. However, the spatiotemporal variability in expression levels of different splicing isoforms of the same gene has not yet been investigated. To gain an insight into the global variation of isoform expression levels during brain development we used RNA-seq data from the BrainSpan to construct a comprehensive spatiotemporal isoform transcriptome.

We observed a greater variability in expression levels of different isoforms across developmental periods than across brain regions. Among ~61,000 brain-expressed isoforms of ~14,000 protein-coding genes, 10% had High temporal expression variability, 35% had Medium and 55% had Low. The High cluster was significantly enriched in transcripts with alternatively spliced microexons (3-27nt) (P=1.96x10-6) and with de novo mutations from the patients with neurodevelopmental disorders (P=0.05) in late fetal period. In contrast, neither Medium nor Low variability cluster was enriched in transcripts with microexons or with de novo mutations. Furthermore, none of the three clusters was enriched in the transcripts with the de novo mutations from healthy controls.

Next, we investigated genes that switch High and Low variability clusters in the adjacent developmental periods. Remarkably, the “switch” genes were highly enriched in neuronal-specific processes (synaptic transmission, P=0.001; nervous system development, P=0.014) and in de novo mutations in cases (P=1.3x10-5), but not in controls. In contrast, the genes that did not switch variability clusters were enriched in the general cellular functions (mRNA metabolism, P=1.27x10-7, gene expression, P=9.45x10-7), and also in housekeeping and ribosomal genes. These results suggest that (1) splicing isoforms of some genes have high expression variability in developing human brain; (2) transcripts with alternatively spliced microexons and de novo mutations are enriched in late fetal developmental period; (3) neuronal-specific isoforms have expression variability that is discordant across brain developmental periods. The relationships between the variability in splicing isoform expression and neurodevelopmental disorders need to be further investigated.

PROBING THE BIOLOGICAL MECHANISMS OF COMPLEX TRAIT ETIOLOGY VIA GENETICALLY PREDICTED ENDOPHENOTYPES

Heather E Wheeler[1], Eric R Gamazon[2], Kaanan Shah[1], Sahar Mozaffari[1], Keston Aquino-Michaels[1], Barbara Stranger[1], Dan L Nicolae[1], Nancy J Cox[2], Hae Kyung Im[1]

[1]The University of Chicago, Medicine, Chicago, IL, [2]Vanderbilt University, Medicine, Nashville, IL

Thousands of well-replicated genetic variants associated with complex traits have been discovered in the last decade, which advanced our understanding of common diseases. However, for a large number of these discoveries the biological mechanisms underlying the associations are not well understood. On the other hand, there has been growing evidence that regulation of intermediate molecular traits (endophenotypes) is likely to play a major role. Thus, much effort has been devoted to investigate the effects of genetic variation on gene expression traits. For example, the GTEx consortium is set to collect and sequence DNA and RNA samples from 900 organ donors across over 40 tissues. Many other molecular traits such as protein levels, telomere length, and methylation are being assayed as well.

To harness the wealth of cross-tissue and tissue specific molecular data generated by the GTEx consortium as well as the ever growing genotype/phenotype data available in dbGaP and other repositories, we propose a function-based association test called PrediXcan that tests the mediating effects of endophenotypes on complex traits.

The approach estimates the component of the endophenotype determined by an individual's genetic profile and correlates the "imputed" endophenotypes with the phenotype under investigation to identify genes or other functional units involved in the etiology of the phenotype. The genetically regulated endophenotype is estimated using whole-genome tissue-dependent prediction models trained with reference omic datasets. PrediXcan enjoys the benefits of gene-based approaches such as reduced multiple testing burden, more comprehensive annotation of gene function compared to that derived from single variants, and a principled approach to the design of follow-up experiments. Since no actual molecular data are used in the analysis of GWAS data - only in silico levels - reverse causality problems are largely avoided. PrediXcan harnesses reference datasets with genomic and molecular traits for disease mapping studies. Our results demonstrate that PrediXcan can detect known and novel genes associated with disease traits and provide insights into the mechanism of these associations.

LIMB LOSS AND THE EVOLUTION OF APPENDAGE ENHANCERS IN SNAKE GENOMES

Carlos R Infante, Alexandra G Mihala, Sungdae Park, Douglas B Menke

University of Georgia, Genetics, Athens, GA

Lizards and snakes (squamates) last shared a common ancestor approximately 150 million years ago in the late Jurassic. Among snakes, complete limb loss evolved approximately 100 million years ago. Using functional and comparative genomics we have identified active cis-regulatory elements (enhancers) in developing vertebrate appendages and studied their fate in snake genomes. To identify active enhancers in developing appendages, we performed ChIP-Seq on embryonic tissues from the forelimb, hindlimb, and genital tubercle of the lizard Anolis carolinensis using an antibody against H3K27ac, which marks active enhancers and promoters. In parallel we performed H3K27ac ChIP-Seq on similar embryonic tissues from the mouse, which last shared a common ancestor with squamates approximately 300 million years ago. In both the lizard and mouse, we identified more than 20,000 genomic regions with significant enrichment compared to background levels. Comparisons between ChIP-Seq datasets revealed substantial overlap of H3K27ac-marked regions between forelimb and hindlimb tissues, as well as between limbs and the genital tubercle in both the mouse and the lizard relative to other non-appendage tissues. Additionally, we found significant enrichment of limb-associated genes near enhancers active in genital tubercle. Our examination of these appendage enhancers in snake genomes revealed extensive sequence conservation in limb enhancers in this group of animals with ancient limb loss. Our results indicate that the regulatory networks active in the developing appendages involve very similar genomic contexts, and may reflect shared developmental control of the vertebrate limbs and external genitalia.

CHARACTERIZING THE COMPLETE METAGENOME, INCLUDING HIGH GC/AT MICROBIAL MEMBERS

Jonathan C Irish, Rachel R Spurbeck, Sukhinder K Sandhu, Laurie Kurihara, Tim Harkins, Vladimir Makarov

Swift Biosciences Inc., Research and Development, Ann Arbor, MI

Next Generation Sequencing (NGS) is revealing new insights into microbial metagenomics; however, there exist three inherent biases that need to be addressed to provide a complete picture of microbial communities. First, the amount of starting material necessary for sequencing can be prohibitive for metagenomics. Second, most library preparation methods cause a bias in the data against high or low GC content sequences, skewing relative abundance of community members. Third, many library preparation kits only adapt double-stranded DNA, excluding any single-stranded phage or viruses present in the community from being detected. Here, we address these issues and present a method to capture the complete metagenome in one NGS library preparation.

To understand the impact of library preparation on GC bias in metagenomic sequence data, we compared library preparation kits on a simplified microbial community model comprised of microbes of differing GC content (19-70% GC). This comparison revealed that the library preparation method can skew the sequence data: some preparations have an inherent GC/AT bias in the genome fragments which are adapted into library molecules, causing organisms at the extremes to be underrepresented in the sequence data. Utilizing the library preparation method with the best relative abundance of all genomes in the artificial community, real mouse microbiome samples were diluted 100-fold (to ~100 pg), sequenced, and the relative species abundance compared to the undiluted samples. The diluted samples produced an equal representation to the undiluted samples, demonstrating that this library preparation method evenly captures the metagenome at extremely low levels of input.

Furthermore, one library preparation method, unique in the ability to adapt single-stranded DNA, enabled a previously unseen view of metagenomes: capturing not only dsDNA from bacteria and eukaryotes, but also the ssDNA from any phage or viruses present. The ability to sequence phage and virus together with other microbes from the same library preparation eliminates the need for separate viral particle extractions and ssDNA genome amplifications prior to sequencing.

These results demonstrate that the method used to prepare DNA libraries for sequencing must be carefully considered when designing NGS-based experiments.

THEORETICAL ANALYSIS INDICATES HUMAN GENOME IS NOT A BLUEPRINT BUT A STORAGE OF GENES, AND HUMAN OOCYTES HAVE AN INSTRUCTION.

Koichi Itoh

The Institute for Theoretical Molecular Biology, The Institute for Theoretical Molecular Biology, Ashiya, Japan

Is Human Genome really a blueprint? If it is not a blueprint, how are human bodies constructed? Firstly, this paper solves this proposition. I indicate 8 examples of important biological pathways and factors among house-keeping genes products and proved that human genome is not a blueprint. Genes of 8 examples are scattered at random in Human Genome. That is why Human Genome is a storage of genes. Secondly, I proved that human oocytes have an instruction for development and differentiation. In this case, I used opened public database for an expression profile of human oocytes. I selected 12700 genes which expressed in human oocytes. Among 12700 genes, more than 800 genes which are related to development and differentiation are expressed. Here I show that human genome is not a blueprint but a storage of genes, and human oocytes have the instructions. Human genome has been thought to be a blueprint, but what type of the blueprint has been a mystery1. Human genome project was over in 2003, and 12 years are already passed, but even the number of human genes still unknown. Analysis of human genomes has been continuously done, but the discussion which a human genome is a blueprint has not been done. Far from that, any traces of a blueprint are not found in human genomes. This may be an evidence that a human genome is not a blueprint. The Watson-Click's DNA double helix is very beautiful. Hence, we life-scientists have been imprinted that a human genome is a blueprint. If we hypothesize that a human genome is a blueprint, what types of absurdity do emerge? And if a human genome is not a blueprint, what must be needed to construct human bodies? To solve these propositions are the aim of this document. In the case of unicellular organisms such as E.coli, their genomes may play a role for blueprints. However, biological mechanisms of multicellular organisms such as Homo Sapiens, are much complex and it is difficult to contain all information as a blueprint in their genomes. Therefore, a human genome plays a role for storage of genes, and I think that human oocytes have the instructions and a fertilized egg selects necessary genes from the storage, and expresses genes for development and differentiation.

NON-CODING SOMATIC MUTATIONS AND REGULATORY VARIATION IN THE GLIOBLASTOMA GENOME

Yunyun Ni[1], Amelia W Hall[1], Anna Battenhouse[1], Max Shpak[2], Matthew C Cowperthwaite[2], Vishwanath R Iyer[1]

[1]University of Texas at Austin, Department of Molecular Biosciences, Center for Systems and Synthetic Biology, Austin, TX, [2]St David's Medical Center, NeuroTexas Institute, Austin, TX

Large-scale sequencing projects have generated extensive catalogs of somatic mutations in cancer genomes. Cancer driver mutations are most commonly analyzed in terms of how they affect the coding regions of genes, but the majority of somatic mutations occur in non-coding regions of the tumor genome. It is still largely unknown to what extent non-coding somatic mutations affect gene expression in cancers, and how their impact compares to the effect of other mechanisms of gene dysregulation such as copy number changes. Using whole-genome sequencing data from The Cancer Genome Atlas (TCGA), and by comparing tumor genomes to matched normal genomes, we identified non-coding somatic mutations across multiple cancer types including glioblastoma multiforme. We then used an expression quantitative trait loci (eQTL) approach to associate somatic genetic variants and copy number changes with transcript level variation across the same tumors. We adjusted the genotype dosage according to the read coverage at each allele to obtain a more accurate, copy-number-adjusted genotype for each somatic variant. Our copy-number-aware eQTL analysis revealed that copy-number alterations show much stronger associations with differential gene expression across tumors than non-coding somatic mutations. However, the non-coding mutations could frequently occur in potential regulatory elements including promoters and other cis-regulatory elements. Clusters of non-coding somatic mutations were also frequently detected, including at the previously known regulatory elements of TERT and other potential regulators like the polycomb complex oncogene BMI1. We also carried out epigenetic profiling to identify active cis-regulatory elements and long-range chromosomal interactions within a smaller number of primary glioblastoma tumors. We defined active enhancers that were clearly associated with genes involved in brain cancer-related pathways, and showed enrichment for transcription factor binding motifs that were not previously known to be regulators in glioblastoma. Our studies reveal considerable genetic and epigenetic heterogeneity across individual cancer genomes that is not well represented in standard genomic analyses of coding mutations and gene expression profiling in cancers.

AFFORDABLE PHASED GENOME REFERENCE SEQUENCES

David B Jaffe[1], Michael Talkowski[1,2], Neil I Weisenfeld[1]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]Center for Human Genetic Research, Massachusetts General Hospital and Harvard Medical School, Cambridge, MA

Historically, genome reference sequences have been the product of expensive, painstaking and often ad hoc projects. Typically homologous chromosomes were collapsed (through inbreeding or computationally). As loss of heterozygosity is generally deleterious, so biological inferences based on single-chromosome data could be misleading.

Here we propose an affordable method for generating affordable, phased genome reference sequences. Our method begins with a microgram of DNA. We make two libraries. The first library is made from 0.5 kb fragments and without PCR. The second library is made from 20-200 kb DNA fragments, which are compartmentalized and randomly amplified using a technology developed by 10X Genomics. Both libraries are deeply sequenced by Illumina.

Once the data are generated, we create an assembly graph from data of the first type using DISCOVAR *de novo*, then use the second data type to walk through the graph, yielding single-chromosome (phased) contigs, and in many cases resolving complex regions whose assembly has not previously been demonstrated using short or long reads, and in fact only using clone-by-clone sequencing.

We test drive this method on human genomes, taking advantage of data generated very recently by 10X. The method should be applicable to a broad range of genomes.

THE MUTATIONAL LANDSCAPE OF HUMAN ADULT STEM CELLS IN CULTURE

M. Jager[1], R. van Boxtel[1], V. Sasselli[1], F. Blokzijl[1], J. de Ligt[1], S. Boymans[1], A. Smouter[2], H. Begthel[1], J. Korving[1], M. Verheul[1], E. de Bruijn[1], P. Toonen[1], L. de la Fonteijne[1], H. Clevers[1], E. Cuppen[1,2]

[1]Hubrecht Institute for Developmental Biology and Stem Cell Research, Utrecht, Netherlands, [2]University Medical Center Utrecht, Medical genetics, Utrecht, Netherlands

Human adult stem cells from various tissues can be grown *in vitro* into transplantable organoids, representing a promising cellular source for regenerative medicine. Accumulation of mutations in the genome is a major risk for the use of stem cells in regenerative medicine, as it may cause cancer. Here, we systematically assessed the mutational load of cultured human intestinal stem cells by whole genome sequencing analysis.

We find that intestinal stem cells acquire ~4 base substitutions per day in culture. The *in vitro* mutation spectrum indicates increased cellular stress in the culture environment. Specifically, we observe a high contribution of C:G > A:T transversions, which can be indicative of oxidative stress. Culturing at physiological oxygen or in the presence of a glycolysis inhibitor, however, does not affect the mutational landscape. In addition, we observe an enrichment of mutations in late replicating DNA, suggesting increased replication stress. In line with this, we occasionally observe deletions at common fragile sites, which is another marker for this type of cellular stress. Importantly, mutations are strongly depleted in transcriptionally active regions, resulting in less than 2 novel base substitutions within protein coding DNA after 5 months of culturing. Our results suggest that functionally important genomic regions in cultured adult stem cells are protected against mutagenic damage accumulation and/or are actively repaired. Moreover, our findings may serve as a basis to further improve stem cell culture conditions.

ASSEMBLY AND ANALYSIS OF 200 COMPLETE HLA HAPLOTYPES

Jacob M Jensen[1,2], The Danish Pangenome Consortium[1,2,3,4], Simon Rasmussen[3], Siyang Liu[4], Palle Villesen[1,2], Mikkel H Schierup[1,2]

[1]Aarhus University, Bioinformatics Research Centre (BiRC), Aarhus, Denmark, [2]Aarhus University, Centre for Integrative Sequencing, iSEQ, Aarhus, Denmark, [3]Technical University of Denmark, Center for Biological Sequence Analysis, Department of Systems Biology, Kgs. Lyngby, Denmark, [4]University of Copenhagen, Department of Biology, Copenhagen, Denmark

The human leukocyte antigen (HLA) covers ~4Mb on chromosome 6 and is the most polymorphic region of the human genome. It harbors more than 200 genes including the classical HLA loci which are under balancing selection and important for self-recognition. The HLA region has a large number of disease associations, particularly for autoimmune diseases. Pinpointing the causes of these associations is hampered by the structural complexity of the HLA, which is poorly characterized with only 8 full haplotypes annotated.
Here we present 200 fully assembled and annotated HLA haplotypes from the Danish pan genome project. In this project 50 trios were sequenced at 80x per individual with multiple insert size libraries up to 20 kb allowing the HLA region of each individual to be de novo assembled into one or a few scaffolds. We used the children of the trios to phase variation in order to determine each of the two HLA haplotypes of the parents. We use the 200 haplotypes of the parents to provide a comprehensive catalogue of structural variation of the HLA region. We demonstrate how this catalogue improves genotype imputation in the region, augmenting what can be gained from GWAS.

The Danish Pangenome Consortium: Anders D. Børglum, Anders Krogh, Arcadio Rubio-García, Christian N. S. Pedersen, David Flores, David Westergaard, Ditte Demontis, Emil Rydza, Esben Nørgaard Flindt, Francesco Lescai, Hans Eiberg, Hao Liu, Jacob Malte Jensen, Jakob Grove, Jette Bork-Jensen, Jihua Sun, John Damm Sørensen, José M. G. Izarzugaza, Jun Wang, Junhua Rao, Laurits Skov, Karsten Kristiansen, Kirstine Belling, Kristoffer Rapacki, Lars Bolund, Mikkel H. Schierup, Ning Li, Ole Lund, Oluf Pedersen, Ou Wang, Palle Villesen, Piotr Chmura, Piotr Dworzynski, Rachita Yadav, Ramneek Gupta, Ruiqi Xu, Rune M. Friborg, Shengting Li, Shujia Huang, Simon Rasmussen, Siyang Liu, Søren Besenbacher, Søren Brunak, Thomas D. Als, Thomas Mailund, Thorkild I. A. Sørensen, Torben Hansen, Weijian Ye, Xiaofang Cheng, Xun Xu et al.

COMPARATIVE GENOMIC ANALYSIS REVEALS THE EVOLUTIONARY DYNAMICS OF NRSF BINDING ACROSS FOUR MAMMALIAN SPECIES

Shan (Mandy) Jiang[1,2], Ricardo Ramirez[1,2], Nicole El-Ali[1,2], Ali Mortazavi[1,2]

[1]University of California, Irvine, Developmental and Cell Biology, Irvine, CA, [2]University of California, Irvine, Center for Complex Biological Systems, Irvine, CA

The evolution of the binding repertoire of transcription factors is one of the key questions of comparative genomics. The transcription factor NRSF (Neuron-Restrictive Silencer Factor, also known as REST) represses many vertebrate neuronal genes in non-neuronal cells through binding to a few thousand canonical 21bp NRSEs (also known as RE1) as well as various non-canonical sites. We conducted a comparative genomic analysis using ChIP-seq for NRSF across four mammalian species (human, mouse, dog and horse) in order to understand the evolution of the NRSF binding repertoire. We found that non-canonical motifs convert to canonical motifs at a higher rate rather than in the reverse direction. We also found 142 species-specific NRSE births as well as 69 species-specific site deaths. We are now in the process of categorizing the mechanisms leading to NRSF binding sites turnover and their relation to pathway level repression of many neuronal processes by this factor.

TETRANSCRIPTS: A PACKAGE FOR INCLUDING TRANSPOSABLE ELEMENTS IN DIFFERENTIAL EXPRESSION ANALYSIS OF RNA-SEQ DATASETS

Ying Jin, Oliver H Tam, Eric Paniagua, Molly Hammell

Cold Spring Harbor Laboratory, Cancer, Cold Spring Harbor, NY

Most RNA-seq analysis software packages are not designed to handle the complexities involved in properly apportioning short sequencing reads to highly repetitive regions of the genome. These regions are often occupied by transposable elements (TEs), which make up between 20 – 80% of eukaryotic genomes. Transposable elements are mobile DNA elements that propagate by multiplying within the genomes of host cells, and can be passed from generation to generation through the germline lineage. Often thought of as "junk" transcripts with little importance for biological phenotypes, TEs can play a large and unexpected role in important processes such as stem cell identity and reprogramming, embryogenesis, neural develpment, aging, and in human diseases such as cancer and neurodegenerative diseases. While TE derived transcripts should be included as part of standard expression analyses, TE-associated reads are often discarded in sequencing data analyses because of the uncertainty in attributing ambiguously mapped reads to these regions. Here, we present a program called *TEtranscripts* that allows users to analyze both gene- and TE-associated reads concurrently in one simplified workflow.

Using simulated reads as well as published datasets that include independent validation, such as qPCR and NanoString, we have shown that *TEtranscripts* outperforms all other published methods in abundance estimation, and concordance between statistical significance estimation and validated alterations in expression. In simulated datasets, we show that *TEtranscripts* performs particularly well at estimating the abundance of young TEs, which are more likely to be mobile and active in cells. In published datasets for both fly and mouse genomes, we show that alterations in TE expression estimated from RNA-seq data by *TEtranscripts* show better overall concordance with external validation data. *TEtranscripts* particularly outperforms other methods for complex mammalian genomes, such as the mouse, which has many more insertions per TE than flies, and a larger diversity in TE families.

Availability: The source code and associated GTF files for TE annotation are freely available at
http://hammelllab.labsites.cshl.edu/software

SPLADDER: INTEGRATED QUANTIFICATION, VISUALIZATION AND DIFFERENTIAL ANALYSIS OF ALTERNATIVE SPLICING

Andre Kahles[1], Cheng Soon Ong[2], Gunnar Rätsch[1]

[1]Memorial Sloan Kettering Cancer Center, Computational Biology, New York, NY, [2]NICTA, Canberra Research Laboratory, Canberra, Australia

Understanding the occurrence and regulation of alternative splicing (AS) is a key task towards explaining the regulatory processes that help shape the complex transcriptomes of higher eukaryotes and that ensure the necessary flexibility in expression from a single locus that is vital for development and regulation. With the advent of high-throughput sequencing of RNA (RNA-Seq) this diversity can be measured at an unprecedented depth. Although the catalog of known AS events has ever grown since, novel isoforms are commonly observed in less well annotated organism, in the context of disease, or within large populations. Whereas an identification of complete isoforms is technically challenging and expensive, focusing on single alternative splicing events as an alternate way to characterize transcriptional diversity is fruitful for differential analysis.

We present SplAdder, a fully integrated analysis framework, that can detect both known and novel AS-events from RNA-Seq alignments, quantify them and differentially test them between given sample sets. AS-events are detected from a given annotation or can be added to it based on the given RNA-Seq evidence. The streamlined, highly efficient and parallelizable pipeline quantifies all events and provides counts as an interface to differential analysis with common tools such as the rDiff package or DExSeq. SplAdder further includes several visualization routines, producing publication ready plots of the quantified splicing graph, displaying one or many events or showing sashimi-like plots for different isoforms.

SplAdder can easily handle several thousand samples of high complexity and has been developed and tested on data from The Cancer Genome Atlas project and the International Consortium of Cancer Genomics. However, SplAdder is not limited to human and we demonstrate applications in the plant A. thaliana as well as the nematode C. elegans. The software is implemented in Python and is available as open source software at www.github.com/ratschlab/spladder. For more information visit www.bioweb.me/spladder.

DECIPHERING FUNCTIONAL MECHANISMS FOR NON-CODING GENETIC VARIANTS ASSOCIATED WITH COMPLEX TRAITS

Cynthia Kalita, Greg Moyerbrailean, Chris Harvey, Roger Pique-Regi, Francesca Luca

Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI

GWAS (Genome wide association study) has identified thousands of regions associated with complex traits. However,it is very challenging to identify the causal genetic variant within a region and characterize its underlying mechanism. Many GWAS signals are in non-coding regions, and may disrupt gene regulatory sequences such as transcription factor (TF) binding sites. Transcription factor activity can be modulated by the cellular environment, where gene-envrioment interactionsmay contribute in explaining a large fraction of unaccounted complex trait variation across individuals.
We have performed fine-mapping of GWAS variants, and followed up with experimental validation with allele specific reporter assays in specific environmental conditions. We focused on functional annotations we previously developed by integrating binding sites predicted by a motif model with DNase I footprinting data. Using an empirical Bayesian framework implemented in the fgwas software, we combined data from GWAS studies with these functional annotations. We observed improved posterior probability of association and increased interpretability of the GWAS signals, as compared to other annotations (e.g. distance to the TSS). Among the variants in enriched motifs active in LCLs, we selected seven from GWAS for LDL, height, mean red cell volume, triglycerides, platelet count, mean cell hemoglobin, and total cholesterol. We show that four of these variants drive allele-specific gene expression ($p<0.05$) at baseline. We then used a relevant panel of treatments that induce gene expression changes in LCLs (dexamethasone, retinoic acid, mono-n-butyl phthalate, copper chloride, sodium selenite, cadmium, perfluorooctanoic acid, and cetirizine). We show that rs12718597, associated with mean red cell volume, and rs2336384 associated with platelet count, induce allele specific expression only in the presence of cetirizine, a common antihistamine allergy medication. rs12718597 is an eQTL for FIGNL1 in the GTEX data and is predicted to alter binding of ATF. rs2336384 is predicted to alter binding of CREB and is an eQTL for MFN2 (GTEX data), a gene differentially expressed in response to cetirizine. In conclusion, an approach that considers gene-environment interactions when combining GWAS signals with functional genomics data, can provide a much better understanding of the molecular mechanisms underlying interindividual variation in complex traits.

LOFTEE: IMPROVING THE DISCOVERY OF PROTEIN-TRUNCATING VARIANTS IN HUMAN GENES

Konrad J Karczewski[1,2], Monkol Lek[1,2], Kaitlin Samocha[1,2], Daniel Birnbaum[1,2], Mark J Daly`[1,2], Daniel G MacArthur[1,2]

[1]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [2]Broad Institute, Medical and Population Genetics, Cambridge, MA

A typical human genome carries at least a hundred variants predicted to cause complete loss-of-function (LoF) of protein-coding transcripts, with approximately twenty observed in the homozygous state. These "experiments of nature" represent unique opportunities to study gene function as human knockouts, and have yielded insights in the identification of potential drug targets. However, discovering LoFs in the human population remains a significant challenge, as these variants are enriched for sequencing and annotation errors, and typically have very low frequency, confounding their discovery and interpretation. We present LOFTEE (the Loss-Of-Function Transcript Effect Estimator), a software package for identifying and filtering loss-of-function variants from whole exome or whole genome sequence data. LOFTEE uses sequence context to prioritize loss-of-function variants and incorporates tissue-specific annotation, using transcript expression patterns derived from GTEx. We validate the LOFTEE filtering approach using the frequency spectrum from over 60,000 exomes from the Exome Aggregation Consortium (ExAC), from analysis of known disease-causing LoF variants, and comparison against multi-tissue RNA sequencing data from over 500 individuals generated as part of the GTEx project. These analyses reveal the pervasive extent of gene inactivation in humans, and reveal features that can be used to distinguish genuine LoF variants that may point to candidate targets for pharmaceutical inhibition.

GENOME-WIDE IDENTIFICATION OF ENHANCERS AT HIGH RESOLUTION IN DROSOPHILA S2 CELLS SUGGESTS THE EXISTENCE OF FUNCTIONAL ENHANCER CORES

Tomas Kazmar[1,2], Cosmas D Arnold[1], Christoph Stelzer[1], Michaela Pagani[1], Martina Rath[1], Alexander Stark[1]

[1]Research Institute of Molecular Pathology (IMP), Stark Group, Vienna, Austria, [2]Institute of Science and Technology Austria (ISTA), Lampert Group, Vienna, Austria

Regulatory DNA sequences, called enhancers, are responsible for defining gene expression levels in a cell type-specific manner. Recently developed methods allow the identification and functional characterization of enhancers in a genome-wide manner, providing sufficiently large and homogeneous sets of enhancers for systematic sequence analyses to uncover how enhancer sequences are organized.

In this work, we delineate enhancers in Drosophila S2 cells at a high resolution using STARR-seq with short enhancer candidate libraries (~150bp compared to 500bp). This reveals enhancer cores that can be sufficient for activity and seem to be key to enhancer function. We find different motif enrichments in these cores compared to the flanking sequences, including motifs known to be important for S2 enhancers in general. The cores are also C/G rich and are flanked by sequences rich in AA/TA dinucleotides reminiscent of nucleosome positioning signals and nucleosomes are indeed predicted to be positioned over the cores. However, in contrast to the prediction, experimentally determined nucleosome occupancy was low at the cores of enhancers that are active in S2 cells. Interestingly though, it was high as predicted in cell types, in which the enhancers are not active.

This suggests an attractive model in which enhancers have a default OFF state with nucleosome-bound enhancer cores and that this nucleosome is displaced by transcription factors when the enhancer becomes activated.

EPIGENOMICS OF COMMON, RARE, AND SOMATIC VARIANTS UNDERLYING DISEASE AND CANCER.

Manolis Kellis[1,2], Gerald Quon[1,2], Melina Claussnitzer[1,2,3], Xinchen Wang[1,2,4], Laurie Boyer[1,4], Richard Sallari[1,2], Roadmap Epigenomics[1,2]

[1]MIT, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Broad Institute, of MIT and Harvard, Cambridge, MA, [3]Harvard, Medical School, Boston, MA, [4]MIT, Biology Department, Cambridge, MA

Perhaps the greatest surprise of genetic studies of human disease is that 90% of top-scoring disease-associated loci lie outside protein-coding regions. This has increased the urgency of mapping non-coding DNA elements and regulatory circuits, in order to understand the molecular basis of human disease. To address this challenge, the Roadmap Epigenomics program has sought to systematically characterize the epigenomic landscape in diverse primary human cells and tissues, resulting in the annotation of 2.3M enhancer elements across 127 tissues and cell types, and tissue-specific regulatory networks linking enhancers to their upstream regulators and target genes. In this talk, we describe the use of these annotations for understanding the molecular basis of genetic differences in common disease and cancer.

First, we use these annotations for interpreting and fine-mapping causal variants from GWAS. We present a new network for identifying master regulators of disease variants, using an expectation maximization framework that simultaneously refines the set of predicted causal variants, the cell types in which they act, and the upstream regulators that target them, even when their motifs are not directly disrupted. We apply this framework in the context of obesity and Alzheimer's disease, resulting in unexpected tissues and regulators, which we validate experimentally by directed perturbation in human primary cells and in mouse, and by studying epigenomic alterations between individuals.

Second, we use these annotations to prioritize weakly-associated variants which we validate experimentally. We develop a machine learning framework that learns combinations of features that define top-scoring variants, and use it to prioritize below-threshold variants. We experimentally validate below-threshold enhancers, their allelic activity, 4C links to target genes, and their cellular and organismal phenotypes in the context of cardiac repolarization phenotypes.

Third, we use our networks to identify new cancer genes based on recurrent somatic mutations in their upstream regulatory elements. We find that cancer-dysregulated genes show an excess of non-coding mutations, and are frequently upregulated by co-option of non-coding elements that are active in other cell types. We predict 16 new prostate cancer driver genes with roles in immune evasion, insulin-androgen signaling, and energy regulation.

COMPUTATIONAL IDENTIFICATION OF NONCODING CANCER DRIVERS FROM WHOLE-GENOME SEQUENCING DATA

Ekta Khurana

Weill Cornell Medical College, Physiology and Biophysics, New York, NY

Whole-genome sequencing of tumors provides an opportunity to investigate sequence variants (including single nucleotide variants and large genomic rearrangements) in noncoding regions. We developed a computational framework to annotate and prioritize cancer regulatory mutations. The framework combines an adjustable data context summarizing large-scale genomics and cancer-relevant datasets with an efficient variant prioritization pipeline. To identify driver elements, we first developed a weighted scoring scheme to score each mutation's impact through analyzing conservation, loss-of and gain-of function events, gene associations and network topology. Then we integrated the scores for variants and their recurrence across multiple tumor samples to develop a functional score for each element. Using this scheme, we identified candidate noncoding drivers in ~1000 samples from ~10 different cancer types. Next, we integrated the functional scores for coding and noncoding regions (promoters, enhancers and ncRNAs) for each gene to explore the fraction of gene expression changes in tumor tissue that can be explained by noncoding mutations vs. those in coding genes themselves. Besides sequence variants in genes and their regulatory elements, we also analyzed the effects of transcription factor rearrangements on the expression of their target genes.

GENE EXPRESSION VARIATION IN HUMAN INDUCED PLURIPOTENT STEM CELLS

Helena Kilpinen[1], Angela Goncalves[2], Dalila Bensaddek[3], Francesco P Casale[1], Daniel Gaffney[2], Angus I Lamond[3], Oliver Stegle[1], on behalf of the HipSci Consortium[1,2,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom, [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, [3]University of Dundee, Centre for Gene Regulation and Expression, Dundee, United Kingdom

Induced pluripotent stem cells (iPSC) are increasingly used to model functional effects of human disease alleles. However, large-scale studies addressing their heterogeneity are lacking. We studied gene expression variability in human iPSC lines from 202 healthy donors, analyzing in total 592 iPSC lines and 202 fibroblast lines from which the iPSCs were derived. We sought to understand the sources of biological variability and heterogeneity in these lines in the context of genetic changes.

Clustering of the iPSC lines based on gene expression profiles (20,273 probes genome-wide) confirmed their pluripotency compared to control cells. Variance component analysis of gene expression levels revealed strong donor effects, whereas the contribution of various technical aspects of the reprogramming process, such as culture media and passage rate, were small.

Mapping of expression quantitative trait loci (eQTL) in cis revealed 2991 and 1423 genes with an eQTL in the iPSC and fibroblast state, respectively (FDR < 0.01 per gene). 63% of the fibroblast eQTL genes were eQTLs also in iPSCs (30% vice versa). Replication of the eQTL map using biological replicate lines of the iPSCs showed very high concordance of eQTL effect sizes (ρ=0.97), suggesting that genetic effects on gene expression can be robustly identified when accounting for confounding experimental variability in reprogramming. Highest overlap of eQTL genes with other published datasets was seen with cell lines (lymphoblastoid, fibroblast) as opposed to tissues (GTex). Further, we identified 1492 genes with a QTL associated with expression variance (varQTLs; FDR < 0.01 per gene) in iPSCs, which were largely non-overlapping with the identified eQTL genes (15% of varQTL genes eQTLs in iPSCs, 14% in fibroblasts). These variants modulate the variability among the replicate iPSC lines and likely arise though interactions between genotypes and the cellular state (GxE).

Finally, LC-MS/MS proteomics data from a small number of donors confirmed the quantitative effect of a subset of the iPSC eQTLs on protein abundance. We are processing proteomics data from ~60 donors to evaluate the full extent to which gene expression variability in iPSCs manifests on the level of protein abundances. Further, cellular imaging phenotypes available within the HipSci resource will allow us to assess the effects of expression variability on the level of whole cells.

CENTRIFUGE: RAPID AND SENSITIVE CLASSIFICATION OF METAGENOMIC SEQUENCES

Daehwan Kim[1,2], Li Song[1,3], Steven L Salzberg[1,2,3]

[1]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [2]Johns Hopkins University, Biostatistics, Baltimore, MD, [3]Johns Hopkins University, Computer Science, Baltimore, MD

Centrifuge is a very rapid and memory-efficient system for the classification of DNA sequences from microbial samples, with better sensitivity and comparable accuracy than the competing software. To achieve this, we designed a novel indexing scheme based on BWT and FM-index and optimized it specifically to address the classification problem. Centrifuge requires a relatively small index (for example, 2.9 GB for ~2,800 bacterial genomes) yet provides a fast classification speed (able to handle a single run of DNA-sequencing reads within an hour). Together these enhancements enable timely and accurate analysis on conventional desktop computers, propelling performance to new levels of efficiency. Centrifuge is available as free, open-source software from http://www.ccb.jhu.edu/software/centrifuge.

TESTING THE GENOMIC ENRICHMENT OF COMMON AND RARE COPY NUMBER BURDEN ASSOCIATED WITH AUTISM

Dokyoon Kim[1], Anastasia Lucas[1], Ruowang Li[1], Alex T Frase[1], Santhosh Girirajan[1], Scott B Selleck[1], Marylyn D Ritchie[1,2]

[1]Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA, [2]Geisinger Health System, Biomedical and Translational Informatics, Danville, PA

Children with autism have a higher frequency of large, rare copy number variations (CNVs) compared with the general population. One conclusion from the extensive genetic analysis of autism thus far is that there are hundreds or perhaps thousands of genes or genomic regions that contribute to autism susceptibility. Thus, total genomic copy number burden, as an accumulation of copy number change, is a meaningful measure of genomic change that may contribute to autism susceptibility. Previously, our group found that autism is associated with increased levels of copy number burden. However, one of the current limitations of this approach is that it is difficult to interpret biological meaning based on the accumulation of copy number change genome-wide associated with autism. In this study, we develop a comprehensive and systematic pipeline for annotating copy number variants into genes/genomic regions and subsequently pathways and other gene groups using Biofilter – a bioinformatics tool that aggregates over a dozen publicly available databases of prior biological knowledge. Next we conduct enrichment tests of biologically defined groupings of CNVs including genes, pathways, Gene Ontology, or protein families. We applied the pipeline to a CNV dataset in the population-based, case-control Childhood Autism Risks fro Genetics and Environment (CHARGE) study. We found not only several significant genes associated with autism such as TRAJ1 ($p<0.01$) but also KEGG pathways of Olfactory Transduction ($p<0.05$) or Jak-STAT Signaling Pathway ($p<0.05$) and protein families of Ankyrin Repeats ($p<0.001$) or Olfactory Receptor ($p<0.05$). Based on the copy number burden analysis, it follows that the more and larger the copy number change, the more likely that one or more target genes will be affected that influence autism risk and phenotypic severity. Thus, our study suggests the proposed enrichment pipeline could improve the interpretability of copy number burden analysis where hundreds of loci or genes contribute toward disease susceptibility via biological knowledge groups such as pathways. This CNV annotation pipeline with Biofilter can be used for CNV data from any genotyping or sequencing platform and to explore CNV enrichment for any traits or phenotypes. Biofilter continues to be a powerful bioinformatics tool for annotating, filtering, and constructing biologically informed models for association analysis – now including copy number variants.

GRD: CURATED GENOMIC-BASED 16S RIBOSOMAL RNA GENE DATABASE

Seok-Won Kim[1,2], Kenshiro Oshima[2], Wataru Suda[2], Suguru Nishijima[2], Sangwan Kim[2], Todd D Taylor[1], Masahira Hattori[2]

[1]RIKEN Center for Integrative Medical Sciences, Laboratory for Integrated Bioinformatics, Yokohama, Japan, [2]The University of Tokyo, Center for Omics and Bioinformatics, Kashiwa, Japan

The 16S ribosomal RNA (16S rRNA) gene is found in all prokaryotes and is mainly used as a phylogenetic marker in environmental microbiota and human microbiome studies. The anti-Shine-Dalgarno sequence (anti-SD), which is the complementary sequence of the Shine-Dalgarno sequence in all mRNAs, is situated at the 3' end of the 16S rRNA gene. However, some complete genomes do not include anti-SD in their GenBank annotation. It is well known that there are some intervening sequences in the 16S rRNA gene, and the existence of introns within 16S rRNA has been reported. Various 16S rRNA databases such as RDP, SILVA, and Greengenes have been developed, and these databases are typically constructed on the basis of 16S sequences annotated in GenBank and PCR-based 16S rRNA studies. We developed a novel genomic-based 16S rRNA database called GRD. It is a highly-curated 16S rRNA gene database consisting of full-length 16S rRNA genes from both completely and partially sequenced bacterial genomes using various bioinformatic tools and manual curation. In this study, we show that the quality of annotation by GRD is more reliable than that by GenBank. Moreover, we determined that the current average copy number of 16S genes in prokaryotes is 3.58. By using GRD, we estimated the level of 16S rRNA sequence similarity necessary to determine the boundaries between taxonomic levels. This led to the identification of boundary identities between taxonomic levels for 16S rRNA genes, which is very useful and highly reliable for 16S-based taxonomical species assignment.

COMPUTATIONAL AND FUNCTIONAL ASSESSMENT OF NON-CODING MUTATIONS IN THE HUMAN GENOME

Martin Kircher[1], Fumitaka Inoue[2], Daniela Witten[3], Gregory Cooper[4], Nadav Ahituv[2], Jay Shendure[1]

[1]University of Washington, Dept. of Genome Sciences, Seattle, WA, [2]University of California (UCSF), Dept. of Bioengineering and Therapeutic Sciences, San Francisco, CA, [3]University of Washington, Dept. of Biostatistics, Seattle, WA, [4]HudsonAlpha Institute for Biotechnology, -, Huntsville, AL

The interpretation of variants identified by exome (WES) or genome (WGS) sequencing remains a paramount challenge for the field. While most variants discovered from WES can be immediately interpreted for their impact on a protein (e.g. missense, canonical splice, nonsense), the effects of non-coding mutations (e.g. non-canonical splice variants, variants in promoters, enhancers, non-coding transcripts or UTRs) – are much harder to interpret. When scaling to WGS, the vast majority of variants discovered fall into the latter category, and currently we have a very limited toolset for their interpretation.

While methods for editing genomes are becoming more widely available (e.g. CRISPR/Cas9), they are limited in throughput and require a prior molecular understanding of variant effects for developing a directed functional assay. It is therefore currently impossible to test thousands of rare variants identified from an individual genome. To address this, we pursue two higher-throughput approaches: (1) computational methods that uniformly score variants in the genome – inside and outside of coding regions – and (2) functional assays that test regulatory variants in a high-throughput and systematic fashion.

For the first goal, we previously developed Combined Annotation Dependent Depletion (CADD), a computational framework which allows us to score variants genome-wide using a broad set of annotations. We are continuing the development of CADD and will present the most recent updates (e.g. improved InDel scoring, an extended feature set and >150x faster model training). For the second goal, we are improving methods for massively parallel reporter assays (MPRAs) for regulatory sequences. Here, we aim to dissect effects of specific variants and to infer universal regulatory models from quantitative measurements of many promoter and enhancer variants as well as the systematic exploration of all possible sequence variants for selected elements. Such models will then be integrated in a CADD-like framework. We have enhanced MPRA to directly quantify effects from RNA tag-sequencing read-outs and transitioned from an episomal context to a lentiviral integration system, in which histone modifications and chromatin structure effects are taken into account. Preliminary experiments suggest that we can achieve 90% concordance of the quantitative measurements with less than 10 tags per insert for this system.

GENOME DATA AT NCBI–EASIER ACCESS, MORE FORMATS, IMPROVED PRESENTATION

Paul Kitts, Michael DiCuccio, Avi Kimchi, Terence Murphy, Kim Pruitt, Tatiana Tatusova

National Center for Biotechnology Information (NCBI), National Library of Medicine, NIH, Bethesda, MD

The National Center for Biotechnology Information (NCBI) databases contain data for over 30,000 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 format consistently 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 that help to differentiate between the multiple genome assemblies for a particular species.

GENETIC AND CLINICAL PREDICTORS OF CD4 RECOVERY DURING SUPPRESSIVE cART: WIHS

Ruth M Greenblatt[1], Kord M Kober[2], Peter Bacchetti[1], Ross Boylan[1], Kathyrn Anastos[3], Mardge Cohen[4], Mary A Young[5], Deborah Gustafson[6], Bradley Aouizerat[2,7]

[1]University of California San Francisco, Clinical Pharmacy, San Francisco, CA, [2]University of California San Francisco, Physiological Nursing, San Francisco, CA, [3]Yeshiva University, Montefiore Medical Center, Bronx, NY, [4]John H. Stroger Jr. Hospital of Cook County, Ruth Rotstein CORE Center, Chicago, IL, [5]Georgetown University, School of Medicine, Washington, DC, [6]State University of New York, Downstate Medical Center, New York, NY, [7]University of California, San Francisco, Institute for Human Genetics, San Francisco, CA

Background: Blood CD4 cell counts fail to recover in a significant minority of virologically suppressed cART recipients. We hypothesized that host genetics, including novel mutations, and key clinical characteristics would predict rapid versus slow recovery in women with HIV RNA levels below detection during cART.
Method: Longitudinal treatment response and clinical data were generated by the WIHS cohort, and used to define rapid vs. slow CD4 cell recovery among women during, at a minimum, the first 2.5 years of cART with virologic suppression. Whole exome sequencing (WES) was conducted on 95 women, with the most consistently slow (n=47) or rapid (n=48) CD4 recoveries. Additive stepwise logistic regression identified statistically significant predictors of rapid recovery.
Results: Each decade increase in age at the start of viral suppression while on cART reduced the odds ratio (OR) of rapid recovery by 0.50 (95% CI 0.27-0.92). Self-reported adherence ≥95% increased the OR of rapid recovery by 2.7 (CI 1.04-7.0). When added as a third predictor, higher CD4 nadir approached statistical significance (OR 1.34 per 100 cells, CI 0.96-1.87). Following WES analysis, rapid CD4 recovery was statistically significantly associated with sequence anomalies aggregated at the gene level for 68 genes (all $p<0.001$). Notably, the host genes identified were enriched for genes (n=14; 20.6%) that encode for proteins that interact with HIV-encoded proteins. An additional 38 genes (55.9%) encode for proteins that in turn interact with other host proteins known to interact with HIV-encoded proteins.
Conclusion: Despite consistent viral loads below detection on cART, self-reported nonadherence and higher age adversely influenced CD4 recovery. The finding that nonadherence and polymorphisms of HIV target genes influence CD4 recovery suggest that intermittent or low grade viral replication contributed to slow CD4 response in this sample of cART recipients with ≥2.5 years of virologic suppression.

GENOME-WIDE SIGNALS OF POSITIVE SELECTION IN STRONGYLOCENTROTID SEA URCHINS.

Kord M Kober[1,2], Grant H Pogson[1]

[\1]University of California, Ecology and Evolutionary Biology, Santa Cruz, CA, [2]University of California, Physiological Nursing, San Francisco, CA

Comparative genomics studies investigating the signals of positive selection among groups of closely related species are still rare and limited in taxonomic breadth. Here, we use the well-annotated genome of the purple sea urchin, *Strongylocentrotus purpuratus*, as a reference to investigate the signals of positive selection at 6,520 single-copy orthologs from nine sea urchin species belonging to the family Strongylocentrotidae. Applying a conservative false discovery rate of 5%, we identified 1,008 (15.5%) candidate positive selection genes (PSGs). Tests for positive selection along the nine terminal branches of the phylogeny identified 824 genes that showed lineage-specific adaptive diversification (1.67% of branch-sites tests performed). No differences were observed in the rates synonymous substitution (*dS*), GC content, and codon bias between the candidate PSGs and those not showing positive selection However, the candidate PSGs had ~68% higher rates of nonsynonymous substitution (*dN*) and ~33% lower levels of heterozygosity, consistent with continued action of selective sweeps and opposite to that expected by a relaxation of selective constraint. Although positive selection was identified at reproductive proteins and innate immunity genes, the strongest signals of adaptive diversification were observed at extracellular matrix proteins, cell adhesion molecules, membrane receptors, and ion channels. The observed patterns of positive selection showed limited associations with ecological variables such as temperature or depth. However, many of our candidate PSGs have been widely implicated as targets of pathogen binding, inactivation, mimicry, or exploitation. Our results suggest that pathogens might represent major drivers of positive selection in the marine environment but more studies are needed to link host-pathogen antagonistic coevolution at specific loci in sea urchins and to expand similar surveys to include other organismal groups.

QUANTITATIVE GENETICS OF GENE EXPRESSION DURING *DROSOPHILA MELANOGASTER* DEVELOPMENT

Enrico Cannavo*[1], Nils Kölling*[2], Dermot Harnett[1], Jacob Degner[1], David Garfield[1], Francesco P Casale[2], Hilary E Gustafson[1], Matt Davis[1], Oliver Stegle[2], Ewan Birney[2], Eileen E Furlong[1]

[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, [2]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom
* These authors contributed equally, listed alphabetically

The processes controlling embryonic development must work within two balancing constraints. To ensure individual viability, development must produce stereotypically patterned embryos in the face of environmental variation and segregating mutations. In spite of these constraints, species adapt through changing these robust developmental programs, showing their flexibility at this timescale. Mutations impacting developmental gene expression therefore play an important role in the evolution of novel phenotypes, and are thought to contribute to inter-individual differences in morphology, behaviour, and disease.

To better understand this dichotomy, we developed the first embryonic gene expression dataset in 80 genetically diverse inbred *Drosophila* lines from a single, natural population (DGRP). This allowed us to collect measurements at three different developmental stages for individuals with identical genotypes. With a novel linear mixed model framework that accounts for developmental stage and population structure, we uncovered extensive genetic variation underlying changes in expression levels (eQTLs) among the 80 lines. Surprisingly, although we found fewer eQTLs for genes known to be involved in development than for others, there were still developmental genes with relatively strong eQTLs. Furthermore we discovered a number of developmental stage specific eQTLs which identify cis-regulatory elements with time point specific activity. A major benefit of the DRGP is that the LD structure is short (on average $r2 < 0.2$ within tens of bp), which means that for around 50% of our eQTLs we can identify a single likely causal variant. This resolution allows us to explore the functional properties of these eQTLs. Many are near promoters, and associated with known transcription factor motifs, however we also see enrichment for eQTLs in known developmental enhancers at some distance from the promoter. We are also able to distinguish alternative polyadenylation sites, and a number of genes change UTR extent without a change in overall expression.

This study shows the robustness of development to genetic variation, the distribution of genetic variants underlying gene expression in a well-studied model organism, and the importance of understanding the developmental window of any specific genetic variant.

PUTTING THE W'S BACK INTO WHOLE-GENOME, WHOLE-TRANSCRIPTOME & WHOLE-EPIGENOME SEQUENCING

Jonas Korlach

Pacific Biosciences, 1380 Willow Road, Menlo Park, CA

Three requirements have to be met for obtaining a comprehensive view of an organism's genome biology: (i) sufficiently long sequence read lengths to resolve repeats, characterize structural variation, phase haplotypes, and identify gene products, (ii) uniform sequencing quality and depth, regardless of the DNA sequence context and complexity, and (iii) sensitivity to detect epigenetic DNA base modifications. Short-read next-generation sequencing technologies have been shown to exhibit deficiencies in these areas; as a consequence our views of genomes, transcriptomes and epigenomes have largely been incomplete and fragmented.

I will describe the most recent advances in long-read, single molecule, real-time (SMRT) DNA sequencing which overcomes these deficiencies. It allows for reference-grade de novo genome assemblies, in conjunction with new assembly algorithms that allow for a resolution of the diploid or polyploid nature of genomes. These new genomes show that many forms of structural variation have been missed until now. Full-length mRNA-seq is leading to the discovery of thousands of new isoforms and new genes, many of which play crucial roles in cellular differentiation and disease. Lastly, I will highlight how the genome-wide analysis of DNA modifications through SMRT sequencing has led to new insights into epigenetic mechanisms in both normal development as well as in the context of drug resistance in cancer.

GENETIC LANDSCAPE OF PRECLINICAL MODELS COMPARED TO PRIMARY TUMORS

Joshua M Korn, Hui Gao, Robert McDonald, Hans Bitter

Novartis Institutes for Biomedical Research, Oncology, Cambridge, MA

Cell lines and patient-derived xenograft (PDX) models are key preclinical models for drug discovery in oncology, and understanding the genetics of these models is critical to assess the translatability of preclinical results to the clinic. Here, we perform a systematic genetic comparison between cell lines, PDXs, and patient (TCGA) data. Consistently across the three data types, we observed a wide variation in mutation rates both between and within lineages (including, for example, a subset of copy-number-stable hypermutators in colorectal cancer) as well as enrichment of particular mutational patterns in certain indications, notably C-->A in lung and C-->T in melanoma. The median and standard deviation of the mutation rate per indication was highly correlated between cell lines, PDXs, and patient tumors, although there is some variability from indication to indication. We note the correlation was highest between PDXs and patient tumors (R=0.94 and R=0.96 for median and standard deviation, respectively; cell lines-tumors R=0.51 and R=0.72; cell lines-PDXs R=0.53 and R=0.60). The weaker correlation in cell lines is primarily due to a lack of hypermutators in cell line derivatives of melanoma tumors, although in general cell lines show more mutations per megabase than PDX models. For five indications—breast cancer, melanoma, colorectal cancer, non-small cell lung cancer, and pancreatic cancer—we compared the mutational and copy number landscape of in vivo and in vitro models to clinical samples both at the gene and pathway levels. The frequency of genetic alterations across the three datasets was remarkably consistent and revealed a relatively broad conservation of driver mutations found in human cancers in preclinical models, although we see somewhat greater correspondence between PDX models and clinical samples than between cell line models and clinical samples. For example in melanoma, there is an excess of MAPK mutations in cell lines (mostly by BRAF mutations), and a relative lack of MDM2 amplifications. However, for all indications we note a lack of certain subsets of driver mutations indicating preclinical models do not cover the complete diversity of patient tumors, and efforts could be made to fill in these gaps.

eQTL ANALYSIS OF MAIZE KERNELS TO DISCOVER FUNCTIONAL REGULATORY VARIATION

Karl A Kremling[1], Edward S Buckler[1,2,3]

[1]Cornell University, Department of Plant Breeding and Genetics, Ithaca, NY, [2]Cornell University, Institute for Genomic Diversity, Ithaca, NY, [3]United States Department of Agriculture, Agricultural Research Service, Ithaca, NY

To find functional variants from among tens of millions of SNPs in *Zea mays* (maize) I am using association mapping (GWAS) to determine the genetic variation that controls tens of thousands of intermediate expression phenotypes. Because natural phenotypic diversity is controlled by altering expression patterns in addition to changing CDS, these expression quantitative loci, or eQTL, are likely to point to true functional variants. These functional intergenic SNPs often go uncharacterized in genetic screens because their phenotypes are more nuanced than knockout mutations. However, by using expression values as quantitative traits in association studies, eQTL can be found on a genomic scale. Using a previously published set of 25.8 billion RNAseq reads (Fu et al 2013), a pipeline was developed to align reads and calculate expression values from the immature kernels of 368 diverse maize lines. After determining that the expression phenotypes had a median narrow sense heritability of 0.32, the expression values were used as phenotypes to conduct 39k GWAS experiments. In addition to calculating covariates (PCs) to account for population structure, hidden factors (HFs) were calculated from the matrix of 368 x 39,621 phenotypes to control for unmeasured sources of confounding. After accounting for the genetic and unmeasured experimental sources of structure using PCs and HFs, tens of thousands of significant cis and trans eQTL were found for thousands of genes. These eQTL point to functional variation in the form of local and distant regulators of expression.

SUPPORTED LIPID BILAYERS TO TURN GENOMIC SCIENCE INTO MATERIALS SCIENCE.

Sam Krerowicz[1,2,3], David C Schwartz[1,2,3], Mahesh Mahanthappa[1]

[1]UW-Madison, Chemistry, Madison, WI, [2]UW-Madison, Genetics, Madison, WI, [3]UW-Madison, Biotechnology Center, Madison, WI

The engineering mentality that guides today's thinking in the construction of DNA nano-structures and materials is heavily dependent on DNA design principles set down over the past few decades by Ned Seeman, Erik Winfree and Paul Rothemund. Such thinking has culminated in the development of "DNA origami," which makes heavy use of long, single-stranded DNA molecules. Using the bioinformatic tools and molecular modalities developed by LMCG we are developing modern approaches to the construction of microscale objects made of DNA that also offer novel routes to dynamic action and control through presentation on supported lipid bilayers.

A DYNAMIC FRAMEWORK FOR METABOLIC ENGINEERING OF THE BRANCHED-CHAIN AMINO ACID BIOSYNTHESIS PATHWAY IN *ESCHERICHIA COLI*

Anna S Kropornicka[1], Devesh Bhimsaria[2], Jennifer Reed[3], Aseem Z Ansari[4,5]

[1]University of Wisconsin-Madison, Genetics, Madison, WI, [2]University of Wisconsin-Madison, Electrical Engineering, Madison, WI, [3]University of Wisconsin-Madison, Chemical and Biological Engineering, Madison, WI, [4]University of Wisconsin-Madison, Biochemistry, Madison, WI, [5]University of Wisconsin-Madison, Genome Center of Wisconsin, Madison, WI

Metabolic engineering, in combination with synthetic biology, has been used to produce a number of valuable resources, including biofuels and pharmaceutical precursors. The goal of metabolic engineering is to harness an organism's metabolism to create a product of interest, which is generally at odds with the cell's primary objective of maximizing biomass production. In the valine biosynthesis pathway, the overproduction of valine indirectly causes a toxic byproduct to accumulate in *Escherichia coli* cells. This opposition makes it extremely difficult for metabolic engineers to manipulate these systems. To bypass the cell's opposition, computational methods are used to predict where regulatory perturbations should be made to maximize the production of a desired molecule and simultaneously optimize cell growth. The modular design principles of natural transcription factors can be harnessed to create artificial transcription factors (ATFs), such as Transcription Activator-Like Effectors (TALEs), that target any specified sequence and perturb metabolic networks with temporal control. Combining *in silico* modeling with a design amenable to high-throughput production of TALEs provides a dynamic method of regulating metabolic networks. As a proof of principle, we have showed that the valine biosynthesis pathway could be optimized with rationally engineered TALEs. Our ATFs have proven to be specific and effective repressors.

THE X EFFECT: REGULATORY VARIATION BETWEEN THE SEXES

Kimberly Kukurba[1,2], Princy Parsana[3], Kevin Smith[2], Zach Zappala[1], Anshul Kundaje[1], Alexis Battle[3], Stephen Montgomery[1,2]

[1]Stanford University, Genetics, Stanford, CA, [2]Stanford University, Pathology, Stanford, CA, [3]John Hopkins University, Computer Science, Baltimore, MD

The unique mode of inheritance and regulatory mechanisms of the X chromosome have resulted in special patterns of evolution that shape its genetic architecture and contribute to differences between males and females. However, despite the influence of the X chromosome and genotype-sex interactions on phenotype, they are largely excluded from genetic studies. We investigated the impact of sex and the X chromosome on regulatory variation using RNA-sequencing from 922 genotyped individuals in the Depression Genes and Networks (DGN) study, currently the largest available cohort with RNA-seq from primary tissue. First, we investigated the contributions of sex to expression variance across the transcriptome. We observed that females have higher variance in expression on the autosomes. Conversely, males exhibit higher expression variance on the X compared to females, a pattern expected in males whose hemizygous state results in more extreme effects and we demonstrate arises from cis genetic variation. Second, we identified cis expression quantitative trait loci (eQTLs) on the autosomes and the X. We identify eQTLs affecting the large majority (74.8% at FDR 0.05) of autosomal genes with quantifiable expression, but fewer eQTLs affecting X-chr genes (43.7% at FDR 0.05). Additionally, eQTL effect sizes on the X are significantly lower than those on the autosomes, especially among genes with strong purifying selection. Similarly, we observe a depletion of splicing QTLs (sQTLs) on the X. These results suggest that cis-regulatory variants on the X are removed from the population faster, consistent with theoretical models indicating that the X undergoes more efficient selection. Third, we investigated genotype-sex interactions and observed that despite selective pressures, the X contains a higher proportion of sex-specific eQTLs compared to the autosomes. Functional enrichment analyses of genes with sex-specific eQTLs suggest that they are enriched in reproductive structure development, hormone response, and regulation of cell death. To further elucidate the functional mechanisms underlying the effect of sex on genetic regulation of gene expression, we detected sex-specific open chromatin regions using the assay for transposase-accessible chromatin followed by sequencing (ATAC-seq, N=20), which allows us to identify the transcription factors and causal mechanisms driving sex-specific eQTLs. Together, this work advances our understanding of how the X and sex-interactions shape human gene regulation and highlights their importance in genetic studies.

A PANEL OF NOVEL STATISTICAL TESTS IDENTIFIES TUMOR SUPPRESSORS AND ONCOGENES FROM PAN-CANCER GENOME SEQUENCING DATA

Runjun D Kumar[1,2], Adam C Searleman[1,2], S. Joshua Swamidass[3], Obi L Griffith[4], Ron Bose[1]

[1]Washington University School of Medicine, Oncology, St. Louis, MO,
[2]Washington University in St. Louis, MD-PhD Program, St. Louis, MO,
[3]Washington University School of Medicine, Pathology, St. Louis, MO,
[4]Washington University in St. Louis, The Genome Institute, St. Louis, MO

One application of cancer genome sequencing data is identifying which human genes can act as cancer genes by promoting tumor formation and growth. This problem consists of two distinct tasks. First, cancer genes must be identified among genes bearing only passenger mutations. Several previously published methods address this problem by detecting genes with elevated mutation rates, or by detecting nonrandom patterns of mutations within genes. The second task is to separate putative cancer genes into likely oncogenes or tumor suppressors (TSG). Few methods exist for this problem. Both tasks are crucial as downstream experiments and therapeutics will require high quality mechanistic predictions.
In this study, we used a pan-cancer dataset of 1.7 million cancer mutations and a panel of 99 manually curated cancer genes to design five novel statistical tests for identifying and separating cancer genes. First, we developed patient and cancer type bias as new approaches for identifying cancer genes. Statistical tests based on these signals isolate known cancer genes (oncogenes and TSGs) better than existing methods with an Area-Under-Receiver-Operator-Characteristic (AUC) of 0.90, compared with AUCs of 0.76 to 0.82 for existing methods (MutSigCV, Oncodrive-fm, OncodriveCLUST). Furthermore, our tests based on mutation clustering and functional impact bias isolated cancer genes as well or better as existing methods. In addition, our statistical test based on the rate of truncating mutations was highly effective at separating known oncogenes and TSGs (AUC=0.92). Finally, we integrated these tests into a random forest model which identified cancer genes and classified them as TSGs or oncogenes simultaneously without loss of performance. We validated our statistical tests and random forest using several independent gene panels. Existing methods performed quite well at isolating TSGs (up to AUC=0.85 against the TSGene panel), but performed less well with oncogenes (up to AUC=0.63 against the KinDriver panel). Individually or in combination, our new methods were as good or better at isolating TSGs (up to AUC=0.88), and provided substantial improvements in isolating oncogenes (up to AUC=0.80). These new statistical tests suggest many new tumor suppressors and oncogenes, allowing for design of targeted downstream analyses and experiments.

TUMOR-NORMAL GENOME ANALYSIS VIA PERSONALIZED GRAPH REFERENCES

Deniz Kural, Kate Blair, Brandi Davis-Dusenbery, Wan-Ping Lee, Vladimir Semenyuk

Seven Bridges Genomics, Dept of Genome Biology, Cambridge, MA

We construct a personalized graph reference and utilize a cancer graph genomics mapping & determination pipeline for more accurate tumor genome reconstruction.

Initially, the normal genome is mapped against a population graph reference and fully determined. Then, the normal genome is incorporated into the population graph to construct a personal reference genome graph. This personal reference genome is then used as the basis for mapping the tumor genome and constructing the Primary Tumor Graph. The Primary Tumor Graph includes all minor allele frequencies, even at low trace amounts, and phase information inherent in the read data. It will thus indicate the sub-clonal structure, and explicitly represent the tumor as what it truly is - an evolving population of cells.

Thus, are able able to trace the evolution of the tumor over time, as allele frequencies change in response to treatment, and will be able to sequence metastatic or circulating tumor cells and incorporate additional branches to the Tumor Graph. This will enable an accurate characterization of both the initial tumor data and the clonal evolution of the tumor, resulting in more accurate molecular histories of cancer evolution, ultimately leading to better diagnosis and treatment.

VARIABLE LYMPHOCYTE RECEPTOR-BASED GLYCOPROTEOMICS OF THE BLOOD-BRAIN BARRIER

Jason M Lajoie[1], Brantley R Herrin[2], Eric V Shusta[1]

[1]University of Wisconsin-Madison, Chemical and Biological Engineering, Madison, WI, [2]Emory University, Pathology and Laboratory Medicine, Atlanta, GA

The blood-brain barrier (BBB) is comprised of specialized endothelial cells (ECs) lining the cerebral vasculature physically separating the blood from the brain. Unlike the vascular endothelium in other organs which contain fenestrae, ECs at the BBB are sealed together by continuous tight junctions that eliminate free diffusion of water soluble molecules into the brain. As a result, the BBB endothelia are tasked with actively regulating the bi-directional transport of nutrients and proteins between the blood and the brain and therefore have a unique phenotype. Many of the proteins responsible for this unique phenotype are plasma membrane (PM) proteins. In addition, BBB PM proteins are often glycosylated and their glycosylation status contributes to BBB physiology in health and disease. Currently the coverage of BBB PM proteins is relatively low given the difficulties of working with these typically hydrophobic, low-abundance proteins in standard proteomics workflows. Furthermore, glycomics of the BBB has been thus far restricted to lectin profiling which likely underrepresents the diversity of BBB glycan expression. Therefore, there remains a substantial need to further profile the BBB PM glycoproteome to gain insight into BBB development, maintenance, and disease.

Here we seek to address this need via creation and screening of an immune antibody library raised against mouse BBB PM proteins. Instead of using traditional mammalian antibodies, we have chosen to employ a recently discovered class of diverse antigen recognition molecules from lamprey called variable lymphocyte receptors (VLRs). Fresh PM protein mixtures were prepared from mouse brain vasculature and used to immunize larval lampreys in order to elicit an antigen-specific VLR immune response. Initial immunohistochemistry on mouse brain sections and glycoarray analysis using serum from immunized lampreys revealed the production of VLR that selectively bind to brain vasculature and label a unique cohort of glycans. Subsequently, a yeast surface display library, termed BBBVLR, was created via subcloning of the elicited VLR response from lymphocytes recovered from the immunized lampreys. High throughput screening of the BBBVLR library via consecutive rounds of magnetic and fluorescence activated cell sorting was employed to enrich the library for VLRs binding to proteins and glycans from fresh mouse BBB PM preparations. Mining of the enriched BBBVLR library is ongoing and should yield a diverse cohort of interesting BBB binding monoclonal VLRs that will be used as novel tools to profile the BBB PM glycoproteome.

TOWARDS UNDERSTANDING THE GENOMIC ARCHITECTURE OF CANCER GENOMES

Ernest T Lam[1], Alex R Hastie[1], Marcin B Imielinski[2], Cheng-Zhong Zhang[2], Jeremiah Wala[2], Zeljko Dzakula[1], Han Cao[1]

[1]BioNano Genomics, Research and Development, San Diego, CA, [2]Broad Institute of Harvard and MIT, Center for Biomedical Informatics, Cambridge, MA

Understanding the genetic architecture of cancer requires whole-genome and integrative approaches. Cancers often feature genomic alterations that range from single-base changes to large-scale structural variation (SV) involving gains and losses, and rearrangement of DNA content. Having a complete catalogue of mutations in cancer is crucial for identifying key drivers and providing accurate diagnosis, prognosis and targeted therapy. Whole-genome sequencing has become more routine and affordable since the introduction of next-generation sequencing (NGS) technologies. However, NGS platforms have limited power to decipher large, complex structural variants frequently observed in cancer. Genome mapping represents a complementary technology that provides critical structural information. It involves high throughput analysis of single molecules spanning hundreds of kilobases in nanochannels. Long-range information is preserved and direct interrogation of complex structural variants made possible. Therefore, leveraging the strengths of these complementary platforms would give a comprehensive view of a cancer genome.

Here, we present our analysis of well-studied and highly rearranged cancer genomes such as the near-tetraploid HCC1143 cell line. We constructed completely *de novo* genome map assemblies with N50 lengths of more than 1 Mb. We derived multi-sample normalized copy number profiles of matched tumor-control pairs based on genome mapping data. We observed that tumor samples had highly variable copy number profiles, corresponding to focal and chromosome-scale changes. Copy number breakpoints were shown indicative of translocation events. We also present a pipeline to integrate NGS and genome mapping data to validate and refine translocation calls. Genome mapping data helped bridge and phase neighboring translocation events. Finally, we present a computational approach to identify translocations by clustering single molecules with abnormal alignment to the reference and by performing local assemblies of these molecules. Overall, integrating NGS and genome mapping data provides a comprehensive view of a cancer genome.

ACCELERATING WRIGHT-FISHER SIMULATIONS ON THE GPU

David S Lawrie

University of Southern California, Molecular and Computational Biology, Los Angeles, CA

Many problems in computational biology can be classified as "embarrassingly parallel", consisting of a vast number of individual computations that are all independent of each other and thus capable of being performed concurrently. The rise of modern Graphics Processing Units (GPU) and programming languages designed to leverage the inherent parallel nature of these processors will allow researchers to dramatically speed up many programs that have high arithmetic intensity and intrinsic concurrency. Forward Wright-Fisher simulations are powerful in their ability to model complex demography and selection scenarios, but have historically been slow and required long compute times, thus limiting their usefulness. The Wright-Fisher forward algorithm is, however, exceedingly parallelizable, with many steps which are "embarrassingly parallel". In accelerating the parallelized code on the GPU, the presented Wright-Fisher forward simulation can be used to model arbitrary selection and demographic scenarios while running >130x faster than its serial, CPU counterpart. With simulations running at such speeds, one can do quick parametric bootstrapping of previously estimated parameters and even use simulated results to calculate the likelihoods and summary statistics of demographic and selection models against real polymorphism data for maximum-likelihood analysis or ABC – all without restricting the scenarios that can be modeled or requiring approximations to the forward algorithm for efficiency. Further, as this speedup can be achieved on even modest GPU hardware, investigators without access to high-end compute clusters will be able to accelerate the population genetics simulations needed for their research, democratizing computation in evolution.

A FLEXIBLE MIXED EFFECTS MODEL FRAMEWORK FOR DIFFERENTIAL DNA METHYLATION ANALYSIS

Amanda J Lea[1], Susan C Alberts[1,2], Jenny Tung[1,2,3,4], Xiang Zhou[5]

[1]Duke University, Biology, Durham, NC, [2]National Museums of Kenya, Institute of Primate Research, Nairobi, Kenya, [3]Duke University, Evolutionary Anthropology, Durham, NC, [4]Duke University, Duke University Population Research Institute, Durham, NC, [5]University of Michigan, Biostatistics, Ann Arbor, MI

Identifying sources of variation in DNA methylation levels is a key step in understanding gene regulation. High-throughput sequencing based approaches, such as whole genome bisulfite sequencing (methyl-seq) and reduced representation bisulfite sequencing (RRBS), are commonly used to obtain genome-wide, base-pair resolution estimates of DNA methylation levels. These methods quantify DNA methylation levels by comparing the number of methylated reads to the number of total reads at each CpG site. Because of the binary nature of these data, the most successful available analysis tools are based on beta binomial models. However, these methods do not account for sources of covariance among samples, such as genetic relatedness or population structure. Here, we present a flexible binomial mixed effects model that can be used to account for covariance between samples, and an accompanying efficient inference method–MACAU–for parameter estimation and hypothesis testing. Unlike other mixed effects programs commonly used for genomic datasets (e.g., EMMAX, GEMMA, FaST-LMM), MACAU works directly on count data and naturally takes variation in read depth among CpG sites and individuals into account. Using both simulations and real datasets, we show that our model exhibits better calibration of type I error in the presence of population structure than other existing approaches (e.g., beta-binomial models or linear mixed effects models after data transformation). This pattern is apparent in analyses of a publicly available Arabidopsis methyl-seq dataset, in which population structure is correlated with the predictor variable of interest. Further, we show that MACAU is better powered to detect truly differentially methylated sites. We applied MACAU to an RRBS dataset we generated from 50 baboons, and found that MACAU detected 2-fold more sites associated with age (at a 10% FDR) than a beta-binomial model (the next best approach). Importantly, these sites are enriched near genes previously shown to be differentially expressed with age in the same population, suggesting that our method accurately identifies age-associated CpG sites. Taken together, these results suggest that our binomial mixed effects approach and corresponding fitting method represent a significant improvement over current approaches, which either fail to account for the discrete nature of bisulfite sequencing data or fail to account for genetic relatedness.

A GRAPH-BASED FRAMEWORK FOR UNIFIED IDENTIFICATION OF SHORT AND STRUCTURAL GENETIC VARIANTS IN WHOLE-GENOME SEQUENCING DATA

Dillon H Lee, Alistair Ward, Gabor Marth

University of Utah School of Medicine, Department of Human Genetics, Salt Lake City, UT

The current practice of variant identification is largely restricted to the detection of short sequence variants, SNPs, and INDELs. Although several state-of-the-art, easy to use tools exist for short-variant detection (e.g. GATK, FREEBAYES, SNPTOOLS), these tools often produce divergent variant calls, especially INDELs, and it is very difficult to reconcile such variants into a single, accurate set. Furthermore, while it would be highly desirable to also detect larger, structural variants (SV), existing SV detector packages are typically difficult to integrate, highly resource-intensive to run, and resulting call sets require expert review to reduce false positive detection rate.

We present a highly innovative approach addressing these shortcomings, and permitting more comprehensive interrogation of genomic sequences for a wide array of researchers. In this approach, we start with a collection of sequence variations resulting in a variant call set that is typically of high sensitivity. We then apply a novel "variant adjudication" procedure to discard false positives, while keeping true positive calls: we construct a graph from these variants, the Variant Graph, representing allelic variants as branches of a graph initialized by the current, linear genome reference sequence. Using a graph mapping algorithm (GLIA, a graph extension of the Smith-Waterman alignment algorithm) we developed earlier, we re-map all reads from each of the samples contributing to the candidate calls. We keep candidate variants confirmed by mappings to those branches in the graph that represent the corresponding variant allele, and discard those candidates that were not confirmed by such mappings. This procedure results in a substantial improvement in sensitivity, while also maintaining and improving specificity. Because the graph construction and mapping approach works for most types of SVs in addition to all short variants, variants of different types can be integrated in a single step.

In addition to integrating variants across different tools, our approach offers a procedure to detect already known variants, whether small or structural, in newly sequenced samples. Because a very high fraction of variants in such an individual are already in sequence databases (e.g. >99% of SNPs in any given sample are in dbSNP), re-detecting these variants yields a high fraction of genetic polymorphisms in that individual. Our approach has been first validated as part of generating the final call set from low-coverage 1000 Genomes Project data. We describe further application of this approach in deep whole-genome sequencing datasets.

DEVELOPMENTAL ENHANCERS REVEALED BY EXTENSIVE DNA METHYLOME MAPS OF ZEBRAFISH EMBRYOS

Hyung Joo Lee[1,2], Rebecca F Lowdon[1,2], Brett Maricque[1,2], Bo Zhang[1,2], Michael Stevens[1,2], Daofeng Li[1,2], Stephen L Johnson[1], Ting Wang[1,2]

[1]Washington University School of Medicine, Department of Genetics, St. Louis, MO, [2]Washington University School of Medicine, Center for Genome Sciences and Systems Biology, St. Louis, MO

DNA methylation plays an essential role in normal development. DNA methylation undergoes dynamic changes during development and cell differentiation, and different tissues and cell types exhibit distinct DNA methylation patterns. Recent genome-wide studies discovered that tissue-specific differentially methylated regions (DMRs) often overlap tissue-specific distal *cis*-regulatory elements. However, developmental DNA methylation dynamics of the majority of the genomic CpGs outside gene promoters and CpG islands has not been extensively characterized. Here we generate and compare comprehensive DNA methylome maps of zebrafish developing embryos by using two complementary techniques: methylation-dependent DNA immunoprecipitation followed by sequencing (MeDIP-seq) and methylation-sensitive restriction enzyme digestion followed by sequencing (MRE-seq). From these maps we identify thousands of developmental stage-specific DMRs (dsDMRs) across zebrafish developmental stages. The dsDMRs contain evolutionarily conserved sequences, are associated with developmental genes, are marked with active enhancer histone post-translational modifications, and strongly enrich for binding motifs of transcription factors involved in important developmental processes. Their methylation pattern correlates much stronger than promoter methylation with expression of putative target genes. When tested *in vivo* using a transgenic zebrafish assay, 20 out of 20 selected candidate dsDMRs exhibit functional enhancer activities. Our data suggest that developmental enhancers are a major target of DNA methylation changes during zebrafish embryo development.

A GRAPH GENOME REFERENCE SIGNIFICANTLY IMPROVES VARIANT CALLING

Wan-Ping Lee[1], Kaushik Ghose[1], Vladimir Semenyuk[1], Deniz Kural[1], Ben Murray[2], Amit Jain[2], Richard Brown[2], John Browning[1], Andrew Stachyra[1], Felix Sung[1], Björn Pollex[2], Nate Meyvis[1]

[1]Seven Bridges Genomics, R&D, Cambridge, MA, [2]Seven Bridges Genomics, R&D, London, United Kingdom

The publication of the first, fairly complete, sequence of the human genome by the International Human Genome Sequencing Consortium in 2001, and the development of high throughput, short read sequencing (HTS) has opened the gate for cheap whole genome analysis and a whole host of research and clinical applications are now within reach.

Projects, such as the thousand genomes project, have painstakingly analysed the genomes of additional individuals, from different populations, allowing us to understand how genomes vary between humans. A key insight from these projects, has been that most of the variants in an individual are shared by the population. This has led to the hypothesis that, by incorporating known variants into the current, linear, reference we can improve alignment of HTS reads and thereby improve variant calling.

We test this hypothesis by developing a graph based whole-genome read mapper. A directed acyclic graph (DAG) is constructed by combining the linear reference genome with a list of variants. Our mapper works in two phases. First, we identify regions where a read is likely to map in the DAG. Secondly, we align the read against the DAG, using a graph-aware extension of a string matching algorithm. We achieve 550 and 6,000 reads per second in the first and second steps. The graph mapper generates standard BAM files ensuring compatibility with the majority of other available informatics tools.

We simulated 18,827 insertions with lengths ranging from 2 to 96 bp. We aligned simulated reads against a DAG constructed with human genome and these variants. We passed the graph aligned BAM files to GATK, Samtools and Freebayes to call variants. The overall insertion detection accuracy improved by 15% (from 78% to 93%) when compared to alignments using a linear mapper (BWA).

Our experiments show that there is a considerable gain in variant calling sensitivity when known variants are incorporated into the alignment step. We demonstrate a working and practical implementation of this idea in our graph aligner.

INTER-INDIVIDUAL VARIATION IN CELLULAR IMAGING DATA BETWEEN INDUCED PLURIPOTENT STEM CELL LINES FROM 157 DONORS

Andreas Leha[1], Helena Kilpinen[2], Davide Danovi[3], Minal Patel[1], Alex Alderton[1], Sally Forrest[1], Rizwan Ansari[1], Nathalie Moens[1], Oliver Culley[3], Mia Gervasio[3], Fiona Watt[3], Oliver Stegle[2], Richard Durbin[1], on behalf of the HipSci Consortium[1]

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, [2]European Bionformatics Institute, EBI, Hinxton, United Kingdom, [3]King's College, HipSci Cell Phenotyping, London, United Kingdom

Human induced pluripotent stem cells (iPSCs) enable the generation in vitro of a wide variety of cell types from many individuals. They therefore have great potential for research into the genetics of basic cellular and developmental processes. To address such questions and build a resource, the Human induced pluripotent Stem cells initiative (HipSci) is producing hundreds of iPSC lines from both healthy and diseased individuals.

Here we describe analysis of imaging data from 375 lines derived from 157 healthy donors. Pluripotency and lineage specific markers have been assessed by immunohistochemistry before and after differentiation towards endoderm, mesoderm and ectoderm (1239 data sets in total). We also study cell behavior and morphology in interaction with extrinsic stimuli. These enable us to assess the effect of the genetic (donor) effects versus line variation and other experimental factors.

First, to quantify cellular phenotypes from the raw imaging data, we have developed a novel statistical approach to extract distributional properties of marker abundance from single cell data, accounting for background variation and batch differences.

Second, we employed a variance component approach to estimate the sources of variation of each assay, attributing variability to donor effects and line effects. We found that between 20 and 30% of the variance observed in the differentiation potential can be explained by genetics. In the cell morphology data the genetic component is especially large in the cell-to-cell variation.

Finally, we considered the cellular phenotyping data as a quantitative trait for genetic mapping, performing GWAS on individual trait and combined measures, aggregating over multiple markers. Despite the moderate sample size, initial analyses suggest some genome-wide significant associations. We are in the process of analysing these in more detail and will link the cellular trait QTLs to other molecular assays such as gene expression or DNA methylation.

In conclusion, our results show consistent genetic effects for a range of cellular and developmental traits, including DNA content, cell morphology, proliferation and differentiation potential. This demonstrates that iPSCs are an effective system for genetic analyses to dissect the genetics of pluripotency and early differentiation in humans.

SCIATICA IN FINNISH STUDY POPULATIONS: ROLE OF LOW FREQUENCY VARIANTS

Susanna Lemmelä[1], Svetlana Solovieva[1], Rahman Shiri[1], Markku Heliövaara[2], Johannes Kettunen[3], Verneri Anttila[4,5], Markus Perola[6,7], Ilkka Seppälä[8], Markus Juonala[9,10], Mika Kähönen[11], Jorma Viikari[9,12], Olli Raitakari[9,10], Terho Lehtimäki[8], Aarno Palotie[4,5,7], Eira Viikari-Juntura[1], Kirsti Husgafvel-Pursiainen[1]

[1]Finnish Inst Occup Health, Health Work Ability, Helsinki, Finland, [2]Natl Inst Health Welfare, Popul Health Unit, Helsinki, Finland, [3]Univ Oulu, Inst Health Sci, Oulu, Finland, [4]Mass General Hospit, Anal Translat Genet, Dept Med, Boston, MA, [5]Broad Inst MIT Harvard, Pgrm Med Popul Genet, Cambridge, MA, [6]Natl Inst Health Welfare, Public Health Genom, Dept Chronic Dis Prev, Helsinki, Finland, [7]Univ Helsinki, Inst Mol Med Finland, Helsinki, Finland, [8]Univ Tampere, Dept Clin Chem, Fimlab Laboratories, Tampere, Finland, [9]Turku Univ Hospit, Div Med Turku, Finland, [10]Univ Turku, Res Centre Applied Prev Cardiov Med, Turku, Finland, [11]Tampere Univ Hospit, Dept Clin Physiol, Tampere, Finland, [12]Univ Turku, Dept Med, Turku, Finland

Sciatica or sciatic syndrome is common (population prevalence ~5%) and often disabling low back disorder in the working-age population. It is complex disorder with relatively high heritability (20-40%) but poorly understood molecular mechanisms. This study was conducted in Finns, population enriched with certain low frequency and rare variants due to small founder population followed by bottleneck events. Meta-analysis (291 cases; 3671 controls) was conducted across 2 Finnish GWAS genotyped or imputed at 8.7 million variants. Some of the imputed variants showing association to sciatica were found to indicate bottleneck effect, for being more frequent in the Finns than in other European populations. We therefore genotyped the variants in larger sample set of 2 cohorts (n~4200). Concordance between genotyped and imputed genotypes was high in both populations (>98.5%) validating imputation. The presence of sufficient number of Finns in 1000 Genomes reference panel is crucial for reliable detection of such low frequency variants. Several lead SNPs had also a low frequency (<5%) in general and could not have been identified without imputing genotype data against 1000 Genomes reference. Our data is in line with previous studies illustrating a high utility of dense genotype imputation based on representative data from multiple populations in search for low frequency variants associated with complex diseases. Our confirmation of the variants by genotyping a large population sample allows us to search for replications in other independent Finnish populations. In conclusion, our study suggests interesting novel genes and molecular signaling pathways for development of sciatica, justifying further investigation.

LEMONS – A TOOL FOR THE IDENTIFICATION OF SPLICE JUNCTIONS IN TRANSCRIPTS OF VERTEBRATES LACKING REFERENCE GENOMES.

Liron Levin*[1], Dan Bar Yaacov*[1], Amos Bouskila[1], Michal Chorev[2,3], Liran Carmel[2], Dan Mishmar[1]

[1]Ben Gurion University of the Negev, Department of Life Sciences, Beer Sheva 8410501, Israel, [2]The Hebrew University of Jerusalem, Department of Genetics, The Alexander Silberman Institute of Life Sciences, Jerusalem 91904, Israel, [3]The Hebrew University of Jerusalemm, School of Computer Science and Engineering, Jerusalem 91904, Israel

*These authors contributed equally to this work.

RNA-seq is growing to be a preferred tool for genomics studies of model and non-model organisms. However, DNA-based analysis of organisms lacking sequenced genomes cannot rely on RNA-seq data alone to design the isolation of most genes of interest, as DNA codes both for exons and introns. To address this caveat we designed a novel tool, LEMONS, that takes advantage of evolutionary conservation of both exon/intron boundary positions and splice junction recognition signals to produce high throughput splice-junction predictions in the absence of a reference genome. When tested on multiple annotated vertebrate mRNA data, LEMONS accurately identified 87% (average) of the splice-junctions. LEMONS was then applied to RNA-seq data from the Mediterranean chameleon, which lacks a reference genome, predicting a total of 90,820 exon-exon junctions. We experimentally verified the splice junction predictions generated by LEMONS by amplifying and sequencing twenty randomly selected genes from chameleon DNA templates. Exons and introns were detected in 19 out of 20 of the positions predicted by LEMONS. To the best of our knowledge, LEMONS is currently the only experimentally verified tool that can accurately predict splice-junctions in organisms that lack a reference genome.

CHARACTERIZING POLYMORPHISMS OF *FACTOR VIII* GENE IN THE 1000 GENOMES

Jiani Li[1,2], Ivenise Carrero[1,2], Jingfei Dong[3,4], Fuli Yu[1,2,5]

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [3]University of Washington, Department of Medicine, Seattle, WA, [4]Puget Sound Blood Center, Seattle, WA, [5]Tianjin Medical University, Tianjin Neurology Institute, Tianjin, China

Hemophilia A (HA) is an X-linked bleeding disorder caused by deleterious mutations in the coagulation factor VIII (*F8*). Over 2000 pathogenic mutations in *F8* have so far been documented predominantly in Europeans and Americans of European descendent. Information on *F8* variants with diverse ethnic background is limited. We analyzed 2535 subjects of 27 ethnicities available in the 1000 Genomes Project phase 3, and characterize novel *F8* variants from different ethnicities and analyze their potential functional impacts.

Our study identified 3030 SNPs and 31 INDELs with Africans having the highest number of variants. Among all F8 variants, 86.4% were rare variants (MAF<0.01), 55.6% were novel (did not identified in dbSNP138 dataset). Eighteen variants previously associated with HA were found in our study, and the majority had allele frequencies consistent with the disease incidence in European subjects, but M2257V were presented in 27% percent of African subjects and it did not affect the carriers' *F8* RNA expression level. The mutation E132D, T281A, A303V and D422H were identified only in males. Twelve novel rare non-synonymous variants were predicted to be deleterious and 11 of them appear to be ethnicity specific. A 497-kb large deletion was discovered in 8 female subjects, and it was probably generated by the homologous recombination between two segmental duplications, *int22h-1* and *int22h-2*. The deletion was recurrent and did not affect the *F8* RNA expression level and the average gene expression level on X chromosome.

These results emphasize the importance of interrogating variants on multiple ethnic backgrounds, and the re-discovery of the rare large deletion led us to recapitulate the underlying molecular mutagenesis process driven by segmental duplications, which also accounts for the mutagenesis of inversion and duplications.

MUTATION SIGNATURE AND INTRATUMOR HETEROGENEITY OF ESOPHAGEAL SQUAMOUS CELL CARCINOMA IN A CHINESE COHORT

Qingxuan Song[1], Mengfei Liu[2], Jian Bai[3], Amir Abliz[2], Wenqing Yuan[2], Zhen Liu[2], Jingjing Li[2], Changqing Zeng[3], Hong Cai[2], Yang Ke[2], Jun Li[1]

[1]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [2]Peking University Cancer Hospital & Institute, Laboratory of Genetics, Beijing, China, [3]Chinese Academy of Sciences, Beijing Institute of Genomics, Beijing, China

Esophageal squamous cell carcinoma (ESCC) is one of the most common and most aggressive cancers. The epidemiological features of ESCC are highly variable among world's populations. Rural Anyang in the Henan Province of China is a well-known high-incidence area, yet the causal factors of ESCC in this population remain elusive. We performed exome sequencing of 81 tumor-normal pairs, identified *TP53*, *PIK3CA* and *NOTCH1* as significantly mutated genes, and observed highly recurrent aberrations in several other genes previously reported for ESCC (*ZNF750, MLL2, FAT1, FAT2,* and *FAT3*). Our catalog of ~7,000 single-nucleotide mutations revealed two main signatures: C>T transitions due to spontaneous deamination of 5-methyl-cytosine, and C>T and C>G mutations at TpCpN attributed to the APOBEC family of cytidine deaminases. Since APOBEC activities are associated with exogenous viruses, the prominence of this signature suggests a role of HPV in ESCC etiology, consistent with our previous studies that detected HPV DNA in tumor samples and anti-HPV-E7 antibody in patient's blood. To characterize intratumoral heterogeneity we applied our newly developed method to summarize the clonal frequencies of copy number alternations (CNAs) and single nucleotide mutations in each tumor. This tool, Clonal Heterogeneity Analysis Tool (*CHAT*), considers a wider range of evolutionary scenarios than existing methods concerning the relative timing and phase relationship between a CNA and a mutation it contains. A majority of the 81 ESCCs show a multi-modal distribution of the clonal frequencies, reflecting extensive within-tumor diversity. To better understand the patterns of growth, migration and metastatic potential among different cells within a tumor we compared multiple sub-samples in 10 tumors. For each, we analyzed 4-6 sectors of the tumor, 2-4 samples of adjacent normal tissue, and 1-2 nearby lymph nodes. The spatial heterogeneity of molecular lesions within each tumor uncovers multiple genes and pathways affected in each patient, as well as the likely temporal progression of tumorigenic events that may have driven the initiation and outgrowth of ESCC. By integrating mutation signatures, intratumoral clonal heterogeneity, and clinical outcomes, our results yield new understanding of the molecular bases and evolutionary path of this lethal disease.

INTEGRATION OF GENETIC AND FUNCTIONAL GENOMICS DATA TO UNCOVER CHEMOTHERAPEUTIC INDUCED CYTOTOXICITY

Ruowang Li[1], Dokyoon Kim[2], Scott M Dudek[2], Marylyn D Ritchie[2,3]

[1]Pennsylvania State University, Bioinformatics and Genomics, University Park, PA, [2]Pennsylvania State University, Biochemistry and Molecular Biology, University Park, PA, [3]Geisinger Health System, Biomedical and Translational Informatics, Danville, PA

Chemotherapeutic drugs are used in treatment of many cancers. However, despite its broad usage, some patients experience severe side effects and drug resistance that limit their utility. Understanding the drug-induced cytotoxicity could potentially lead to personalized treatment. However, finding the causal genetic variants that influence the drug's cytotoxicity has been challenging. Using lymphoblastoid cell lines of the HapMap CEU (Utah residents with Northern and Western European ancestry) and YRI (Yoruba individuals) populations, we measured cell lines' cytotoxicity responses for chemotherapeutic drugs: capecitabine, carboplatin, cisplatin, cytarabine, and paclitaxel. To identify variants that are key for each drug, we performed an analysis that jointly analyzed DNA sequence variations, gene expressions and functional genomics data in CEU and YRI HapMap populations. Using whole genome sequencing data from the 1000 Genomes Project and RNA sequencing data from the GEUVADIS project, we identified candidate genetic variants and gene expression variables that are associated with drug cytotoxicity. In addition, meta-analyses of the drugs identified common pathways, proteins, and gene ontologies associated to cytotoxicity indicating potential shared mechanism of chemotherapeutic induced cytotoxicity. To uncover potential interactions between candidate genetic variants and gene expression factors, we integrated them using grammatical evolution neural network implemented in ATHENA. The integration analysis identified unique sets of genetic variants and gene expression factors in interaction models in both CEU and YRI population with high predictive power. For models that have similar predictive power, we ranked the models according to hundreds of functional genomic annotations from the ENCODE project, including genome segmentations and DNase-I hypersensitive sites. Based on the consistency and enrichment of these annotations, we found potential functional models for the chemotherapeutic drugs that could be further tested in follow-up studies.

TRACKING THE EFFECTS OF HUMAN GENETIC VARIATION THROUGH THE GENE REGULATORY CASCADE

Yang I Li[1], Bryce van de Geijn[2], Allegra Petti[3], Yoav Gilad[2], Jonathan K Pritchard[1,4]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]University of Chicago, Department of Human Genetics, Chicago, IL, [3]Washington University in St Louis, Genome Institute, St Louis, MI, [4]Stanford University, Howard Hughes Medical Instite, Stanford, CA

Genetic variation affects gene regulation at many stages, including at the chromatin, RNA, and protein levels. A single DNA variant may affect multiple cellular phenotypes as a consequence of the intimate connection between different levels of the regulatory cascade. However, the levels at which variant effects first manifest themselves and the extent to which these differences percolate through to protein level differences are as yet unknown. Addressing this gap will provide a better understanding of how genetic variation impacts gene regulation and, ultimately, phenotypic traits such as disease risk. Here, we present novel insights on the impact of genetic variation on gene regulation at the systems level by jointly mapping genetic variants that affect the levels of four histone modifications, DNA methylation, transcription rate, stable mRNA levels, RNA decay rate, ribosome occupancy and protein.

COMPARISON OF NORMALIZATION AND DIFFERENTIAL EXPRESSION ANALYSES USING RNA-SEQ DATA FROM 726 INDIVIDUAL *DROSOPHILA MELANOGASTER*

Yanzhu Lin[1], Kseniya Golovnina[2], Zhen-Xia Chen[2], Yazmin L Serrano Negron[1], Hina Sultana[2], Brian Oliver[2], Susan Harbison[1]

[1]Laboratory of Systems Genetics, Center for Systems Biology, National Heart Lung and Blood Institute, Bethesda, MD, [2]Laboratory of Cellular and Developmental Biology, Developmental Genomics Section, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD

We sequenced the RNA of 726 individuals from the *Drosophila* Genetic Reference Panel in order to quantify differences in gene expression among single flies. Until recently, whole-transcriptome studies in *Drosophila melanogaster* were conducted on homogenates of many flies, which restricted our ability to characterize the variability of gene expression among individuals. One of our experimental goals was to identify the optimal analysis approach for the detection of differential gene expression among genotypes, sexes, environments, and their interactions. Here we investigate the effects of three different filtering strategies, eight normalization methods, and two statistical approaches. We also performed a statistical power analysis using the eight biological replicates per sex/genotype/environment condition in our data set. We found that at least three biological replicates per condition were required in order to have sufficient statistical power to detect expression differences among the three-way interaction of these conditions. Some common normalization methods, such as Total Count, Quantile, and RPKM normalization, could not align the data across samples. Furthermore, the mere removal of low-expressed genes from the analysis of Median, Quantile, and Trimmed Mean of M-values normalization methods changed the number and identity of differentially expressed genes. We found that the most critical considerations for the analysis of RNA-Seq read count data were the biological replication of samples, normalization method, and assumption of the underlying data distribution. The best analysis approach to our data was to first normalize the read counts using the *DESeq* method and then apply a generalized linear model assuming a negative binomial distribution using either *edgeR* or *DESeq* software. Genes having very low read counts were removed after the statistical analysis. This strategy properly aligned our data across samples and accommodated its large dynamic range.

YOU MAY HAVE SEQUENCED, BUT HOW WELL DID YOU DO?

Stephen Lincoln[1], Justin Zook[2], Marc Salit[2], the Genome in a Bottle Consortium[2]

[1]Invitae, San Francisco, CA, [2]National Institute of Standards and Technology, Gaithersburg, MD

In clinical sequencing accuracy is paramount. False negatives or false positives could result in a patient receiving an ineffective or unnecessary procedure, or could result in a missed opportunity to improve outcome. Thus, clinical laboratories thoroughly validate their methods and often confirm positive findings with an orthogonal technique (e.g. Sanger).

To support these efforts, the Genome in a Bottle (GIAB) consortium developed a pilot reference material based on Coriell sample NA12878. Fourteen data sets generated using 5 technologies were integrated, producing high confidence genotypes for 85% of the NA12878 genome. Homozygous reference calls were also produced, indicating 2.2 billion locations that we confidently believe are invariant. The integrated call set is highly sensitive and specific, showing 100% concordance with multiple clinical lab sequences of genes in NA12878. The GIAB is also highly concordant with high-accuracy SNP genotypes, with fosmid sequences and with phased pedigree calls.

Twenty papers have already used the GIAB data. In one study of a gene panel in 1105 individuals, the GIAB and 6 similar samples contributed 41% of the useful variants. Nevertheless the GIAB has limitations: About 10-20% of genes in clinical diagnostic panels have few, in any, exonic variants in the GIAB. The GIAB also includes few technically challenging variants, and there is no high-quality copy number variant (CNV) data yet. By contrast, in a review of clinical test results, over 5% of reported pathogenic variants were technically challenging small (single-exon) copy number variants or large (>10bp) indels.

The high concordance of clinical NGS with both GIAB and traditional data has potential implications for the long term role of Sanger confirmation. In one set of 2000 variants observed by clinical NGS, no high confidence calls failed confirmation. Analysis is ongoing to characterize properties of these calls (genomic contexts, variant types, and NGS data quality) to possibly define classes of variant calls which are least likely to be false positives.

Work continues to improve the reference materials. Five additional samples from 2 families are being characterized using high-depth Illumina, Complete Genomics, and Ion Torrent short-read data, as well as Moleculo, PacBio, and BioNano Genomics long-read data. These data will provide an accurate assessment of not just small variants but also CNVs and structural variants in “easy” and some “hard” regions of the genome (segmental duplications, low-complexity sequence, and hyper-variable loci). All of these input data are available for download and analysis by the community.

UNVEILING THE GENETIC AND CAUSAL RELATIONS BETWEEN NICOTINE AND ALCOHOL DEPENDENCE VIA LARGE SCALE META-ANALYSES

Dajiang Liu, for the GWAS&Sequencing Consortium of Alcohol&Nicotine use GSCAN

Pennsylvania State University, Public Health Sciences, Hershey, PA

Nicotine and alcohol use and dependence are known to be influenced by genetic factors, and they co-occur at rates far above chance. In order to unveil the genetic architecture of coding variants underlying nicotine and alcohol use and their co-occurrence, the GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) was formed. As part of the GSCAN effort, we aggregated data from 18 cohorts with ~100,000 individuals genotyped with exome-array. A total of five traits were analyzed including cigarettes per day, packs per year, smoking initiation, drinks per week and age of first smoking. After extensive quality control, 463,184 variants are segregating in the dataset. Among these variants, 158,636/3,731/1,473 are rare non-synonymous/loss-of-function/splice variants with MAF<1%. To aggregate information between cohorts and enlarge sample sizes, we generated summary association statistics and performed meta-analyses of genetic associations with RAREMETAL. For cigarettes per day, we re-discovered a known locus within the nicotinic receptor cluster CHRNA5-CHRNA3-CHRNB4. After fine-mapping this locus we discovered no novel independently associated variants. We also identified an additional signal at STARD3. Functional roles of these identified variants are being validated in mouse studies. Current efforts in GSCAN represent a major step toward understanding the genetic architecture of nicotine and alcohol dependence.

GENOME ADAPTATION OF INDUSTRIAL YEAST TOLERANCE IN SACCHAROMYCES CEREVISIAE AGAINST LIGNOCELLULOSIC BIOMASS CONVERSION INHIBITORS

ZongLin L Liu[1], Yang Zhang[2], Mingzhou Song[3]

[1]USDA-ARS, NCAUR, Peoria, IL, [2]NMSU, Computer Science, Las Cruces, NM, [3]NMSU, Computer Science, Las Cruces, NM

Industrial yeast Saccharomyces cerevisiae has been recognized as a promising candidate for the next-generation biocatalyst development for advanced biofuels production due to its more robust responses to harsh environment conditions. In overcoming toxic chemical compounds derived from pretreatment of lignocellulosic materials for low-cost cellulosic ethanol production, ARS scientists developed a tolerant industrial yeast strain S. cerevisiae NRRL Y-50049 (U.S. Patent) through environmental engineering in a laboratory setting, that is able to in situ detoxify major inhibitory compounds, such as 2-furaldehyde (furfural) and 5-(hydroxymethyl)-2-furaldehyde (HMF), while producing ethanol. Many genes, including a novel aldehyde reductase gene, were identified to be involved in the yeast tolerance and detoxification. New pathways were defined for the in situ detoxification by strain NRRL Y-50049. Genome sequencing analysis uncovered many single nucleotide variations for strain NRRL Y-50049 in comparison with its wild type parental strain. Signature protein expression patterns were also identified for the tolerant Y-50049 in response to challenges of furfural and HMF. Comparative transcriptome analyses using a newly developed computational method ChiNet revealed at least 44 downstream pathways affected by rewired transcription subnetworks against furfural and HMF. Results of our investigation suggest tolerance of strain NRRL Y-50049 adaptation against furfural and HMF occurred at the genome level and thus obtained tolerance is inheritable.

DNA METHYLATION DYNAMICS IN PIGMENT CELL DEVELOPMENT

Rebecca F Lowdon, Hyung Joo Lee, Stephen L Jonshon, Ting Wang

Washington University in St. Louis, Genetics, Saint Louis, MO

Organismal development relies on billions of cells acquiring correct cellular identities. Understanding sequence-specific DNA demethylation events that activate cell type-specific enhancers can elucidate how the epigenome regulates cell fate choice and ultimately contributes to development and disease. We hypothesized that cell type-specific DNA demethylation events drive cell fate decisions by regulating transcription factor binding site (TFBS) accessibility in progenitor cells. We addressed this question using the model organism *Danio rerio* by analyzing the differentiation of two neural crest-derived pigment cell types, melanocytes and iridophores. First, we generated whole-genome methylomes and transcriptomes from neural crest progenitors and differentiated pigment cells. Second, we analyzed the methylomes to examine DNA methylation dynamics over TFBS motifs at predicted enhancers. Our data revealed cell fate-associated demethylation events that are specific to each pigment cell type, as well as progenitor cell-associated demethylation events. The methylome for each developmental stage of pigment cell development was associated with a set of demethylated regions enriched for a specific set of TFBS motifs. Further analysis revealed that demethylation events at TFBS-containing enhancer elements are related to the expression or binding of their cognate transcription factors, elucidating the role of DNA methylation in regulating pigment cell enhancer accessibility. This study investigates how DNA methylation drives development by describing a model for how DNA methylation regulates transcription factor binding at developmental enhancers. Ultimately, this line of study will elucidate how epigenome changes influence development and disease.

A GENOMIC VIEW OF LOCAL ADAPTATION

David B Lowry

Michigan State University, Plant Biology Department, East Lansing, MI

Understanding how organisms adapt to different habitats is of fundamental importance to biology. The endeavor is difficult, however, because local adaptation often involves many phenotypes, each of which has a complex genetic basis. The plant, *Mimulus guttatus*, is a model system for understanding the genetic and genomic basis of local adaptation. Inland populations of *M. guttatus* typically have an early flowering annual life-history, which has evolved in response to hot summer drought conditions. In contrast, coastal populations are always late-flowering perennials, and individuals within those populations are very large morphologically due to local adaptation to year-round soil water availability maintained by coastal summer fog. A large chromosomal inversion polymorphism, *DIV1*, is involved in the evolutionary transition between the annual and perennial *M. guttatus* ecotypes. *DIV1* is a textbook example of how inversions can play a major role in evolutionary adaptations because the phenotypes it affects are well understood and it has been confirmed to be adaptive in nature. My current research program uses a combination of multi-population whole genome resequencing and gene expression analyses to identify the genetic mechanisms by which *DIV1* contributes to the major annual/perennial life-history transition. These data sets are also being used to identify the mechanisms of salt tolerance adaptations by coastal populations and to link gene expression with contrasting physiological responses to water deficit by the ecotypes. Overall, this research is well on its way to gaining a comprehensive genomic view of the mechanisms of local adaptation.

ASSEMBLING MAIZE INBRED CML247: THE MAIZE PAN-GENOME TAKES OFF

Fei Lu[1], Robert Bukowski[1], Qi Sun[1], Edward S Buckler[1,2]

[1]Cornell University, Institute for Genomic Diversity, Ithaca, NY, [2]USDA, ARS, Ithaca, NY

Maize has a complex genome, exhibiting highest amounts of structural variations (SVs) among the major crop species. For example, only about half of the genome is shared between any two maize varieties. The single B73 reference genome is insufficient to represent all of the genomic content of maize, which leads to underrepresented diversity and spurious SNP calls during variation discovery. Hence, a pan-genome, including multiple reference genomes of representative maize varieties, is needed to capture those untapped genetic variations and improve the quality of current variation discovery. Initializing the maize pan-genome construction, we recently started sequencing and assembling maize inbred line CML247 as a pilot project, to optimize maize de novo assembly approach by testing different sequencing platforms and assembly algorithms. Currently, the NRGENE approach, DISCOVAR de novo, Nanopore MinION device, and BioNano Irys System are being tested. In addition, many representative maize de novo assemblies are being collected from our collaborators. A set of ultrahigh density genetic markers (8.1M) is used to assess the quality of assemblies. Now we are constructing the whole genome alignment database of maize and the pan-genome based variation discovery will be performed.

SYSTEMATIC IDENTIFICATION OF GXE DETERMINANTS OF GENE EXPRESSION

Roger Pique-Regi[1], Gregory Moyerbrailean[1], Chris Harvey[1], Omar Davis[1], Donovan Watza[1], Xiaoquan Wen[2], Francesca Luca[1]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI

Adaptations to local environments have played major roles in shaping allele frequency distributions in human populations. Yet, a mismatch between genotype and environment may be responsible for higher disease risk.Recent studies have shown that GxE interactions can be detected when studying molecular phenotypes that are relevant for complex traits (e.g. infection response eQTLs in immune cells). Despite these relevant examples, the extent to which the environment can modulate genetic effects on quantitative phenotypes is still to be defined.

Here we have taken a high-throughput approach to achieve a comprehensive characterization of GxE interactions in humans. To this end we have investigated the transcriptional response to 50 treatments in 5 different cell types (for a total of 250 cellular environments). We then selected the environmental conditions (cell type/treatment) with relevant changes in gene expression and collected deep sequencing data to fully characterize the transcriptional response to environmental changes and to identify genes showing allele specific expression (ASE). We observe that treatments with similar biochemical properties (e.g. nuclear receptor ligands or metal ions) tend to cluster together in principal component analysis (PCA), which also demonstrates cell-type specific transcriptional changes upon environmental perturbation.

We analyzed allele specific expression (ASE) using QuASAR, which allows for joint genotyping and allele specific analysis on RNA-seq data. Across 56 cellular environments we discovered 9548 instances of ASE (FDR<10%), corresponding to 8923 unique ASE genes. We found that in an individual sample, on average, 0.5% of genes with heterozygous SNPs are ASE genes. We then used a Bayesian model across treatments and cell types to identify genes regulated through GxE interactions (conditional-ASE). Consistent with previous analyses of condition-specific eQTLs, we observe that the majority of ASE is consistent across conditions. For a given gene, the probability of ASE is negatively correlated with average expression, with a 4.2 fold decrease per 10x increase in FPKMs. On the other hand, we find a 1.3 fold increase in probability of ASE per 2x change in expression in response to treatments. When we consider a Bayes factor measuring evidence in support of GxE interaction, we find 227 control-only ASE and 245 treatment-only ASE genes. We observe also a trend for increasing evidence of conditional-ASE for genes with larger differential expression. These results provide the first characterization of ASE across a large number of environmental exposures and will contribute to the understanding of how GxE interactions have shaped human phenotypes in different environments and underlie variation in complex traits.

THE BRACHIOPOD GENOME OF *LINGULA ANATINA* PROVIDES INSIGHT INTO THE EVOLUTION OF LOPHOTROCHOZOANS AND CALCIUM-PHOSPHATE-BASED BIOMINERALIZATION

Yi-Jyun Luo[1], Takeshi Takeuchi[1], Ryo Koyanagi[2], Lixy Yamada[3], Miyuki Kanda[1], Mariia Khalturina[1], Manabu Fujie[2], Shinichi Yamasaki[2], Kazuyoshi Endo[4], Noriyuki Satoh[1]

[1]Okinawa Institute of Science and Technology Graduate University, Marine Genomics Unit, Onna, Japan, [2]Okinawa Institute of Science and Technology Graduate University, DNA Sequencing Section, Onna, Japan, [3]Nagoya University, Sugashima Marine Biological Laboratory, Toba, Japan, [4]University of Tokyo, Department of Earth and Planetary Science, Tokyo, Japan

An abundance of the Silurian Period, *Lingula* fossils with morphology very similar to that of extant species inspired Darwin with the idea of "living fossils". Although they superficially resemble bivalve molluscs, lingulid brachiopods show unique features, including radial cleavage and dorsoventrally oriented shells. In particular, their shells are composed of calcium phosphate and collagen fibers, characters shared only by evolutionarily distant vertebrates, one of the biggest mysteries of metazoan evolution. To gain insights into brachiopod evolution, we decoded the 425-megabase genome of *Lingula anatina*. Comprehensive phylogenomic analyses place Lingula close to molluscs, but distant from annelids. Among lophotrochozoans, *Lingula* shows the slowest evolutionary rate of genes associated with basic metabolism. Its gene number increased to ~34,000 by extensive expansion of gene families, especially those associated with shell formation. In addition, we found that Lingula shared shell formation-related genes and mechanisms similar to molluscs, such as chitin synthase and bone morphogenetic protein (BMP) signaling. Although *Lingula* and vertebrates share similar hard tissue components, our genomic, transcriptomic, and proteomic analyses showed that *Lingula* lacks genes involved in bone formation, indicating a classical example of convergent evolution. Furthermore, we showed that *Lingula* has experienced domain combinations to produce shell matrix collagens with epidermal growth factor (EGF) domains and carries lineage specific shell matrix proteins, such as alanine-rich fibers. We propose that gene family expansion, domain shuffling, and co-option of genes appear to be the genomic background of *Lingula*'s unique biomineralization.

DISCOVERY AND GENETIC CHARACTERIZATION OF NEW NEUROPSYCHIATRIC SYNDROMES FROM FAMILY-BASED STUDIES.

Gholson J Lyon[1], Jason O'Rawe[1], Yiyang Wu[1], Han Fang[1], Laura Jimenez Barron[1], Giuseppe Narzisi[1], Michael Schatz[1], Min He[2], Kai Wang[3]

[1]CSHL, Genetics, Cold Spring Harbor, NY, [2]Marshfield Clinic, Human Genetics, Marshfield, WI, [3]USC, Genetics, Los Angeles, CA

We study the breadth and depth of genetic variants in Utah, where there is a large founding population, large family structures and good genealogical records, which enables well powered family-based genetic studies for rare diseases. We use exome and whole genome sequencing (WGS) to identify mutations that segregate with various idiopathic syndromes, and we undertake comprehensive functional studies of many of the newly identified mutations. This has led to the discovery of many new genetic syndromes, including Ogden Syndrome, RBCK1 Syndrome, and most recently RykDax Syndrome. This latter syndrome presents with severe intellectual disability (ID), a characteristic intergluteal crease, and very distinctive facial features. Using WGS datasets from 10 members of one family, we can increase the reliability of the biological inferences with an integrative bioinformatics pipeline, including a new algorithm, Scalpel, developed for more accurate identification of indels. We find a 2 to 5-fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data and analysis pipelines, and we derive greater accuracy when more pipelines are used in conjunction with data encompassing a larger portion of the family. We show that 60X WGS depth of coverage from the Illumina HiSeq platform is needed to recover 95% of indels detected by Scalpel. We also developed SeqHBase, a big data-based toolset for analysing family-based sequencing data, and we demonstrated SeqHBase's high efficiency and scalability on several disorders, including with RykDax Syndrome, where we identified a maternally inherited missense variant in an X-chromosomal gene, TAF1. A "genotype-first" approach led us to other families with variants in TAF1 and containing individuals having a remarkably similar clinical presentation. We continue to advocate for more comprehensive and accurate whole genome analyses in large pedigrees, and we have collected ~2000 DNA samples to date from dozens of families in Utah, including detailed phenotyping information. Some of these samples have undergone exome or whole genome sequencing, and we are currently analyzing these data. This includes the ongoing analysis of whole genomes from 3 families with singleton cases of autism, and an analysis of nine whole genomes from a pedigree with Prader–Willi Syndrome (PWS), Hereditary Hemochromatosis, Familial Dysautonomia (FD), and Tourette Syndrome.

INTEGRATED ANALYSIS OF PROTEIN-CODING VARIATION IN OVER 60,000 INDIVIDUALS

Daniel G MacArthur[1,2,3]

[1]Exome Aggregation Consortium, MA, [2]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [3]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA

Deep reference panels of human genetic variation are critically required for clinical variant interpretation and disease gene discovery, but assembling such panels requires overcoming challenges of data heterogeneity and scale. To explore the patterns of variation across human protein-coding genes we have jointly analyzed exome sequencing data from over 60,000 individuals as part of the Exome Aggregation Consortium (ExAC) using new haplotype-based approaches to variant-calling. We demonstrate that these approaches substantially improve the accuracy and sensitivity of variant detection, especially for small indels, and scale readily to tens of thousands of samples.

Our results provide an unprecedented view of the spectrum of human functional genetic variation down to extremely low frequencies. By comparing these data to a null mutational model we are able to quantitate the depletion of variation in the human population for each gene and variant class. We identify over 4,500 genes with extreme depletion of protein-truncating variants (PTVs), and show that these genes are highly enriched for known disease phenotypes in human and mouse; we also identify many genes with severe functional constraint and no known function. Finally, we describe a new method for inferring regional constraint against missense variation, identifying regions enriched for disease-causing variation.

We show that ExAC population-specific frequency information can be used to flag many previously reported pathogenic mutations as likely benign polymorphisms. In addition, as a case study in the value of large-scale reference data for variant interpretation, we combine data from ExAC, over 500,000 genotyped individuals from 23andMe, and 13,000 sequenced prion disease cases to quantitate the penetrance of reported dominant mutations in the PRNP gene, showing that these variants confer lifetime disease risks ranging from <0.1% to ~100%. Finally, we examine predicted PTVs across all genes, and show that the discovery of human “knockouts” will benefit enormously from large-scale sequencing across multiple populations.

PHOSPHO-PROTEOMIC ANALYSIS OF *SACCHAROMYCES CEREVISIAE* REGULATORY MUTANTS REVEALS NOVEL REGULATOR-TARGET INTERACTIONS IMPORTANT FOR NACL STRESS RESPONSE

Matthew MacGilvray[1], Evgenia Shishkova[2], Josh Coon[2], Audrey Gasch[1]

[1]University of Wisconsin-Madison, Laboratory of Genetics, Madison, WI,
[2]University of Wisconsin-Madison, Department of Chemistry, Madison, WI

Saccharomyces cerevisiae adapts to stressful conditions by altering the expression of condition-specific genes as well as a common stress-activated expression program termed the environmental stress response (ESR). While gene expression alterations and their role in acclimation to many stresses have been fairly well characterized, there is only a rudimentary understanding of the upstream regulatory network and the post-translational protein modifications that control these expression changes. Specifically, the identities of many key regulators (e.g. kinases and phosphatases) involved in this regulatory network and the corresponding target proteins are incomplete. Recently, our lab developed a computational and experimental pipeline that accurately predicted the NaCl signaling network in *S. cerevisiae* to better understand stress-activated signaling networks. This pipeline integrates gene fitness contributions, RNA-seq and phospho-proteomic data, and high-throughput protein interaction data sets to generate an inferred stress network. Here, we harnessed the predictive power of our inferred NaCl network to select regulators thought to be important for NaCl-dependent signaling and investigated their downstream targets. To identify candidate target proteins of the selected regulators, we performed quantitative phospho-proteomic analyses on the wild-type strain BY4741 and derivative mutants of the kinase Hog1, the PKA phosphodiesterase Pde2, and the cell cycle protein Cdc14 before and after NaCl exposure. We scored direct candidate targets by identifying phospho-peptides exhibiting a NaCl-dependent decrease in phosphorylation in the *HOG1*Δ and *PDE2*Δ mutant strains compared to wild-type or a NaCl-dependent increase in phosphorylation in the *CDC*14 mutant strain compared to wild-type. The novel regulator-target pairs identified in this study provide new information about how cells regulate physiology during NaCl stress.

A COMPRESSED SUFFIX ARRAY IMPLEMENTATION OF A POPULATION REFERENCE GRAPH, WITH APPLICATIONS TO *P. FALCIPARUM*

Sorina Maciuca[1], Pf3k Consortium[2], Dominic Kwiatkowski[1,2], Gil McVean[1], Zamin Iqbal[1]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [2]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, United Kingdom

The central position in genome analysis of a single reference genome per species creates considerable problems in regions of high diversity and structural variation. Although very general graph-based alternatives have been proposed, it is hard to scale them to whole genomes and to maintain a connection with a core coordinate system, on which much downstream analysis is based. We have developed a data structure that incorporates the variation seen in a population by embedding alternative sequences into the linear reference genome. This defines an implicit reference graph that stores homologous sequences in regions with variation. Homologous alternative alleles are placed together with their corresponding reference, so coordinates in our structure can be projected onto the primary reference genome. The graph is encoded in a compressed suffix array, allowing us to use a Burrows-Wheeler Transform approach for mapping sequences of any length. We use a modified version of the backwards search algorithm to align reads to our structure in time that scales linearly with the length of the read. We can map reads across multiple sites containing variation and output corresponding paths through the reference structure, with the ability to discover recombination between already known haplotypes. By doing this, we enable re-use of standard downstream analyses based on mapping pileups, and allow easier incorporation of annotation.

We apply the method to the malaria parasite, *P. falciparum*, where there are some regions of the genome of size up to 3kb, where it is common for no sequence reads to map to the reference. These regions correspond to coding sequences of surface antigens, which are exposed to circulating antibodies. Among these highly polymorphic regions are the MSP3.4 and MSP3.8 genes, with most of the diversity being concentrated in their DBL domains. These genes are dimorphic, each having an ancient diverged alternate haplotype that predates the split between *P. falciparum* and *P. reichenowi*, which occurred several million years ago. We can detect recombination just outside the dimorphic region, but not inside. There is as yet no explanation for this apparent long-term balancing selection and there is no clear understanding or representation of the genetic variation in the region. We apply this method to the MSP3.4 and MSP3.8 genes, using data from the Pf3k project (http://www.malariagen.net/projects/parasite/pf3k) and evaluate its performance in building a comprehensive catalogue of genetic variation therein.

TWO NOVEL LIBRARY PREPARATIONS FOR SOMATIC MUTATION DETECTION AND HYPOMETHYLATION PROFILING OF CIRCULATING, CELL-FREE DNA

Vladimir Makarov, Cassie Schumacher*, Catherine Couture*, Juluie Laliberte*, Sukhinder Sandhu, Jonathan Irish, Timothy Harkins, Laurie Kurihara, Sergey Chupreta

Swift Biosciences, Inc, Ann Arbor, MI

Circulating, cell-free DNA (cfDNA) is a powerful, non-invasive sample source that contains tumor associated DNA. This study describes two novel library preparations for Illumina platforms that are sensitive, and specific to assess the genome-wide hypomethylation and detect clinically relevant mutations in 56 oncogenes from cfDNA to monitor tumor burden and treatment.

To characterize the methylation status of cfDNA, NGS libraries were generated utilizing a chemistry that sequentially ligates adapters to single-stranded DNA. Since the library is generated after bisulfite treatment, a high recovery of DNA library is obtained compared with methods where the library is prepared prior to bisulfite conversion. Using 5 ng of cfDNA isolated from healthy subjects and cancer patients, including colon, breast, ovarian, and pancreatic cancer, we prepared and sequenced bisulfite DNA libraries to a depth of 10 million reads. Hypomethylation was analyzed using the approach developed by Dennis Lo lab (PNAS, 2013, v.110, pp. 18761-18768). Upon analysis, hypomethylation of cfDNA from the cancer subjects ranged from 0.4% to 43.4%,where the minimum threshold established for healthy controls was 1.1%.

To detect somatic mutations in cfDNA, we developed a two-step library method that employs multiplexed PCR with hundreds of primer pairs followed by a 10 min adapter ligation step, which results in amplicons 120-160 bp in length, enabling amplification and variant calling from cfDNA. Using this technology, an oncology panel was developed to target known, clinically relevant mutations in 56 genes. The panel was validated using a cohort of control and clinical samples with pre-validated genotypes where robust detection of 5% mutant frequency was observed, and where the limit of detection was as low as 1%. The percent on-target bases and coverage uniformity were comparable between high molecular weight gDNA and 165 bp sized cfDNA. Amplicon libraries were prepared from the 5 ng aliquots of cfDNA used for methylation analysis. Sensitive detection of somatic mutations in cfDNA was observed.

These results indicate that both hypomethylation analysis and multiplexed amplicon-based mutation detection are good candidates for non-invasive cancer detection and monitoring using cell-free DNA.

GENETIC RISK VARIANTS FOR IBD SHAPE THE GUT MICROBIOME IN HEALTHY INDIVIDUALS

Rob Mariman, Mahmoud Elansary, Julia Dmitrieva, Elisa Docampos, Ming Fang, Emilie Theatre, Myriam Mni, Latifa Karim, Wouter Coppieters, Eduard Louis, Michel Georges

Ulg, Unit of Animal Genomics, Liège, Belgium

The host immune system plays an critical role in maintaining homeostasis with resident microbial communities, therefore ensuring that the complex symbiotic relationship is maintained. At the same time, resident microbiota contribute to host nutrition and ensuring a robust immune system. Dysbiosis of the microbiota is associated with various immunological disorders, including inflammatory bowel diseases (IBD). Both genetic and environmental factors are implicated in this disturbance; however, the relative contributions of these two factors, and the mechanism by which they interact remain unclear.

A large fraction of GWAS-identified risk loci encompass genes encoding proteins thought to be involved in host-microbe interactions. This study aimed to investigate the effect of IBD associated risk variants on the microbiome composition. The effect of these common risk variants are better studied in healthy than in sick individuals, as in the latter secondary effects resulting from the actual pathology (i.e inflammation and barrier disruption) are likely to overrule these effects.

For that reason, we made use of the already established CEDAR cohort of 323 healthy individuals that provides integrated genetic (SNP genotypes) and transcriptome data (circulating immune cells subset, as well as samples from various anatomical locations in the intestine). This dataset was further expanded by profiling the gut microbiome. Therefore, we obtained 16S ribosomal RNA (rRNA) gene sequences (covering six variable regions; V1+2, V3+4, V5+6) from both the small -and large intestine biopties.

We have been testing the effect of 167 risk variants for IBD, that were recently established by fine mapping, on the relative abundance of bacterial taxa. We identified several IBD risk variants that affect one or more bacterial taxa in the intestine during conditions of homeostasis. Sixteen variants had an effect on the ileal microbiome, and eight in the rectum. Of those, four associated SNPs were conserved across the two intestinal locations. Genes in close proximity of these SNPs include, *BANK1*, *PTPN2*, *THADA*, and several members of the *STAT*-family. For several IBD risk variants that affect bacterial taxa, cis-eQTLs were found in either blood lympocytes and/or intestinal tissue (i.e cis-eQTLs in *GALC*, *IL2RA*, and *CD40*). Ongoing research will be presented that focuses on understanding these findings in the context of dysbiosis in IBD. This approach may reveal novel connections between the genetic risk factors for IBD and the microbiota, thereby shedding new light on the pathogenesis of IBD.

IOBIO: INTERACTIVE, VISUALLY DRIVEN, REAL-TIME ANALYSIS OF GENOMIC BIG DATA

Chase Miller, Yi Qiao, Tonya Di Sera, Gabor Marth

USTAR Center for Genetic Discovery, University of Utah School of Medicine, Department of Human Genetics, Salt Lake City, UT

Genome-scale data analysis is extremely powerful for discovering disease-causing genetic variations, identifying pathogenic organisms, and to investigate complex biological pathways, but genomic analysis today requires expensive hardware, technical IT expertise, and bioinformatics skills. Because current systems have been optimized for analyses at a large scale, e.g. end-to-end processing of a whole genome or exome, they cannot adequately address rapid, focused, low-cost analyses e.g. in the region of a single gene, which is what many biomedical researchers need.

We developed a new, web-based data analysis system, IOBIO (http://iobio.io), to empower all biological researchers, but especially bench scientists without computational experience, to easily, interactively, and in a visually driven manner, analyze those portions of vast genomic datasets that are essential for their research. IOBIO analysis apps can access data in any desired genomic region from large data files, rapidly process these small data segments across a chain of "atomic" analysis tools implemented as online servers, and display updated results within seconds. Our new gene data inspector app (http://gene.iobio.io) allows scientists to navigate to a given gene, annotate their variants in a few seconds, and examine them in the context of the underlying sequence alignments. Users can call additional variants "on the fly", to find critical variants that might have been missed by the primary analysis.

Another class of IOBIO analysis apps provide researchers with an insightful overview of their genomic big data files in seconds, by sampling random, representative "slices" of these files. Our sequence alignment inspector (http://bam.iobio.io) and variant file inspector (http://vcf.iobio.io) apps provide informative views of these two fundamental genomic data types, and allow users to further investigate chromosomal subregions. We are currently building a metagenomic analysis app that will rapidly display the overall composition of a soil specimen, a blood or mucus sample, and continually refine the composition as additional genomic or transcriptomic reads are transmitted and analyzed. The researcher will be able to "zoom in" on component organisms and further investigate sample composition e.g. at the kingdom, phylum, and taxon levels.

Using these apps, scientists will be able to carry out many meaningful genomic analyses using nothing more than a browser window or a tablet device to visually examine their results and iteratively perform their analyses with changed parameters until they are satisfied with the results. IOBIO will be an open development platform where not only we, but also all 3rd party developers, will be able to rapidly build real-time, visually oriented analysis apps for many general and niche genomic analysis tasks.

EIGHT THOUSAND YEARS OF NATURAL SELECTION IN EUROPE

Iain Mathieson[1], Nick Patterson[1,2], Iosif Lazaridis[1,2], Nadin Rohland[1,2], Swapan Mallick[1,2], David Anthony[3], Dorcas Brown[3], Joseph Pickrell[4], Bastien Llamas[5], Wolfgang Haak[5], David Reich[1,2,6]

[1]Harvard Medical School, Department of Genetics, Boston, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Hartwick College, Department of Anthropology, Oneonta, NY, [4]New York Genome Center, New York, NY, [5]University of Adelaide, Australian Centre for Ancient DNA, Adelaide, Australia, [6]Harvard Medical School, Howard Hughes Medical Institute, Boston, MA

We generated genome-wide polymorphism data for 65 ancient Europeans, spanning a period of eight thousand years. Combining this with data from modern humans, and 20 previously published ancient samples, we performed the first genomic scan for selection to use ancient DNA. We find five genome-wide significant signals of recent selection at LCT, SLC45A2, HERC2, FADS1 and NADSYN1. These loci are associated with lactase persistence, skin and eye pigmentation, fatty acid metabolism and vitamin D levels. Using time series of allele frequencies, we are able to estimate both the strength and date of selection at these loci and at other sites with weaker evidence of selection. Of the small number of loci under strong positive selection, LCT is the clearest example, with a selective coefficient of 2% and a rapid increase in frequency over the past four thousand years. Similarly, the skin-lightening allele of SLC45A2 increased steadily in frequency from approximately zero to approximately one over the past eight thousand years. Other alleles have more complex histories. For example, the eye color locus HERC2 displays evidence of balancing selection, and the light skin allele of SLC24A5 appears to have been absent in Mesolithic hunter-gatherers before being introduced to Europe at high frequency by the first farmers who migrated from the Middle East. Finally, we make use of our genome-wide data to search for evidence of polygenic selection on complex traits, and to localise this in both time and space. In particular, the distribution of genetic height in Europe is known to have been shaped by recent selection and we show that this is likely to be explained by two independent selective events; one for decreased height in Southern Europe in the Middle Neolithic, around six thousand years ago, and one for increased height in a Steppe population who migrated into Eastern Europe around four thousand years ago.

GENE EXPRESSION CONTAINS POPULATION STRUCTURE

Shannon McCurdy[1], Nicolas Bray[3,5], Brielin C Brown[2], Lior Pachter[2,4,5]

[1]UC Berkeley, Simons Institute for Theoretical Computer Science, Berkeley, CA, [2]UC Berkeley, Electrical Engineering and Computer Science, Berkeley, CA, [3]UC Berkeley, Innovative Genomics Institute, Berkeley, CA, [4]UC Berkeley, Mathematics, Berkeley, CA, [5]UC Berkeley, Molecular and Cell Biology, Berkeley, CA

Population structure in genotypes is revealed via a linear projection of the data onto the first few principal components, using principle component analysis (PCA). This fact is useful for determining the geographic origin of individuals, and also reveals PCA to be a useful tool for removing population structure as a confounder in GWAS studies.

A key question has been whether population structure is present in other high-dimensional genomic data. In the case of gene expression, PCA has not revealed obvious population signatures in the first few principle components. However we show that the population structure in gene expression can be revealed through an alternate linear projection. The key to finding the projection is the coupling of the genotypic and gene expression data through canonical correlation analysis, preceded by dimensional reduction of the data.

We present **F**actored **A**ssociation **A**nalysis, which implements the method in a computationally efficient way, and includes statistical procedures for model selection to avoid over-fitting. We demonstrate the uses of FAA on GEUVADIS data, which consists of gene expression measurements from LCLs for hundreds of genotyped individuals from 1000 Genomes. We show how gene expression data to be used in conjunction with genetic data to infer ancestry informative genes and demonstrate for the first time how joint population structure can be properly accounted for in eQTL studies.

EVOLUTION OF MODULATORY REGULATORY PROGRAMS IN TISSUE-SPECIFIC EXPRESSION OF CICHLIDS

Tarang K Mehta[1], Wiktor Jurkowski[1], Jeffrey T Streelman[2], Sushmita Roy[3], Federica Di-Palma[1]

[1]The Genome Analysis Centre, Vertebrate & Health Genomics, Norwich, United Kingdom, [2]Georgia Institute of Technology, School of Biology, Atlanta, GA, [3]Wisconsin Institute for Discovery, Systems Biology, Madison, WI

Unravelling the genetic basis of functional diversification is fundamental for our understanding of the origins of vertebrate diversity and can also have significant implications for animal and human health. Lake Malawi cichlid fish, one of the largest adaptive radiations on earth, represent a remarkable group of over 500 genetically similar species that have evolved within the last few million years from common ancestry. Notably, despite sharing high levels of genetic similarity, these species occupy a large diversity of ecological niches and differ dramatically in phenotypic traits, including skeletal morphology, dentition, colour patterning, and a range of behavioural traits. The number and phenotypic diversity of distinct but closely related cichlid species is unparalleled among vertebrates. Thus, they have the ability to serve as important models for investigating the genomic basis of vertebrate diversification, as well as functional interactions between genes, regulatory elements, and phenotypes.

Comparative functional genomics (mRNA levels, chromatin organization, transcription factor occupancies) is a powerful tool to study the evolution of tissue and species specific divergence. However computational methods to systematically compare these functional patterns across complex phylogenies are in their infancy. We have recently developed Arboretum, an algorithm to identify modules of co-expressed genes across multiple species in a phylogeny. By incorporating transcriptome data from five East African Cichlid species as well as other types of regulatory data, including predicted cis-regulatory elements and miRNA profiles we were able to identify drivers of tissue-specific regulation and evolution in cichlids. Analyses of identified modules (co-expressed genes) across and within cichlid tissues demonstrate interesting patterns of conservation, divergence, gene ontology and motif enrichment. We associated modules with tissue-specific functions and reconstructed evolutionary gene regulatory networks underlying key cellular processes. We plan to associate certain gene regulatory networks to traits of phenotypic diversity amongst analysed cichlid species through integration of further datasets like, for example, coding and putative regulatory SNPs. This unique computational approach will allow us to interrogate the complex regulatory interplay underlying tissue-specific expression as well as the genetic components driving distinct morphological traits in cichlids.

THE HUMAN TRANSCRIPTOME ACROSS TISSUES AND INDIVIDUALS

Marta Melé*[1,2], Pedro G Ferreira*[3], Ferran Reverter*[1], David S DeLuca[4], Jean Monlong[1], Michael Sammeth[1], The GTEx Consortium[1,2,3,4], Emmanouil Dermitzakis[3], Kristin G Ardlie[4], Roderic Guigó[1]

[1]CRG, UPF, Barcelona, Spain, [2]Harvard University, Cambridge, MA, [3]UG, iGE3, SIBi, Geneva, Switzerland, [4]Broad Institute, Cambridge, MA
*contributed equally

The pilot phase of the Genotype-Tissue Expression (GTEx) project has produced RNASeq from 1,641 samples originated from up to 43 tissues from 175 post-mortem donors, and constitutes a unique resource to investigate the human transcriptome across tissues and individuals. Clustering of samples based on gene expression recapitulates tissue types, separating solid from not solid tissues, while clustering based on splicing separates neural from non-neural tissues, suggesting that post-transcriptional regulation plays a comparatively important role in the definition of neural tissues. About 47 % of the variation in gene expression can be explained by variation across tissues, while only 4% can be explained by variation across individuals. We find that the relative contribution of individual variation is similar for lncRNAs and for protein coding genes. However, genes that vary with ethnicity are enriched in lncRNAs, whereas genes that vary with age are mostly protein coding. Among genes that vary with gender, we find novel candidates both to participate and to escape X-inactivation. In addition, by merging information on GWAS we are able to identify specific candidate genes that may explain differences in susceptibility to cardiovascular diseases between males and females and different ethnic groups. We find that genes that decrease with age are involved in neurodegenerative diseases such as Parkinson and Alzheimer and identify novel candidates that could be involved in these diseases. In contrast to gene expression, splicing varies similarly among tissues and individuals, and exhibits a larger proportion of residual unexplained variance. This may reflect that stochastic, non-functional fluctuations of the relative abundances of splice isoforms may be more common than random fluctuations of gene expression. By comparing the variation of the abundance of individual isoforms across all GTEx samples, we find that a large fraction of this variation between tissues (84%) can be simply explained by variation in bulk gene expression, with splicing variation contributing comparatively little. This strongly suggests that regulation at the primary transcription level is the main driver of tissue specificity. Although blood is the most transcriptionally distinct of the surveyed tissues, RNA levels monitored in blood may retain clinically relevant information that can be used to help assess medical or biological conditions.

GENE.IOBIO : A STREAMLINED, WEB APPLICATION FOR INVESTIGATING POTENTIAL, DISEASE-CAUSING VARIANTS

Chase A Miller, Tonya DiSera, Yi Qiao, Gabor Marth

University of Utah School of Medicine, USTAR Center for Genetic Discovery, Salt Lake City, UT

Algorithms to determine potential disease-causing variants from genome sequences have made tremendous progress, however identifying the singular causative variant often remains elusive, requiring extensive manual analysis that is both tedious and expensive. Often, each potential causative variant must be investigated by hand, requiring experts who must navigate through multiple (often more than five) different software programs and databases, interpret the differently formatted results from each, and then integrate and distill the results into a single coherent report. They also need to ensure that variants previously implicated in the phenotype under examination were not missed in their analysis due to e.g. insufficient base coverage in the corresponding gene's coding sequences.

Here we introduce gene.iobio (available at http://gene.iobio.io), a web application that performs critical real-time analysis of sequencing and variation data in the genes and variants of interest. These include the annotation of the coding regions of genes (e.g. using GenCode gene models), predicting the functional impact of the variant (with e.g. SnpEff), visualization of sequence alignment data to determine if sufficient evidence for the called variants exists (to eliminate false positives) and if variants may have been missed due to insufficient coverage (to identify potential false negatives in the initial analysis). The app allows scientists to interrogate prior knowledge of variants through their presence in the appropriate public databases (e.g. ClinVar), and examining the inheritance patterns of a given variant in affected and unaffected family members in pedigrees. Combining these analyses and visualizations into a single streamlined display and analysis driver interface lowers the time, specialized expertise, and expense required to identify the disease causing variants, in addition to broadening the array of questions that can be easily asked and answered by the researcher.

RAPID PHOSPHOPROTEOMIC EFFECTS OF ABA ON WILDTYPE AND ABA RECEPTOR-DEFICIENT ARABIDOPSIS MUTANTS

Benjamin B Minkoff[1], Kelly E Stecker[1], Michael R Sussman[1,2]

[1]University of Wisconsin-Madison, Biochemistry, Madison, WI,
[2]University of Wisconsin-Madison, Biotechnology Center, Madison, WI

Abscisic acid (ABA) is a plant hormone that controls many aspects of plant growth including seed germination, stomatal aperture size, and cellular drought response. ABA interacts with a unique family of 14 receptor proteins. This leads to the activation of a family of protein kinases, SnRK2s, which in turn phosphorylate substrates involved in many cellular processes. The family of receptors appears functionally redundant–to observe a measurable phenotype, 4 of the 14 receptors had to be mutated to create a multilocus loss-of-function quadruple receptor (QR) mutant. Herein, we examine rapid wildtype and ABA receptor-deficient mutant ABA response using isotope-assisted quantitative mass spectrometric-based phosphoproteomic studies. We found that ABA-induced phosphorylation changes occur as early as 5 minutes after hormone treatment, including 3 SnRK2 autophosphorylation sites as well as their substrates. The majority of robust ABA-dependent phosphorylation changes observed were partially diminished in the QR background, whereas many smaller changes observed in the wild type were completely absent in the mutant. Some of the discovery data was validated with follow-up studies using Selected Reaction Monitoring (SRM)-based targeted measurements on a triple quadrupole mass spectrometer; however, there were inconsistencies between the discovery and targeted data, most of which occurred between smaller magnitude fold changes. The intrinsic variability of the targeted processing and analysis pipeline was tested and showed that about a 20% fold change is the minimum that the described targeted pipeline can reproducibly quantify. Altogether, these data expand our understanding of the current model by which the family of ABA receptors directs rapid phosphoproteomic changes.

THE SECRETS OF A TWO-BILLION-YEARS MARRIAGE: MITO-NUCLEAR COEVOLUTION AFFECTS PROTEIN-PROTEIN INTERACTIONS, HUMAN HEALTH AND SPECIATION

Dan Mishmar

Ben-Gurion University of the Negev, Life Sciences, Beer-Sheva, Israel

Mitochondrial activity became essential for cell life ever since the emergence of Eukaryota, more than 2 billion years ago. This activity requires interactions between factors encoded by the mitochondrial DNA (mtDNA) and by the nuclear genome (nDNA). Since the mutation rate of animal mtDNA exceeds that of the nDNA by an order of magnitude, mito-nuclear coevolution occurs to maintain mitochondrial function. We argue that mito-nuclear coevolution is important for three main biological aspects: (A) To maintain mitochondrial function and protect the health of the individual, to (B) protect the structure and function of machineries relying on mito-nuclear physical interactions and (C) to protect the fitness within a given species in diverse environments. We thus hypothesize, that interfering with mito-nuclear co-evolution will affect the health of the individual, disrupt physical interactions between mtDNA and nDNA-encoded factors, and may even lead to speciation events. Firstly, we demonstrate that interfering with mito-nuclear coevolution leads to genotype combinations that alter the risk to develop type 2 diabetes mellitus, thus underlining the importance of mito-nuclear interactions for human health. Secondly, while mutating mtDNA and nDNA-encoded protein subunits of NADH ubiquinone oxidoreductase (complex I) that coevolve, we show that protein-protein interactions were disrupted. Finally, we found that morphologically identical vertebrates (chameleons) differentiated into two populations across an ancient marine barrier (the Jezreel Valley), that sharply diverged in mito-nuclear genotype combinations, which comprise of highly functional mutations. This latter finding was explained by a mathematical model as the emergence of hybrid breakdown – the first stage of a speciation event. Taken together mito-nuclear interactions and co-evolution are involved in human health and in major evolutionary transitions.

INTEGRATIVE PERSONAL OMICS PROFILING (IPOP) DURING WEIGHT GAIN AND LOSS

Tejaswini Mishra[1], Wenyu Zhou[1], Brian Piening[1], Kimberly Kukurba[1], Kevin Contrepois[1], Gucci Gu[1], Colleen Craig[2], Rui Chen[1], George Mias[1], Jennifer Li-Pook-Than[1], Lihua Jiang[1], Siddhartha Mitra[1], Tracey McLaughlin[2], Michael Snyder[1]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]Stanford University, Department of Medicine, Stanford, CA

Obesity-associated insulin resistance has long been recognized as a fundamental aspect of the etiology of type 2 diabetes. However, despite several reports showing the dysregulation of metabolic and inflammatory pathways in obesity, it is unclear if, and how these pathways and their interplay link obesity to insulin resistance. To understand these mechanisms and define the molecular pathways involved in the development of insulin resistance at a systems-wide level, we have used a targeted and dynamic multi-omics approach called iPOP (integrated Personal Omics Profiling) that combines information about the genome, transcriptome, metabolome, microbiome, proteome, methylome and cytokines. We performed iPOP in a cohort of twenty individuals during a period of weight gain (high-caloric diet for 30 days) followed by weight loss (low-caloric diet for 60 days) with the goal of identifying signatures distinguishing resistance to insulin from sensitivity, and determining pathways that characterize weight gain and loss, and the interplay between sensitivity to insulin and overfeeding. Integrative analysis reveals dynamic system-wide changes in an array of diverse biological components and biochemical pathways associated with weight stress and sensitivity to insulin.

RARE NON-CODING VARIATION IN A POPULATION ISOLATE FROM SARDINIA

M Pala[1,2], Z Zappala[1], M Marongiu[2], X Li[1], J Davis[1], A Mulas[2], R Cusano[2], F Crobu[2], K Kukurba[1], C Jones[3], A Battle[4], S Sanna[2], C Sidore[2], A Angius[2], D Schlessinger[5], G Abecasis[6], F Cucca*[2,7], S B Montgomery*[1]

[1]Stanford University, Pathology and Genetics, Stanford, CA, [2]Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy, [3]CRS4, Advanced Genomic Computing Technology, Pula, Italy, [4]John Hopkins University, Center for Computational Biology, Baltimore, MD, [5]University of Michigan, Center for Statistical Genetics, Ann Arbor, MI, [6]National Institute of Aging, Laboratory of Genetics, Baltimore, MD, [7]Università di Sassari, Dipartimento di Scienze Biomediche, Sassari, Italy

Focused on Ogliastra, an isolated population of 60,000 people on the island of Sardinia, the SardiNIA project has benefited from high altruism of participants, genetic homogeneity and minimal admixture that enables high-powered studies of both rare and common genetic variants. We use whole genomes and peripheral blood transcriptomes of 624 individuals to identify genetic variants which contribute to both population and individual phenotypic differences. We identify 20,325 independent expression quantitative trait loci (eQTLs), 1994 lncRNA-eQTL, 72 miRNA-eQTL and 11,055 splicing quantitative trait loci. Comparing Sardinian eQTLs to those previously found in two large studies of mainland Europeans identifies an enrichment of large effects. In addition, for highly-differentiated eQTLs between Sardinia and Europe, we identify enrichment in genes related to malarial resistance and multiple sclerosis – diseases with known prevalence in Sardinia. As study individuals come from 67 families, we further identify segregating patterns of outlier gene expression and allele-specific expression within distinct families. For these effects, we identify the role of rare non-coding variants and validate their effect using the Crispr/Cas9 system. Our work provides insight into how genetic studies of gene expression in isolated populations can inform population history, epidemiology and individual genetic risk factors outside of protein-coding genes. (*co-senior authors)

DERIVING THE REGULATORY NETWORK CONTROLLING THE TRANSCRIPTIONAL RESPONSE TO IFN-I

Sara Mostafavi[1,2], ImmVar & ImmGen Consortium[3]

[1]University of British Columbia, Statistics, Vancouver, Canada, [2]University of British Columbia, Medical Genetics, Vancouver, Canada, [3]., ., ., MA

Recent genome-wide genetic and transcriptomic studies have highlighted the dysregulation in immune-related pathways in varied pathologies. In particular, type I interferons (IFN-I) are a major family of immune cytokines that have been implicated in autoimmune, immunodeficiency, and neuropsychiatric disorders. Induction of IFN-I leads to downstream antiviral and inflammatory programmes, which depending on the genetic, cellular, and environmental context is orchestrated by dozen to hundreds of IFN-I induced genes. Although IFN-I are one of the most studied immune cytokines, much remains unknown about the precise regulatory network that governs the induction of shared and context-specific IFN-induced genes, and the role of genetic factors (cis and trans) that modulates the transcriptional response. As part of the Immunological Genome (ImmGen) and the Immune Variation (ImmVar) projects, we used a varied collection of genomic data from multiple immune cells types, and across multiple conditions (including baseline and IFN stimulation) and species (human and mouse), to comprehensively characterize the transcriptional regulatory network of IFN-I. Here, we quantify cell- and species-specificity of the transcriptional response to IFN-I across eleven distinct immune cell types, and across two species, identifying shared- and cell-specific components of the regulatory network. We describe an approach for statistically integrating these heterogeneous data collection, including 1400 gene expression profiles, to construct a combined regulatory co-expression network for IFN-I. This predicted regulatory network, which we validate using RNAi and ATAC-seq assays, highlights the role of canonical and non-canonical transcriptional regulators, and modular organization for the induced genes. Further, using genotyping data from the ImmVar project, we identify a dozen significant trans-genetic factors (i.e., trans-eQTLs) that impact the predicted regulatory relationships. These trans-eQTLs had not been previously identified as genome-wide significance due to the multiple hypothesis-testing burden, and their identification was enabled by the considerably reduced search space implied by the learnt regulatory network. These data and accompanying analyses present a valuable resource for interpreting IFN-I induced genes, and distinct dysregulated components in disease.

ESTIMATING SUBNUCLEAR BODIES AS HOLES AND CAVITIES IN THE 3D SHAPE OF DNA

Yuichi Motai[1], Masahiko Kumagai[2], Ryohei Nakamura[2], Hiroyuki Takeda[2], Shinichi Morishita[1]

[1]The University of Tokyo, Department of Computational Biology, Chiba, Japan, [2]The University of Tokyo, Department of Biological Sciences, Tokyo, Japan

What is the precise shape of a subnuclear body such as nucleolus, Cajal body and Polycomb body in the nucleus? Which part of the DNA molecule is in contact with a subnuclear body? To address these questions, one may stain and observe individual subnuclear bodies using microscope; however, it is unclear which genomic regions interact with each subnuclear body.

We here explore an approach of using chromatin conformation capture methods to estimate the shape of subnuclear bodies. The Hi-C method enables us to perform a comprehensive analysis of the spatial architecture of DNA. The method outputs pairs of interacting genomic positions, and there have been proposed various approaches that utilize multi-dimensional scaling to reconstruct the 3D shape of DNA from the interactions. Although the reconstructed 3D shape of DNA is a series of points and has no volume, we treat DNA together with its binding proteins as the union of balls along all points in the 3D shape of DNA. Furthermore, assuming the excluded volume effect of subnuclear bodies, we represent these subnuclear bodies as holes or cavities of the union of balls along the 3D shape of DNA.

Detecting holes and cavities has been studied in topology but is subject to the instability under small changes and noise. To overcome this problem, we adopted persistent homology that has been developed in topology for detecting variously sized holes and cavities that are robust against noise. Because we calculate the shape of subnuclear bodies from the 3D shape of DNA, the integration of genome annotation data (e.g. histone and DNA modifications) allows us to predict the biological entity of each hole and cavity such as subnuclear bodies. We applied this approach to Hi-C data collected from budding yeast (*S.cerevisiae*), mouse, human and medaka (*Oryzias latipes*) embryos and mature cells to estimate subnuclear bodies as holes and cavities of a reconstructed 3D shape of DNA.

PARENET: A TOOL FOR DEGRADOME ASSISTED DISCOVERY AND VISUALIZATION OF SMALL RNA/TARGET INTERACTION NETWORKS

Leighton Folkes[1], Matthew Stocks[2], David Swarbreck[1], Tamas Dalmay[3], Vincent Moulton[2], Simon Moxon[1]

[1]The Genome Analysis Centre, Bioinformatics, Norwich, United Kingdom, [2]The University of East Anglia, Computing Sciences, Norwich, United Kingdom, [3]The University of East Anglia, Biological Sciences, Norwich, United Kingdom

Small RNAs (sRNAs) are an important class of short (20-24nt) non-coding RNAs which regulate gene expression in both plants and animals. Recent studies on sRNA interactions have shown that many do not operate independently, but instead can form part of a larger, more complex, regulatory interaction network. However, these studies have been carried out on only a tiny subset of all sRNAs, such as miRNAs, or were based on computational predictions using sequence complementarity and not empirical evidence. A new technique called Parallel Analysis of RNA Ends (PARE) or more commonly known as degradome sequencing can be used to capture a snapshot of the mRNA degradation profile on a genome-wide scale from which we are able to extract clear signals of position specific sRNA mediated mRNA cleavage. Computational methods now exist to rapidly analyse the degradome to identify and validate sRNA/target interactions for all sRNAs obtained from a next-generation sequencing experiment. The resultant sRNA/mRNA interactions evidenced through the peaks in mRNA degradation signal can be used to identify regulatory interaction networks on a genome-wide scale. Several computational methods have been described and used to identify such networks. However, these methods have relied upon in-house computational pipelines that are not publicly available. Here we describe a new publicly available, user-friendly, interactive software tool that allows users to build, visualize and investigate sRNA interaction networks which are supported by genome-wide degradome analysis.

sRNAs play important roles in diverse processes such as pathogen response, development, reproduction and stress response and we reason that large scale regulatory networks of sRNA interactions are also involved in such diverse processes. By using our approach, we have been able to discover new regulatory interaction networks as well as provide a new tool that requires no computational expertise to use.

WHICH GENETIC VARIANTS IN DNASE I SENSITIVE REGIONS ARE FUNCTIONAL?

G Moyerbrailean[1], C Harvey[1], C Kalita[1], X Wen[2], F Luca[1], R Pique-Regi[1]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Michigan, Biostatistics, Ann Arbor, MI

Despite large experimental efforts to characterize regulatory regions in the genome, we are still missing a systematic definition of functional and silent genetic variants in the regulatory genome. Here, we integrated DNaseI footprinting data with sequence-based transcription factor (TF) motif models to predict the impact of a genetic variant on TF binding across 653 samples and 1,372 TF motifs. In total we identified 5,810,227 genetic variants in footprints, 3,831,862 of which are predicted by our motif models to affect TF binding. Comprehensive examination using allele-specific binding (ASB) reveals that only a small fraction (3%, using Storey q-value method) of genetic variants in footprints show evidence for ASB. In contrast, functional genetic variants predicted by our integrative model are highly enriched for ASB (56%), representing a large improvement over other existing annotations that do not consider whether the sequence change is silent or not. The high-resolution provided by the footprints combined with the sequence prediction allowed us to characterize evolutionary features of TF binding sites at an unprecedented resolution and to fine-map the causal variants in eQTL and GWAS studies. The rich meta information provided by the TF-motif / tissue-type identity should also provide a more solid lead towards identifying the mechanism underlying the association. For complex polygenic traits, selective pressure may have acted on a group of binding sites for the same TF. Using a modified version of the McDonald-Kreitman test for selection, we find 84 TF motifs with binding sites enriched for fixed functional differences, suggestive of positive selection. To further investigate the functional significance of groups of binding sites, we used fgwas to integrate our annotations with 18 complex traits from GWAS meta-studies. Using results from a lipid meta-study, we find that the enrichment for LDL level-associated SNPs predicted to affect HNF4 binding sites is 9.1-fold higher than in a background model that includes baseline regulatory annotation. For a different trait, we find that the enrichment for height-associated SNPs predicted to affect OCT4 binding sites is 6.6-fold higher than background. In terms of fine-mapping, across all 18 traits, we identified 86 SNPs within GWAS hit regions with at least a 2-fold increase in the posterior odds of being the potentially causal SNP in the region. To validate the predicted allelic effects on gene expression, and the underlying molecular mechanism, we performed reporter gene assays for seven motif-disrupting SNPs in GWAS regions. We find that five of the SNPs tested have enhancer/repressor activity and that four have allele-specific activity. Our results underscore the importance of having well-calibrated models to predict the effect of a sequence change on TF binding, and serve as a resource to identify and characterize non-coding functional genetic variation.

IMPROVEMENTS TO GENCODE ARE TRANSFORMING THE INTERPRETATION OF VARIATION

Jonathan M Mudge[1], Adam Frankish[1], Nathan Boley[2], James Wright[3], Jyoti Choudhary[3], Jennifer Harrow[1]

[1]Wellcome Trust Sanger Institute, Informatics, Hinxton, United Kingdom, [2]Stanford University, Genetics, Stanford, CA, [3]Wellcome Trust Sanger Institute, Proteomics and Mass Spectrometry, Hinxton, United Kingdom

The correct identification and interpretation of disease-linked variants is arguably the major goal of human genomics at the present time. While numerous computational pipelines are available to perform such work, most use existing gene annotation to both capture mutations and to infer their functional impact, e.g. on translation or splicing. GENCODE – produced by HAVANA / Ensembl - is the most comprehensive geneset available in terms of transcript count; it aims to capture all transcripts associated with protein-coding genes, lncRNAs and pseudogenes. GENCODE, however, remains incomplete; thousands of experimentally identified transcripts have yet to be incorporated, while the functional annotation of most of our models is inferred rather than confirmed. Indeed, there remains much debate as to the proportion of the transcriptome that simply represents 'noise', and such functional ambiguities can place caveats on variant interpretation. However, both the scope and usability of GENCODE is being improved by the integration of next generation datasets. Here, we will focus on the reappraisal of our transcription start sites based on 5'cap-selected RNA libraries, and the incorporation of highly conservative, low FDR proteomics data into our annotation of coding regions. In the former case, it is now apparent that few – if any – genes utilize a single TSS even within the same first exon, and this granularity is almost entirely absent from current genebuilds. However, the description of alternative TSS based on CAGE and RAMPAGE data can allow us to decipher the true relationship between our models and regulatory features such as uORFs, promoter regions and TF-binding sites. Secondly, mass spectrometry and Ribo-seq data are allowing us to identify entirely novel coding regions, within existing protein-coding genes and pseudogenes as well as in intergenic space. However, we believe that previous reports may have exaggerated the number of putative missing protein-coding genes, and we will emphasize the strict criteria required to allow these annotations to be made with confidence.

On behalf of the GENCODE consortium.

CO-FACTOR DEPENDENCIES OF TRANSCRIPTIONAL ENHANCERS

Felix Muerdter, Łukasz M Boryń, Alexander Stark

Research Institute of Molecular Pathology (IMP), Stark Lab, Vienna, Austria

Transcriptional enhancers are key gene regulatory elements encoded in our DNA. They establish cellular fates in development and can misregulate gene expression in disease. Nevertheless, enhancers only come to life through the action of DNA binding transcription factors and the co-factors they recruit.

Recently, co-factors have become attractive targets for therapeutic intervention in cancer. Many tumors are caused and maintained by misexpression of oncogenes, yet, transcription factors are rarely valid targets for therapy due to the lack of small molecule inhibitors. Co-factors and chromatin regulators can be efficiently targeted and their inhibition can show strong and cancer-specific effects. Nonetheless, it is unclear how this specificity is achieved and whether or not it is encoded at the sequence level.

We have begun to dissect the interplay between co-factors and their target enhancers using the high-throughput enhancer activity assay STARR-seq. Measuring the activity of enhancers while simultaneously disrupting the function of co-factors gives us the ability to ask several questions. How does co-factor function relate to enhancer activity? Do different enhancers rely on different combinations of co-factors, and how can co-factor inhibition have enhancer-specific effects? Answering these questions is invaluable for the discovery of future therapeutic targets and for the understanding of current approaches.

INCOMPLETE LINEAGE SORTING REVEALS PREVALENCE OF SELECTIVE SWEEPS IN GREAT APE EVOLUTION.

Kasper Munch, Mikkel H Schierup, Thomas Mailund

Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

The frequency of hard selective sweeps in recent human history is under scientific debate. Among the great apes, patterns of genetic polymorphism offer a window for possible detection of sweeps into the past few hundred thousand years only. We present a more powerful approach that exploits the patterns of incomplete lineage sorting along the genome to identify regions subject to selective sweeps on an evolutionary time scale. This is particularly applicable to the human-chimpanzee ancestor and the human-orangutan ancestor, which show the same high levels of incomplete lineage sorting. We show that the average impact of such sweeps reaches hundreds of kilo bases away from genes. We further show that the frequency of sweeps varies across the great ape phylogeny, even in ancestral species with similar effective population sizes. We identify many strong sweeps on the X chromosome in particular. These have caused the removal of incomplete lineage sorting in regions of several Mb that together represent about half of the chromosome.

MODELING POPULATION SIZE CHANGES LEADS TO ACCURATE INFERENCE OF SEX-BIASED DEMOGRAPHIC EVENTS

Shaila Musharoff[1], Suyash Shringarpure[1], CAAPA Consortium[2], Carlos D Bustamante[1], Sohini Ramachandran[3]

[1]Stanford University, Genetics, Stanford, CA, [2]JHU, School of Medicine, Baltimore, MD, [3]Brown University, Ecology & Evolutionary Biology, Providence, RI

Sex-biased demographic events ("sex-bias"), e.g. male-biased migration or female-biased mating, can be inferred by comparing X chromosomal to autosomal genetic variation. The conventional sex-bias estimator Q assumes constant population size. However, due to effective size differences, X chromosomes and autosomes recover genetic variation at different rates after size changes, so Q-based inference on a population of non-constant size biases conclusions about sex-bias and the female fraction of the effective population size (pF). Our novel sex-bias framework models population size changes and estimates pF from site-frequency spectra: it estimates demographic parameters from autosomal data and tests sex-bias models on X chromosome data. It detects sex-bias better than Q for data simulated with human-relevant demography. For example, for recent growth, our test has more power (15% for mild bias, 40% for extreme bias) and accurately estimates pF whereas Q does not.

Our method applied to Thousand Genomes phase 3 high-coverage Complete Genomics whole genomes implies female bias in Africans and Europeans. The Q-based estimator that assumes constant size applied to an African population (YRI) gives an unrealistic estimated female proportion (pF) of 1.02; our method using an old growth model gives a pF of 0.63 (i.e. 63% breeding females). For a European population (CEU), the Q-based estimator gives a pF estimate of 0.80; our method using a growth model gives a pF of 1.22 while the more likely complex model with multiple bottlenecks and recent rapid expansion gives a pF of 0.77. Our method tests for specific sex-biased scenarios in addition to estimating pF, and joint tests on CEU and YRI imply a male-biased migration out of Africa. We further applied our method to "Consortium on Asthma among African-ancestry Populations in the Americas" (CAAPA) whole-genome 35x sequence data and find evidence for varying amounts and timings of recent sex-biased admixture in populations including African-Americans, Jamaicans, Brazilians, and Hondurans. These results give insight into human sex-bias and demonstrate that modeling size changes is essential to estimating sex-bias parameters. Our novel approach can clarify pervasive signatures of sex-bias in sexual species and provide null models for selection scans.

GENOME SEQUENCING ELUCIDATES SARDINIAN GENETIC ARCHITECTURE AND AUGMENTS GWAS FINDINGS: THE EXAMPLES OF LIPIDS AND BLOOD INFLAMMATORY MARKERS.

Ramaiah Nagaraja[1], Carlo Sidore[2,3,4], Fabio Busonero[2,3,5], Andrea Maschio[2,3,5], Eleonora Porcu[2,3,4], Magdalena Zoledziewska[2], Maristella Steri[2], Hyun M Kang[3], Vicente Diego Ortega del Vecchyo[6], Charleston W.K. Chiang[7], Robert Lyons[5], Chris Jones[8], Andrea Angius[2,8], John Novembre[9], Serena Sanna[2], David Schlessinger[1], Francesco Cucca[2,4], Gonçalo Abecasis[3]

[1]National Institute on Aging, Laboratory of Genetics and Genomics, Baltimore, MD, [2]CNR, Monserrato, Istituto di Ricerca Genetica e Biomedica, Cagliari, Italy, [3]University of Michigan, Center for Statistical Genetics, Ann Arbor, MI, [4]Universita degli Studi di Sassari, Sassari, Italy, [5]University of Michigan, DNA Sequencing Core, Ann Arbor, MI, [6]University of California, Interdepartmental Program in Bioinformatics, Los Angeles, CA, [7]University of California, Department of Ecology and Evolutionary Biology, Los Angeles, CA, [8]Parco Scientifico e Tecnologico della Sardegna, Center for Advanced Studies Research and Development in Sardinia (CRS4), Pula, Italy, [9]University of Chicago, Department of Human Genetics, Chicago, IL

Studies of common genetic variants have provided entry points to analyse the mechanisms underlying many complex traits and diseases. Extension of studies to the large reservoir of rare and population-specific variants could accelerate biological understanding. Rare variants can be analyzed via DNA sequencing, but designing studies in which enough copies of each variant can be observed to detect genetic associations is challenging. Families and founder populations, where variants rare elsewhere can occur at moderate frequencies, help overcome these limitations. Here, we use sequencing to assess the contribution of genetic variation to quantitative traits, using, as exemplars, levels of low-density lipoprotein cholesterol (LDL-c) and five circulating inflammatory biomarkers. We report ~17.6M genetic variants from whole-genome sequencing of 2,120 Sardinians; 32% are absent from prior sequencing studies. Furthermore, ~129K variants common in our sample (frequency >5%) are rare elsewhere (<0.5%). Six GWAS signals, including a major new locus, were observed for LDL-c, and 19 signals, five novel, for inflammatory markers. All new associations would be missed using 1000 Genomes data alone, underlining the advantages of large-scale sequencing in Sardinia.

HETERO-DGF: A NOVEL ALGORITHM TO DECOMPOSE HETEROGENEOUS BINDING FOOTPRINTS OF TRANSCRIPTION FACTORS

Ryo Nakaki, Shuichi Tsutsumi, Hiroyuki Aburatani

Genome Science Division, RCAST, University of Tokyo, Meguro-ku, Tokyo, Japan

Transcription factors (TFs) recognize different genomic sites, thereby regulating gene expressions in a state-specific manner. The recognition of the state-specific binding sites by TFs is closely controlled by various patterns of binding motifs, where various transcription complexes of the TFs are recruited. In addition to recognizing of its particular binding motif, interactions of different co-factors may select the TF binding motifs distinct at the level of nucleotides. Recently, genome-wide DNase I hypersensitivity profiles obtained by DNase-seq experiments enable us to predict the TF binding footprints at the nucleotide level. Although there are some computational algorithms to identify TF footprints, they are limited to assigning a particular DNase I cleavage pattern to each binding motif. We propose a novel algorithm to predict the heterogeneous TF footprints, through decomposing the binding motifs with different patterns of DNase I cleavages. Compared with the existing methods to predict TF footprints, our algorithm was about 1.5-times more accurate in identification of the footprints of TF recognizing short binding motifs (65~75% at average true positive rates), such as ETS and GATA factors. Subsequently, this algorithm was applied to identify the cell-specific binding footprints of GATA2 in HUVEC (human umbilical vein endothelial cells) and K562. We found that GATA2 cell-specifically recognized diverse binding motifs including ‘GATAA’-sequence, each of which corresponds to different DNase I cleavage patterns. Moreover, we identified that GATA2 heterogeneously recognize cell-specific and common binding motifs in each cell-type. These results have indicated that hetero-DGF is useful to predict state/site-specific DNA-recognition mechanisms of TFs.

PACBIO LONG READ SEQUENCING AND STRUCTURAL ANALYSIS OF A BREAST CANCER CELL LINE

Maria Nattestad[1], Sara Goodwin[1], Timour Baslan[1], James Gurtowski[1], Karen Ng[2], Timothy Beck[2], Yogi Sundaravadanam[2], Melissa Kramer[1], Eric Antoniou[1], John McPherson[2], James Hicks[1], Michael C Schatz[1], Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Ontario Institute for Cancer Research, Genome Technologies, Toronto, Canada

Genomic instability is one of the hallmarks of cancer, leading to widespread copy number variations, chromosomal fusions, and other structural variations in many cancers. 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, along with numerous other amplifications and deletions. 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 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 60x coverage of the genome with average read lengths of 9-13kb (max: 60kb). PacBio read coverage is highly correlated with the copy number assignments made using short read sequencing technologies, although the long reads provide more consistent coverage across repetitive elements. Furthermore, using the structural variation analysis program LUMPY and our new hybrid mapping and de novo assembly algorithm for analyzing split-read alignments, we have developed a detailed map of structural variations in this cell line. We have tentatively identified more than 900 intra-chromosomal and 300 inter-chromosomal variations, including many of the previously known gene fusions in SK-BR-3. Taking advantage of the newly identified breakpoints, we have developed an algorithm to reconstruct the mutational history of this cancer genome. From this we have characterized the amplifications of the HER2 region, discovering a complex series of nested duplications and translocations between chr17 and chr8, two of the most frequent translocation partners in primary breast cancers. To our knowledge, this establishes the most complete cancer reference genome to date, and all data will be released publicly under the Toronto Agreements so that researchers around the world can utilize this important resource.

A GENOME-WIDE EXPLORATION OF THE ANTAGONISTIC PLEIOTROPY THEORY OF SENESCENCE SUPPORTS ITS ROLE IN SHAPING HUMAN AGEING AND DISEASE

Juan A Rodriguez[1], Arcadi Navarro[1,2,3]

[1]Institute of Evolutionary Biology (Universitat Pompeu Fabra-CSIC), Barcelona, Spain, [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [3]Center for Genomic Regulation (CRG), Barcelona, Spain

Human senescence has long been a mystery, with no universally accepted theory accounting for its ultimate evolutionary causes. Perhaps the most popular evolutionary explanation proposed so far is the antagonistic pleiotropy hypothesis (known as the pleiotropic theory of senescence), suggested by G. Williams in 1957. This theory poses that senescence is linked to mutations with multiple (i.e. pleiotropic), antagonic effects that occur in different periods of life. Mutations that are damaging for the organism late in life might be favored by natural selection if they are advantageous early in life, when they can result in increased reproductive success of their carrier.

This theory is consistent with a handful of examples, in most cases speculative, coming from certain genes (such as mTOR) or specific conditions (such as Huntington's disease or Haemochromatosis), but it is largely accepted by default, since it has been supported by any systematic study. In particular, there is no assessment of the impact of antagonistic pleiotropy on human senescence and disease.

Using data from Genome-Wide Association Studies (GWAS) we quantified the global extent and evolutionary implications for our species of antagonistic pleiotropy. We obtained several relevant results. First, we observed a clearly significant excess of early-late antagonistic pleiotropy candidates in our genomes. This excess of alleles that protect from early-onset diseases while increasing the risk of late-onset, age-related pathologies constitutes the first validation this theory. Second, we detected that two independent sets of experimentally validated, ageing-related genes are involved in far more pleiotropies than expected by chance. Next, we scrutinized human genomic regions whose methylation patterns present age-related differences (usually interrelated to gene expression) and found a significant overlap with regions harboring pleiotropies. This was further supported by evaluating the abundance of pleiotropies in an independently ascertained set of genes whose expression varies with ageing. Additionally, we found sound examples of the action of natural selection in favor of haplotypes that fulfill the predictions of the theory. Finally, an additional benefit from our approach is that it unveiled some putative relationships between diseases that so far have not been detected by comorbidity analysis and that may be relevant to diagnosis and treatment of age-related pathologies.

EXPLORING POPULATION STRUCTURE THROUGH LARGE PEDIGREES.

Dominic Nelson, Claudia Moreau

McGill University, Human Genetics, Montreal, Canada

Large pedigrees contain a wealth of information about the genetic history of a population. For pedigrees containing thousands to millions of individuals, such as the province-wide pedigree that exists in Quebec, Canada, the joint analysis of genetic and pedigree data is numerically challenging. We introduce a suite of tools for the joint modeling of genetic and pedigree data, and then apply these tools to several problems in large-scale population genetics. We will discuss four applications: 1-Inference of disease history through tracking of ancestral carriers; 2-Modeling of long-range linkage induced by fine-scale population structure; 3-Reconstruction of missing pedigree information; 4-Optimal sampling strategies for building large haplotype panels in spatially inhomogeneous populations. These discussions will be based on the BALSAC genealogical database, recessive diseases in the French-Canadian population, and genome-wide association studies.

INVESTIGATING AXOLOTL REGENERATION VIA SINGLE CELL TRANSCRIPTOMICS

Jeffrey D Nelson[1], Ron Stewart[1], Colin N Dewey[2], James A Thomson[1]

[1]Morgridge Institute for Research, Regenerative Biology, Madison, WI, [2]University of Wisconsin, Computer Sciences, Madison, WI

The mexican axolotl salamander (*Ambystoma mexicanum*) possesses regenerative capabilities that are extraordinary among vertebrates. Following injury, these animals can regenerate body parts ranging from regions of the brain and heart to entire limbs and tails. While this phenomenon was first observed about 250 years ago, the biological mechanism by which they accomplish such incredible feats is still relatively poorly understood. Following limb amputation, a cluster of cells known as a blastema forms at the wound site. This blastema then goes on to proliferate and fully regrow the missing limb. Previous work in our lab has identified coordinated gene expression programs that likely govern various stages of limb regeneration via blastema outgrowth. However, these experiments were performed on bulk blastema tissue samples, leaving open the critical question of just how individual adult cells coordinate themselves to complete such elaborately patterned tasks. Are blastemas made up of de-differentiated multipotent progenitor cells? Or, are they populated by fate-restricted tissue-specific adult cells that migrate from stump muscle, cartilage, nerve, and blood? Luckily, recent technological advances have allowed us to address these questions. Using single cell microfluidic capture platforms we have isolated and extracted total RNA from dissociated individual blastema cells in regenerating axolotl limbs. We then performed next generation sequencing on these samples, yielding axolotl transcriptomes at a single cell resolution. With bioinformaticist collaborators, we are currently developing methods by which we may identify gene expression patterns that represent sub-populations of cells participating in the regeneration process over time. Ultimately, we hope that the knowledge gained through this work will someday be applied in a biomedical setting to encourage tissue regeneration after injury or disease.

JDN was supported by an NHGRI training grant to the Genomic Sciences Training Program 5T32HG002760.

EVOLUTION OF GENE EXPRESSION IN GIANT ISLAND MICE

Mark J Nolte[1,2], Melissa Gray[1], Michelle Parmenter[1], Colin Dewey[3], Bret Payseur[1]

[1]University of Wisconsin-Madison, Laboratory of Genetics, Madison, WI, [2]Supported by NHGRI training grant to the Genomic Sciences Training Program 5T32HG002760, UW-Madison, Madison, WI, [3]University of Wisconsin-Madison, The Department of Biostatistics and Medical Informatics, Madison, WI

Approximately one hundred and fifty years ago house mice were introduced onto Gough Island, a remote island in the South Atlantic. Since that time Gough Island (GI) mice evolved the largest body size of any wild house mouse population in the world, with individuals being nearly twice the size of their mainland counterparts. The rapid evolution of GI gigantism is a striking example of a widespread biological phenomenon recognized by researchers for decades: island-colonizing species often evolve extreme body sizes, either dwarfism or gigantism. Other notable examples of this phenomenon include evolved dwarfism in an extinct insular elephant and hominin, and acquired gigantism in an insular parrot and lizard, the Kakapo and San Esteban chuckwalla, respectively. Although the generality of the phenomenon suggests common evolutionary mechanisms, the genetic basis of extreme body size evolution on islands remains poorly understood. The powerful genomic resources and vast genetic toolkit developed for laboratory mice can be brought to bear on GI mice, making them a compelling system for understanding the genetic mechanisms underlying island gigantism. We used genetic mapping in a large F2 intercross between GI mice and WSB/Eij (WSB), a wild derived, inbred line with size representative of mainland mice, to identify 19 quantitative trait loci (QTL) regulating body weight and growth rate. Most of these QTL act in the first four weeks after birth. To assess the contribution of gene regulatory evolution to the organ-specific changes underlying GI gigantism we characterized differential gene expression between GI and WSB mice in three metabolically essential organs: the liver, gonadal fat depot and the hypothalamus. We obtained poly-A-selected transcriptome sequence reads from a large cohort of male and female GI and WSB mice at 4 weeks of age; additionally, for the liver, we obtained sequence reads at two weeks and embryonic day 16.5. Our analysis is the first to link multi-organ differential gene expression with island gigantism. Organ-specific, differentially expressed genes located within the QTL we found are especially promising candidates for the loci responsible for the rapid evolution of body size. Identification of these genes lays the foundation for future functional genetic assays that will serve to pinpoint the molecular genetic causes of size variation in nature

EXCESS OF AFRICAN ANCESTRY IN THE MHC REGION OF A RURAL BRAZILIAN ADMIXED POPULATION.

Kelly Nunes[1], Lilian Kimura[1], Juliana P Carnavalli[1], Regina C Mingroni-Netto[1], Bruce Weir[2], Diogo Meyer[1]

[1]University of São Paulo, Department of Genetics and Evolutionary Biology, São Paulo, Brazil, [2]University of Washington, Department of Biostatistics, Seattle, WA

The MHC (Major Histocompatibility Complex) is the genomic region with the highest concentration of immune function genes. Among these are the *HLA* genes (Human Leucocyte Antigens), whose proteins play a central role in the adaptive immune response, and are well documented targets of long-term balancing selection. However, few studies have identified recent selection signals (i.e. in last few thousand years) at *HLA* genes. This identification is possible through the analysis of admixed populations, whose genome is a mosaic of different ancestries. In the present study we test for shifts in ancestry proportions at *HLA* genes of admixed individuals, with respect to genomewide ancestries. To this end, we analyzed 110 unrelated highly admixed individuals, coming from rural communities in Brazil (Quilombos from Vale do Ribeira, in the State of São Paulo). Genotype data, obtained with the Affymetrix Axiom Human Origins 600K array, were used to estimate global and local ancestry via ADMIXTURE and RFMix, respectively. The population's genomewide ancestry is 43% African (AFR), 42% European (EUR), and 15% Native American (NAM). For each SNP, we computed the difference for a particular ancestry with respect to genomewide ancestry (Δ_{ANC}). We compared the mean Δ_{ANC} of the extended MHC region (encompassing all HLA loci) to the genomewide average, and tested for significance by a permutation approach (matching for size and/or SNP density to the MHC). We found an excess of African ancestry in the MHC region, when compared to the rest of the genome ($\Delta_{ANC-AFR}$= 0.103; p=0.001), and a deficit of Native American ancestry ($\Delta_{ANC-NAM}$ = -0.082; p=0.003). These results are consistent with recent selection favoring variants with African ancestry, but may also reflect biases of local ancestry inference in regions of high LD.

QUALITY CONTROL AND PHASING PIPELINES FOR THOUSANDS OF HIGH-COVERAGE WGS SAMPLES

Jared M O'Connell[1], Shankar Ajay[2], Sajani Swamy[2], Anthony J Cox[1], Michael A Eberle[2]

[1]Illumina, Computational Biology Group, Cambridge, United Kingdom, [2]Illumina, Bioinformatics, San Diego, CA

Until recently, it has been financially prohibitive to perform high-coverage whole-genome sequencing (WGS) on thousands of individuals. Hence the 1000 genomes project pioneered an approach combining low-coverage sequencing with genotype imputation to assay cohorts at scale. Recent drops in sequencing costs are now driving large-scale sequencing studies and soon high-coverage cohorts at a similar scale (and larger) to 1000 Genomes will become available. These high-depth studies will improve sensitivity and accuracy for rare variants as well as enabling improved calling of CNVs and SVs. At the same time, large cohorts of WGS individuals will allow us to use population based metrics to assess variant quality and infer very accurate haplotypes using population genetics inspired phasing routines. With this in mind, we report a number of analyses on a high-coverage cohort of 2,324 individuals of European descent, including approximately 100 mother-father-child trios.

Whole genome sequencing will identify many variants that have not been previously observed. Of these new variants, some may occur in parts of the genome that are more prone to errors and so developing ways to assess variant confidence will be vital for cohort studies. To assess variant quality, we evaluated a number of different filtering strategies for variant QC including machine learning approaches as well as masks derived from low-complexity regions and atypical depth. These relatively simple masking approaches proved surprisingly effective, identifying sets of variants enriched for divergence from Hardy-Weinberg equilibrium and with inflated numbers of Mendelian inconsistencies suggesting that they are effective at removing false positive variants.

The availability of high-depth sequence data will enable the creation of very high quality haplotype reference panels that can be used for imputation and informatic phasing. To assess the quality of reference panels generated from large numbers of unrelated samples sequenced to high-depth, we phased the well known NA12878/NA12877 samples together with our WGS samples using Beagle 4 (excluding other members of the pedigree). We then evaluated the phasing accuracy treating the pedigree derived haplotypes as the truth. The statistically phased haplotypes were extremely accurate with a switch error rate less than 0.25%. This is about three times lower than published results and demonstrates the efficacy of population based phasing as sample sizes increase, particularly without the additional complication of uncertain genotypes produced by low coverage sequencing data.

COMPARING STATISTICAL APPROACHES FOR BIOLOGICALLY BINNED VARIANTS FOR ASSOCIATION ANALYSIS OF LOW FREQUENCY VARIANTS

A Okula[1], J Wallace[1], J Leader[2], L Mirshahi[2], T Mirshahi[2], R Dewey [3], J Reid[3], J Overton[3], C O'Dushlaine[3], A Shuldiner[3], S Pendergrass[1], D Carey[2], D Ledbetter[2], M Ritchie[1,2]

[1]The Pennsylvania State University, Center for Systems Genomics, University Park, PA, [2]Geisinger Health System, Danville, PA, [3]Regeneron Genetics Center, Tarrytown, NY

The advent of next-generation sequencing technology has presented an opportunity for rare variant discovery. However, there are challenges in rare variant analysis as statistical power for detection is often low. Multiple analytic strategies have been developed for combining rare variants to increase statistical power to detect associations. BioBin is an automated tool developed to perform multi-level collapsing or binning of variants in a biological manner based on user designated features such as genes, regulatory regions and pathways. BioBin expands on collapsing methods that aggregate variants by using a flexible, biologically informed binning strategy using an internal biorepository, the Library of Knowledge (LOKI). LOKI combines over a dozen databases from the public domain including KEGG, Gene Ontology, and Protein Families. These databases provide variant details, regional annotations and pathway interactions used to generate bins of biologically-related variants, thereby increasing the power of any subsequent statistical test. In this study, we expand the framework of BioBin to incorporate a dispersion-based statistical test, SKAT, thereby providing the option of performing a unified collapsing and statistical rare variant analysis in one tool. Extensive simulation studies performed on gene-coding regions using 1000 Genomes CEU population data showed a BioBin-SKAT analysis to have greater power than BioBin-Logistic regression (standard burden test) in all simulated conditions, including variants influencing the phenotype in the same direction, a scenario where burden tests often retain greater power. The use of Madsen-Browning variant weighting increased power in the burden analysis to that equitable with BioBin-SKAT; but overall BioBin-SKAT retained equivalent or higher power under all conditions. BioBin-SKAT was applied to analyze LDL cholesterol in 7,354 unrelated individuals from the Geisinger MyCode biorepository sequenced at high depth after whole-exome capture by Regeneron Genetics Center. We binned variants with a frequency<0.05 using BioBin-SKAT to look for genes associated with LDL as a quantitative trait while adjusting for sex, age and BMI. As expected, we identified variants in known genes of lipid metabolism including APOB and PCSK9 with Bonferroni corrected p-values<0.01. This study demonstrates that BioBin-SKAT is a powerful tool for the identification of genes harboring low frequency variants for complex phenotypes.

ANALYSIS OF HLA LOCI IN NARCOLEPSY

Hanna M Ollila[1], Jean-Marie Ravel[1], Fang Han[2], Ling Lin[1], Juliette Faraco[1], Xiuwen Zheng[3], Giuseppe Plazzi[4], Yves Dauvilliers[5], Jacques Montplaisir[6], Steven J Mack[7], Michael Mindrinos[8], Emmanuel Mignot[1]

[1]Stanford University, Center for Sleep Sciences, Palo Alto, CA, [2]Peking University People's Hospital, Department of Pulmonary Medicine, Beijing, China, [3]University of Washington, Department of Biostatistics, Seattle, WA, [4]University of Bologna, Department of Biomedical and Neuromotor Sciences (DIBINEM), Bologna, Italy, [5]Gui-de-Chauliac Hospital, Department of Neurology, Montpellier, France, [6]University of Montréal, Department of Psychiatry, Montréal, Canada, [7]CHORI, Children's Hospital Oakland Research Institute, Oakland, CA, [8]Stanford University, Stanford Genome Technology Center, Palo Alto, CA

Introduction: Narcolepsy is caused by a lack of hypocretin (orexin) and it strongly associated with HLA class II DQA1*01:02~DQB1*06:02 (DQ0602) but also effects of other HLA-DQ alleles have been reported consistently across multiple ethnic groups. In narcolepsy, the destruction of hypocretin producing neurons is suggested to be triggered by autoimmunity. Especially upper airway infections such as pandemic 2009 influenza H1N1 are known triggers for narcolepsy.
Methods: We used a large sample of over 3,000 narcoleptics and 10,000 controls of European and Asian background and examined the effects of other HLA loci in narcolepsy.
Results: Conditional analysis after controlling for HLA-DR/DQ revealed a strong association with DPA1*01:03~DPB1*04:02 (DP0402) (OR=0.45 [0.38-0.55] $P=8.99*10^{-17}$) and DP0501 (OR=1.38 [1.18-1.61], $P=7.11*10^{-5}$). HLA class II independent associations were also seen across ethnic groups in the HLA class I region with HLA-A*11:01 (OR=1.32 [1.13-1.54], $P=4.92*10^{-4}$), B*35:03 (OR=1.96 [1.41-2.70], $P=5.14*10^{-5}$), and B*51:01 (OR=1.49 [1.25-1.78], $P=1.09*10^{-5}$). Similar effects were also seen after careful matching for HLA-DR/DQ with DP0402 (OR=0.51 [0.38-0.67], $P=1.01*10^{-6}$), DPA1*01:03~DPB1*04:01 (DP0401) (0.61 [0.47-0.80], $P=2.07*10^{-4}$) and predisposing effects of DPB1*05:01 in Asians (OR= 1.76 [1.34-2.31], $P= 4.71*10^{-5}$).
Conclusions: These effects may reflect modulation of autoimmunity, or indirect effects of HLA class I and DP alleles on response to viral infections such as influenza.
Support: Wake Up Narcolepsy, NIH NS23724, Sigrid Juselius Foundation, the Paivikki and Sakari Sohlberg Foundation, Orion Research Foundation and 973 Program 2015CB856405 and NSFC81420108002.

HUMAN DISEASE: FINDING THE BEST MOUSE MODEL

H Onda, A Anagnostopoulos, SM Bello, H Dene, M Knowlton, B Richards-Smith, CL Smith, M Tomczuk, LL Washburn, JT Eppig

The Jackson Laboratory, Mouse Genome Informatics, Bar Harbor, ME

The Human-Mouse Disease Connection (HMDC, www.diseasemodel.org) facilitates clinical and translational research by providing integrated access to human and mouse data, by assisting investigators in identifying appropriate mouse models for human disease and in locating repositories from which the models can be obtained. HMDC's highly-accessible platform allows search by human or mouse genes (symbols, names or IDs), gene lists, or genome location ranges, and using mouse phenotype terms or human disease (OMIM) terms. In addition, using Variant Call Format (.vcf) files, investigators can search HMDC for candidate disease genes in mouse by positional gene sequence variations in human. Results are returned as a grid displaying in each row matching human and mouse homologs, mouse phenotype terms, and OMIM diseases. Cells of the grid are active links leading to underlying data details. Additional views, accessed via tabs, display data in disease- or gene-focused tabular format. Links to the Mouse Genome Informatics (MGI, www.informatics.jax.org) web pages provide gene details for mouse and homology data for human, details of mouse phenotypes, publications on the mouse gene and human disease model, disease models summaries, and links to OMIM records (www.omim.org). Links to the International Mouse Strain Resource (IMSR, www.findmice.org) provide information and access to repositories holding specific mouse models. Exploration of data in HMDC can support hypothesis building through examination of potential candidates for human disease where phenotypic matches or overlapping genome locations suggest new disease gene candidates. Conversely, data for a known human disease gene may prompt the engineering of a mouse model where none currently exist. Supported by NIH grant HG000330.

DIVERSE MOLECULAR PROFILING MAPS OF SKELETAL MUSCLE REVEAL MECHANISTIC INSIGHTS ABOUT TYPE 2 DIABETES

Stephen C Parker[1], Jeroen R Huyghe[2], Michael R Erdos[3], Heikki Koistinen[4], Peter S Chines[3], Ryan Welch[2], Laura J Scott[2], D L Taylor[3], Brooke N Wolford[3], Hui Jiang[2], Xiaoquan Wen[2], Narisu Narisu[3], Timo Lakka[5], Richard M Watanabe[6,7], Karen Mohlke[8], Jaakko Tuomilehto[4], Michael Boehnke[2], Francis Collins[3]

[1]Univ. of Michigan, Computational Medicine & Bioinformatics and Human Genetics, Ann Arbor, MI, [2]Univ. of Michigan, Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, [3]National Institutes of Health, National Human Genome Research Institute, Bethesda, MD, [4]National Institute for Health and Welfare, Metabolism Group, Helsinki, Finland, [5]Univ. of Eastern Finland, Institute of Exercise Medicine, Kuopio, Finland, [6]Univ. of Southern California Keck School of Medicine, Preventive Medicine, Los Angeles, CA, [7]Univ. of Southern California Keck School of Medicine, Physiology and Biophysics, Los Angeles, CA, [8]Univ. of North Carolina, Genetics, Chapel Hill, NC

Type 2 diabetes (T2D) results from environmental and genetic factors interacting across time and multiple tissues. More than 90% of >100 variants associated with T2D and related traits through genome wide association studies (GWAS) occur in non-coding regions, suggesting a strong regulatory component to disease susceptibility. To understand the full spectrum of genetic variation and regulatory element usage in T2D-relevant tissues and across disease progression, the Finland United States Investigation of NIDDM Genetics (FUSION) Study obtained skeletal muscle biopsies from 278 well-phenotyped Finns with normal and impaired glucose tolerance and with T2D. We performed dense genotyping and imputation and constructed strand-specific mRNA-seq libraries and sequenced a total of 25.3 billion fragments (mean 91.3M read pairs per sample). We identified >13k genes with expression and/or splicing quantitative trait loci (e/sQTL) (5% FDR). We produced reference chromatin state maps across 30 cell types and reference transcriptome maps across 16 tissue types. Integrative analyses show that the genetic regulatory architecture of skeletal muscle specific gene expression is encoded in skeletal muscle stretch enhancers and not typical enhancers or stretch enhancers from unrelated tissues. Our eQTL analyses identify target genes and direction of effect of GWAS-identified risk alleles (including T2D). Comparison to the reference transcriptome map reveals the skeletal muscle expression selectivity of traits studied by GWAS. This rich data resource enables identification of diverse molecular processes involved in muscle based insulin resistance, changes in transcription with progression towards T2D, and reveals mechanistic insights about disease predisposition.

MEIOTIC ADAPTATION TO WHOLE GENOME DUPLICATION

Kirsten Bomblies

Harvard University, Organismic and Evolutionary Biology, Cambridge, MA

Whole genome duplication (WGD) is a major factor in evolution, adaptation and speciation, and WGD events predate some of the major radiations in eukaryotes. Yet by doubling the number of homologous chromosomes that can pair and recombine, WGD challenges reliable meiotic chromosome segregation. Neopolyploids often show defects in meiotic chromosome segregation and low fertility. However, numerous polyploid species in nature, particularly among plants, and most of these have stable, often diploid like chromosome segregation, showing that evolution can overcome the early problems that face polyploids. The molecular basis of this stabilization remains mysterious. We use Arabidopsis arenosa, an outcrossing relative of A. thaliana with extant diploid and polyploid populations, to better understand the genic basis of adaptation to WGD. Using a genome scan, we find eight functionally interacting meiosis genes,¬ which encode proteins that together coordinate chromosome pairing, synapsis, and the number and distribution of chiasmata. We show evidence for one gene that the derived allele has a strong functional consequence in tetraploid meiosis. We hypothesize that these genes represent a polygenic solution to WGD-associated chromosome segregation challenges, and that likely the derived tetraploid alleles condition increased reduced crossover numbers, perhaps by increasing crossover interference.

THE EVOLUTION OF PRDM9 MOTIFS IN HUMANS AND MICE

Robert W Davies[1], Afidalina Tumian[2], Simon Myers[1,3]

[1]Oxford University, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]International Islamic University Malaysia, Department of Electrical and Computer Engineering, Kuala Lumpur, Malaysia, [3]Oxford University, Department of Statistics, Oxford, United Kingdom

In both mice and humans, the protein PRDM9 binds to DNA at specific locations in the genome to initiate the recombination process. However, if an individual hotspot has a heterozygous mutation within the motif that reduces SNP binding efficiency, PRDM9 will preferentially bind to the strong motif, and through gene conversion, may replace the strong motif with the weak one. This unequal exchange of DNA leads to transmission distortion and an increased rate of fixation for variants which disrupt PRDM9 binding sites. We used this principle to identify the binding targets of PRDM9 alleles, often existing millions of years ago, by searching for short sequences of DNA that were lost faster on a given lineage than others. For both mice and primates, we find PRDM9 motifs closely matching known active motifs, confirming the power of this approach, along with hundreds of novel motifs which imply a universally rapid evolution of the recombination landscape. We used motif loss sites to build ancient recombination maps and compared these to each other and present day maps, showing factors which influence broad scale recombination rates are stable over millions of years, as well as specific genomic regions with more pronounced deviations consistent with broad scale genomic rearrangements, including the ancestral fusion event leading to human chromosome 2. Most identified motifs show evidence of GC biased gene conversion. Fitting this conversion signal using a mixture of exponentials implies both crossover and non-crossover events likely show GC-bias, and we obtain estimates for their tract lengths, as well as their relative proportions, results which are similar in between human and mice.

BMD LOCI UNDERLIE DEVELOPMENTAL DETERMINATION OF ETHNIC DIFFERENCES IN SKELETAL FRAGILITY ACROSS POPULATIONS DUE TO SELECTION PRESSURES

Carolina Medina-Gómez[1], Alessandra Chesi[2], Denise H Heppe[1], Babette S Zemel[3], Jia-Lian Yin[1], Heidi J Kalkwarf[4], Albert Hofman[1], Joan M Lappe[5], Andrea Kelly[3], Manfred Kayser[6], Sharon E Oberfield[7], Vicente Gilsanz[8], André G Uitterlinden[1], John A Shepherd[9], Vincent W Jaddoe[1], Struan F Grant[2], Oscar Lao*[6,10], Fernando Rivadeneira*[1]

[1]Erasmus University Medical Center, The Generation R Study Group, Rotterdam, Netherlands, [2]Children's Hospital of Philadelphia, Division of Human Genetics, Philadelphia, PA, [3]Children's Hospital of Philadelphia, Division of GI, Hepatology, and Nutrition, Philadelphia, PA, [4]Cincinnati Children's Hospital Medical Center, Department of Pediatrics, Cincinnati, OH, [5]Creighton University, Department of Medicine, Omaha, NE, [6]Erasmus University Medical Center, Department of Forensic Molecular Biology, Rotterdam, Netherlands, [7]Columbia University, Division of Pediatric Endocrinology Diabetes and Metabolism, New York, NY, [8]Children's Hospital of Los Angeles, Department of Radiology, Los Angeles, CA, [9]UCSF, Department of Medicine, San Francisco, CA, [10]CNAG, Centro Nacional de Análisis Genómico, Barcelona, Spain

*authors contributed equally

Bone mineral density (BMD) is a highly heritable trait used both for the diagnosis of osteoporosis in adults and to assess bone health in children. Ethnic differences in BMD have been documented, with markedly higher levels in individuals of African descent, which partially explain disparity in osteoporosis risk across populations. To date, 63 independent genetic variants have been associated with BMD in adults of Northern-European ancestry. We demonstrate that 61 of these variants are predictive of BMD early in life by studying their compound effect within two multiethnic pediatric cohorts. Furthermore, we show that within these cohorts and across populations worldwide the frequency of those alleles associated with increased BMD is systematically elevated in individuals of Sub-Saharan African ancestry. The amount of differentiation in the BMD genetic scores among Sub-Saharan populations together with neutrality tests, suggest that these allelic differences are compatible with the hypothesis of selective pressures acting on the genetic determinants of BMD, providing a new example of polygenic adaptation in a human trait.

POPULATION STRUCTURE IN AFRICAN-AMERICANS.

Soheil Baharian[1], Maxime Barakatt[1], Christopher R Gignoux[2], Suyash Shringarpure[2], Brian K Maples[2], Eimear E Kenny[3], Carlos D Bustamante[2], Melinda C Aldrich[4], Simon Gravel[1]

[1]McGill University, Department of Human Genetics, Montreal, Canada, [2]Stanford University, Department of Genetics, Stanford, CA, [3]The Charles Bronfman Institute for Personalized Medicine, Department of Genetics and Genomic Sciences, New York, NY, [4]Vanderbilt University, Department of Thoracic Surgery, Nashville, TN

We present a detailed population genetic study of 3 African-American cohorts comprising over 3000 genotyped individuals across US urban and rural communities: two nation-wide longitudinal cohorts, and the 1000 Genomes ASW cohort. Ancestry analysis reveals a uniform breakdown of continental ancestry proportions across regions and urban/rural status, with 79% African, 19% European, and 1.5% Native American/Asian ancestries, with substantial between-individual variation. The Native American ancestry proportion is higher than previous estimates and is maintained after self-identified Hispanics and individuals with substantial inferred Spanish ancestry are removed. This supports direct admixture between Native Americans and African Americans on US territory. Local ancestry patterns and variation in ancestry proportions across individuals are broadly consistent with a single African-American population model with early Native American admixture and ongoing European gene flow in the South. The size and broad geographic sampling of our cohorts enable detailed analysis of the geographic and cultural determinants of finer-scale population structure. Recent identity-by-descent analysis reveals fine-scale geographic structure consistent with the routes used during slavery and in the great African-American migrations of the twentieth century: east-to-west migrations in the south, and distinct south-to-north migrations into New England and the Midwest. These migrations follow transit routes available at the time and are in stark contrast with European-American relatedness patterns.

BETTER, FASTER, STRONGER: MIXED MODELS AND PCA IN THE YEAR 2015

Alkes L Price[1,2,3]

[1]Harvard T.H. Chan School of Public Health, Epidemiology, Boston, MA, [2]Harvard T.H. Chan School of Public Health, Biostatistics, Boston, MA, [3]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA

Linear mixed models (LMM) and principal components analysis (PCA) are widely used tools for analyzing genetic data and correcting for population stratification in genome-wide association studies (GWAS). We describe recent improvements to these methods, focusing on both speed and power. Previous LMM and PCA methods require building a genetic relationship matrix at time cost O(MN^2) (where M = #SNPs and N = #samples), but the new methods avoid this computation. First, the BOLT-LMM method uses conjugate gradient and variational iterations to compute a retrospective score statistic in a small number of O(MN) iterations. The method models non-infinitesimal genetic architectures via a Bayesian mixture prior on marker effect sizes, attaining higher power than previous methods that model infinitesimal (Gaussian) genetic architectures. Theory, simulations, and application to real phenotypes show that the boost in power increases with cohort size, making BOLT-LMM appealing for GWAS in large cohorts. Second, the FastPCA method computes top PCs using a random low-rank matrix approximation algorithm, generalizing the power iteration approach. The method is extremely accurate in simulated and real data sets, and requires only O(MN) time. Applying the method to a very large European American data set, the top PCs distinguish NW European, SE European, Ashkenazi Jewish, Eastern European and Irish ancestry. Assessment of unusual population differentation along these top PCs detects both known and novel signals of natural selection at genome-wide significance.

FREQUENCY OF MOSAICISM POINTS TOWARDS MUTATION PRONE EARLY CLEAVAGE CELL DIVISIONS

Chad Harland[1], Carole Charlier[1], Latifa Karim[2], Nadine Cambisano[2], Wouter Coppieters[2], Michel Gerorges[1]

[1]University of Liège, Unit of Animal Genomics, GIGA & Faculty of Veterinary Medicine, Liège, Belgium, [2]University of Liège, GIGA Genomics Core Facility, Liège, Belgium

Next generation sequencing (NGS) of nuclear families allows for the direct identification of de novo mutations. The use of parent-offspring trios has resulted in estimation of the average de novo mutation rate of $\sim 10^{-8}$ in humans, and of a ~4-fold higher mutation rate in the male than in the female germline.
We are sequencing the genome of > 750 related bovine that should allow us to reliably identify de novo mutations in at least two sperm cells of 100 bulls and two oocytes of 100 cows. We herein report initial results based on the analysis of (i) three four-generation pedigrees comprising the four grand-parents and two parents of a cow (proband) with five of her half-sib offspring, and (ii) two three-generation pedigrees comprising the two parents of a bull (proband) with, respectively, 11 and five half-sib offspring. All animals were sequenced at ≥ 20-fold depth. We sequenced blood of cows and sperm of bulls, except for one of the three-generation pedigrees for which blood and sperm were sequenced for the bull proband.
Our major findings at this point are (i) that ~30% of the de novo mutations detected in sperm and ~20% of the de novo mutations detected in female blood (absent in the parents), occurred during the development of the proband rather than being inherited from the parents, and (ii) that ~34% of de novo mutations transmitted by a sperm cell are detectable in the sperm of the father and ~52% of de novo mutations transmitted by an oocyte are detectable in the blood of the mother. In human studies, the first type of de novo mutations would have been erroneously assigned to either the paternal or maternal germline, while the second type would have been ignored.
We show by simulation that the observed high degree of germline and somatic mosaïcism cannot be reconciled with the general assumption that all cell divisions are equally prone to de novo mutations. It rather suggests that a large proportion of de novo mutations occur during early embryonic development.
Our findings have important implications with regard to proper estimation of de novo mutation rate from NGS data and understanding of the molecular mechanisms underlying the mutational process. Latest results with regards to mutation rate, sex effect and mutation types will be presented.

INTER-INDIVIDUAL VARIATION IN EPIGENETIC MARKS BETWEEN HUMAN INDUCED PLURIPOTENT STEM CELL LINES

Angela Goncalves[1], Natsuhiko Kumasaka[1], Andrew Knights[1], Francesco Casale[2], Jose Garcia-Bernardo[1], Daniel Gaffney[1], on behalf of the HipSci Consortium[1]

[1]Wellcome Trust Sanger Institute, Bioinformatics, Cambridge, United Kingdom, [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom

Pluripotent stem cells vary at molecular and functional levels, particularly in their ability to differentiate into alternative cell lineages. Previous work in our lab and others has highlighted that the transcriptome of IPSCs appears to be remarkably homogeneous. To elucidate the origin of the molecular and functional variability between iPS cells, the HIPSCI project is generating a large collection of human iPS cells and collecting data on differentiation efficiency, genetic sequence, gene expression and key epigenetic marks.

Here we describe the analysis of Illumina 450k methylation array data from 176 IPSC lines derived from 136 healthy donors, and ChIP-seq for H3K4me3, H3K27ac and H3K27me3 histone modifications in 44 IPSC lines derived from 39 healthy donors. We employed variance component analysis to estimate the sources of variation in the methylation assay, attributing variability to donor, line, gender, donor age and other technical effects such as culture media, passage rate or batch. We found that on average most of the variance observed in methylation probe levels can be explained by genetic differences between individuals. Next, we considered our methylation and histone modification data as quantitative traits for genetic mapping, and performed association analysis on individual trait and combined measures. Individually, for the methylation data we found QTLs for over 2000 genes at an FDR of 5%. For the histone modification data we employed a unified probabilistic model recently developed by our group that combines allele-specific and population level sequencing data, to boost the power for association detection. Despite the small sample size, we also detected hundreds of genome-wide significant associations.

Our results reveal extensive genetic effects on iPSC epigenetic traits. Further analysis will reveal how these genetically mediated changes to IPSCs epigenomes drive variation in important stem cell properties including maintenance of pluripotency, cell growth and differentiation. Our results will also highlight how epigenetic changes are propagated to alternative cell layers, including the transcriptome and proteome.

LINKING IMMUNE RESPONSIVE REGULATORY VARIATION AND POPULATION ADAPTATION TO PATHOGEN PRESSURE

Maxime Rotival*[1], Helene Quach*[1], Eddie Loh*[1], Julien Pothlichet[1], Etienne Patin[1], Guillaume Laval[1], Nora Zidane[1], Christine Harmant[1], Marie Lopez[1], Geert Leroux-Roels[2], Frédéric Clément[2], Jean-François Deleuze[3], Lluis Quintana-Murci[1]

[1]Human Evolutionary Genetics Unit CNRS URA3012, Institut Pasteur, Paris, France, [2]CEVAC, Ghent University Hospital, Ghent, Belgium, [3]Centre National de Génotypage, CNG, Evry, France

The immune system is one of the most adaptive systems that exist, as immunity genes recurrently appear as privileged targets of natural selection. Still, little is known about the relationship between genotypic and phenotypic diversity in immune responses in the human population, and how immunological mechanisms have been subject to natural selection. Here, we used RNA-sequencing, coupled with genome-wide SNP data and whole exome sequencing, to define the responses of primary monocytes from 200 individuals of European- and African-descent to four immune stimuli. We generated 970 transcriptional profiles, obtained in resting monocytes and after exposure to three TLR ligands – Pam3CSK4 (TLR1/2), LPS (TLR4), and R848 (TLR7/8) – and Influenza A virus (IAV). Using coexpression analyses, we showed that the activation of the three TLR pathways is primarily associated with a strong inflammatory response driven by the TFs C/EBP and TATA, while IAV infection is associated with a strong, specific activation of HLTF targets, together with an SRY-driven up-regulation of MHC class II genes. In addition, both IAV and the activation of TLR7/8 pathway trigger IRF-driven antiviral responses. Interestingly, these co-expression modules exhibit strong differences in the amplitude of their responses between Africans and Europeans. We next investigated the genetic basis of these responses, and found over 7650 genes associated with a cis-eQTL in at least one condition, 46% of which were response eQTLs (reQTL, showing a context-specific effect in at least one population). Many infection-specific population differences in gene expression appeared to result from strong differences in the allelic frequency of reQTLs between Africans and Europeans, some of which presented strong signals of positive selection. We also detected 4692 genes associated with trans eQTLs, with some co-expression modules being under the control of a response trans-eQTL. Finally, we identified several cases of allele-specific-expression (ASE), the magnitude of which was found to be context-dependent. Together, this study increases our understanding of how regulatory variation influences the heterogeneity of immune response phenotypes, including ASE, and reveals specific responses to immune stimuli that differ between human populations and have been crucial for our past adaptation to pathogen pressure.

GENETIC ANALYSIS OF PARALLEL LOCAL ADAPTATION TO SERPENTINE AND MINE SOILS IN *MIMULUS*

Kevin M Wright*[1], Jessica Selby*[2], Annie Jeong [2], Uffe Hellsten[3], Daniel S Rokhsar[3], John H Willis[2]

[1]Harvard University, Cambridge, MA, [2]Duke University, Durham, NC, [3]DOE Joint Genome Institute, Walnut Creek, CA

*Authors contributed equally

A major challenge in 21st century biology is to understand how organisms adapt to complex and often unpredictable environments. Evolutionary ecologists have studied plant adaptation to extreme edaphic environments such as serpentine soils and heavey metal contaminated mine tailings for decades and this classic work provides some of our best examples of "natural selection in action." These extreme soils impose such strong selection that plant populations occurring on serpentine/mine soils and adjacent "normal" soils are often locally adapted over a scale of meters despite substantial gene flow. Throughout western North America local populations of the yellow monkeyflower *Mimulus guttatus* have repeatedly adapted to patches of serpentine soils or toxic copper mine tailings. Using a combination of reciprocal transplants in the field and lab, QTL mapping, physiological experiments, and population genomic approaches, we are beginning to identify the most important evolutionary genetic changes that have enabled this plant species to survive and reproduce on serpentine soils or mine tailings throughout its range. Serpentine soils and copper mine soils are patchily distributed and vary substantially in their physiochemical properties, and it is not known whether widespread species, such as *M. guttatus*, repeatedly adapt to different patches via the same or different molecular mechanisms. Even if the same gene is repeatedly used, is this due to new mutations or repeated use of standing variation? Are serpentine or Cu mine adapted alleles and pathways selectively equivalent on soils from different serpentine or mine regions, or uniquely suited to each particular habitat? This talk will highlight some of our initial discoveries that answer some (but not all!) of these basic questions about evolutionary plant solutions to ecological challenges.

MARSUPIAL-SPECIFIC GENOMIC IMPRINTING IN THE OPOSSUM, *MONODELPHIS DOMESTICA*

Andrew G Clark[1], Xu Wang[1], Kory C Douglas[2], Paul B Samollow[2]

[1]Cornell University, Molecular Biology and Genetics, Ithaca, NY, [2]Texas A&M, Veterinary Integrative Biosciences, College Station, TX

Genomic imprinting is largely restricted to therian mammals, and while we know a great deal about imprinting in the placenta of eutherian mammals, information on marsupials is fragmentary. A motivation to explore marsupial imprinting is the radical difference in the tissues supporting fetal growth, and the expectation that the evolution of genomic imprinting relates to the function of these tissues. We profiled genome-wide allele-specific expression (RNA-seq), histone modifications and DNA methylation in fetal brain and extra-embryonic membranes from reciprocal crosses of two semi-inbred opossum lines, providing an unbiased survey of parent-of-origin effects. Among 68 genes known to be imprinted in eutherians (and having an opossum ortholog), 52 were covered with informative SNPs to score allelic expression. Only 3 (<6%) were found to be imprinted in opossum, and 48 display biallelic expression, reflecting a striking lack of conservation of imprinting status. We also discovered and validated 8 novel imprinted genes that are not known to be imprinted in any other species, 5 of which are protein-coding, and 3 are non-coding lincRNAs. Two of the protein-coding imprinted genes have a 1-to-1 ortholog in eutherians, and the other 3 coding genes experienced a gene family expansion in opossum, with one shared paralog in eutherians. We find no evidence for any homology between the 3 imprinted opossum lincRNAs and any eutherian mammal genome. However, these lincRNAs are present and are highly conserved within marsupials and in non-mammalian vertebrates, including chicken. We discovered and confirmed promoter DMRs for all 5 protein-coding genes, suggesting their imprinting is regulated by DNA methylation. The lincRNA genes lack a DMR, and chIP-seq results suggest they are regulated by histone modifications. These results contrast with our previous work on the ~400 opossum X-linked genes subjected to imprinted X-chromosome inactivation (XCI), where DNA methylation does not play a role, but instead H3K27me3 is associated with the paternal allele silencing. Unlike eutherian mammals, marsupials achieve imprinting and XCI through distinct mechanisms as assessed by global comparison of histone marks between imprinted and XCI genes. Our study provides the first comprehensive catalog of parent-of-origin expression status in a marsupial, opens the door to mechanistic analysis of marsupial-specific imprinted genes and sheds light on both the regulation and evolution of genomic imprinting in mammals.

A FINE-SCALE MAP OF RECOMBINATION RATES AND HOTSPOTS IN THE ZEBRAFINCH GENOME

Sonal Singhal*[1], Ellen Leffler*[2], Isaac Turner[3], Oliver Venn[3], Alva Strand[1], Brian Raney[4], Qiye Li[5], Chris Balakrishnan[6], Simon Griffith[7], Gil McVean[3], Molly Przeworski[1]

[1]Columbia University, Biology, New York, NY, [2]University of Chicago, Biology, Chicago, IL, [3]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [4]UC Santa Cruz, CBSE, Santa Cruz, CA, [5]Beijing Genomics Institute, China National Genebank, Guangdong Sheng, China, [6]East Carolina Institute, Biology, Greenville, NC, [7]Macquarie University, Biology, Sydney, Australia

Recombination is a major force shaping genetic variation in populations, yet, outside of a few model organisms, we know little about its underlying mechanisms. One pattern to emerge is that, in most but not all species, meiotic recombination is concentrated in hotspots, or kilobase segments that experience 10 to 100 times more recombination than the background. In mammals but not yeast, these hotspot regions evolve quickly across genetically differentiated groups, due to changes in the binding site of the zinc finger protein, PRDM9.

Although PRDM9 has a pivotal role in controlling recombination in mammals, many species do not have PRDM9. Of special interest are birds, because along with having compact and largely syntenic genomes across millions of years of evolution, no PRDM9 ortholog has been identified in any bird genome sequenced to date. This observation raises a few basic questions about recombination in bird genomes: in the absence of PRDM9, what are broad scale recombination patterns across the genome? Is recombination concentrated in hotspots? If there are hotspots, how quickly do they evolve? To address these questions, we generated whole genome resequencing data from wild populations of two closely-related species of birds (the zebrafinch *Taeniopygia guttata* and the long-tailed finch *Poephilia acuticauda*). By inferring fine-scale recombination maps from linkage disequilibrium data and comparing maps for both species, we show the following: (1) at the broad scale, and in contrast to what is seen in mammals and *Drosophila*, recombination in the genome is localized largely to the telomeres, with intervening recombination deserts, (2) these finches do appear to have hotspots, which evolve much more slowly than they do in mammals, and (3) recombination rates peak near the transcription start site, similarly to what is observed in other vertebrate species without PRDM9 and yeast. Beyond the insights into birds, this study serves as a template for how we can collate patterns of diversity and recombination to explore the mechanisms that generate and structure genetic variation across a much broader taxonomic swath than has been previously accessible.
*equal contribution from authors

GENOME-WIDE ASSOCIATION AND LOCAL ANCESTRY ANALYSES OF HIGH-ALTITUDE ADAPTATIONS IN TIBETANS

Anna Di Rienzo[1], Choongwon Jeong[1], Buddha Basnyat[2], Geoff Childs[3], Sienna R Craig[4], Maniraj Neupane[5], David B Witonsky[1], John Novembre[1], Cynthia M Beall[6]

[1]University of Chicago, Human Genetics, Chicago, IL, [2]Patan Hospital, Oxford University Clinical Research Unit, Kathmandu, Nepal, [3]Washington University, Anthropology, St. Louis, MO, [4]Dartmouth College, Anthropology, Hanover, NH, [5]Mountain Medicine Society of Nepal, Kathmandu, Nepal, [6]Case Western Reserve University, Anthropology, Cleveland, OH

Indigenous human populations in Tibet show distinct physiological traits from those of acclimatized lowlanders, e.g. unelevated hemoglobin concentration (Hb) and extremely low arterial blood oxygen saturation level (SaO2). Previous studies of Tibetans identified several candidate genes, *EGLN1*, *EPAS1* and *PPARA*, which harbor signatures of positive selection. However, the genetic basis of Tibetan high-altitude physiology remains poorly understood.
We performed a GWAS of Hb and SaO2 in a group of 880 ethnic Tibetan women born and raised at high altitudes in Nepal. Both phenotypes were controlled for known covariates. No genome-wide significant association was found after Bonferroni correction. However, *EPAS1* SNPs are associated with Hb (linear mixed model (LMM) $p \geq 2.27 \times 10\text{-}5$) and *HIF1A* SNPs are associated with SaO2 (LMM $p \geq 2.72 \times 10\text{-}4$), confirming a role for oxygen homeostasis in Tibetan adaptations. Next, we conducted a local ancestry analysis in a subset of 337 unrelated Tibetans, to find loci with excess high-altitude ancestry represented by the Sherpa. Both *EPAS1* and *EGLN1* show marked enrichment in high-altitude ancestry: 73.1% (+6.1 SD) for *EPAS1* and 60.1% (+3.3 SD) for *EGLN1*. The genome-wide mean high-altitude ancestry is 45.1% with SD = 4.6%. To refine our local ancestry signals, we scanned the Tibetan genome for signals of adaptations and focused on those that fall in regions of excess local ancestry. These regions identify strong signals in genes, including *PPARA, SLCO1B1/SLCO1B3, OCA2/HERC2, HFE*, and *PON1*, which are known to play a role in hypoxia response, or in blood traits or in adaptations to local environments in East Asians. Enrichment analyses of the genes overlapping selection and local ancestry signals using DAVID v6.7 identified significant enrichments in immune-mediated disease pathways (e.g. asthma, type I diabetes), in cardiovascular disease classes, in immunity and defense, lactation and mammary development biological processes, and in genes expressed in peripheral blood and liver. These findings suggest that combining phenotype mapping, selection scans and local ancestry analyses may shed new light on the genetics and physiology of high altitude adaptations and on the biology of hypoxia response.

THE EVOLUTION OF RATTLESNAKE VENOM

Noah L Dowell, Matthew W Giorgianni, Sean B Carroll

University of Wisconsin and Howard Hughes Medical Institute, Cellular and Molecular Biology, Madison, WI

The mechanisms underlying the origin of biological novelties are not well understood. Prey-killing venoms have evolved independently multiple times across the animal kingdom. Approximately half of the 6000 snake species are venomous but the evolutionary origins of snake venom proteins are unclear. Within rattlesnakes, there appears to have been a relatively rapid evolution of at least two distinct venom types. *Crotalus atrox*, a large bodied generalist, possesses hemorrhagic venom with a relatively high LD_{50} while *Crotalus scutulatus*, a small-bodied snake possesses a more potent, neurotoxic venom with a low LD_{50}. Using a comparative genomics approach, we sought to identify the genetic mechanisms through which these divergent venom types evolved. Sequence analysis of *C. atrox* and *C. scutulatus* venom loci revealed gene gains through duplication events and gene losses via mutation but the unexpected retention of the genetic capacity to produce a neurotoxic venom in *C. atrox*. Phylogenetic analysis combined with the transcriptome data allowed us to infer the order of molecular events that lead to the evolution of neurotoxic venom. First, gene duplication produced multiple copies of toxin genes (acidic and basic phospholipase A2s) with appropriate transcriptional regulatory elements, and subsequent neofunctionalization of an acidic-Pla2 yielded a neurotoxin in *C. scutulatus*. In contrast, in *C. atrox*, pseudogenization is likely to have erased the acidic-Pla2 while the basic-Pla2s remain. These results indicate that there has been dynamic evolution of rattlesnake venom genes over a relatively short time-scale.

AN EARLY MODERN HUMAN WITH A RECENT NEANDERTAL ANCESTOR

Qiaomei Fu*[1,2,3], Mateja Hajdinjak*[3], Silviu Constantin[4], Oana T Moldovan[5], Swapan Mallick [2,6,7], Pontus Skoglund[2], Nick Patterson[6], Iosif Lazaridis[2], Birgit Nickel[3], Bence Viola[3], Kay Pruefer[3], Matthias Meyer[3], Janet Kelso[3], David Reich[2,6,7], Svante Pääbo[3]

[1]Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, IVPP, CAS, Beijing, China, [2]Harvard Medical School, Genetics, Boston, MA, [3]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany, [4]"Emil Racovita" Institute of Speleology, 010986 Bucharest 12, Romania, [5]"Emil Racovita" Institute of Speleology, Cluj Branch, 400006 Cluj-Napoca, Romania, [6]Broad Institute of MIT and Harvard, Cambridge, MA, [7]Harvard Medical School, Howard Hughes Medical Institute, Boston, MA

Neandertals became extinct in Europe about 40,000 years ago but contributed between 1 and 2 % of the DNA to present-day people in Europe and Asia. In order to better understand the interactions between modern and archaic humans we have studied a number of early modern humans from Europe. Among these is a mandible that is ~37,000-42,000 years old from Peştera cu Oase, Romania. Although the specimen contains small amounts of endogenous DNA we have used DNA capture to isolate variable sites that are informative with respect to its relationship to present-day humans and Neandertals.

We find that Neanderthals share substantially more derived alleles with the Oase mandible than with present-day people in Eurasia. We estimate that the proportion of Neandertal DNA in its genome is between 8.4 and 11.3% and observe three genomic segments that are over 50 cM in size and several that are over 5 cM, suggesting that the Neanderthal contribution to this individual was recent. From the size distribution of these segments we estimated that this individual had a Neandertal ancestor four to six generation back. Thus, the admixture between modern humans and Neanderthals was not limited to the Middle East or to the first ancestors of present-day people to leave Africa; it occurred in Europe as well.

INSIGHTS INTO RECOMBINATION AND SEX CHROMOSOME EVOLUTION FROM WHOLE-GENOME SEQUENCING OF PLATYPUS

Hilary C Martin[1], Elizabeth Batty[1], Julie Hussin[1], Portia Westall[2], Tasman Daish[3], Tom Grant[4], Rory Bowden[1], Frank Grutzner[3], Jaime Gongora[2], Peter Donnelly[1,5]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [2]Faculty of Veterinary Science, University of Sydney, Sydney, Australia, [3]School of Molecular and Biomedical Science, University of Adelaide, Adelaide, Australia, [4]University of New South Wales, Sydney, Australia, [5]Department of Statistics, University of Oxford, Oxford, United Kingdom

As a member of the most basal mammalian group, the platypus is ideal for studying the evolution of different biological processes. We have sequenced 61 platypus samples from 14 rivers from across the species' range between 16°S and 43°S in eastern Australia, including Tasmania. This is the first large-scale whole-genome sequencing (WGS) project to look at diversity in this species. One of the most interesting aspects of platypus biology is the sex chromosome system: there are five X and five Y chromosomes which, in male meiosis, form a multivalent chain connected by nine pseudoautosomal regions (PARs). It is thought that the chain evolved from the X5Y5 end, progressing through serial translocations to the X1Y1 end, and thus that the different pairs are of varying ages. This makes the platypus a unique animal in which to test the theoretical predictions about the evolution of PARs, which may not be possible in the highly differentiated sex chromosomes of other species. The Y chromosomes are missing from the reference genome, since it was made from a female. However, we can recover Y sequences by *de novo* assembly of male-specific reads and from examining sequence divergences where they map to the female reference. Through the latter, we find evidence for Y-specific sequences similar to their gametologs on the Xs. These regions appear to range from a few hundred base pairs to over 100kb long and to localize at the PAR boundaries. Some seem to be segregating between or within river systems, possibly suggesting that the divergence of X and Y chromosomes is ongoing. In addition to this work on the sex chromosomes, we are also investigating the population structure and history of the species, to inform conservation effects, and estimating fine-scale recombination rates. We cannot find clear evidence that platypus have *PRDM9*, the product of which controls hotspot placement in primates and mice, and we anticipate that this study will shed light on the evolution of the important process of meiotic recombination in mammals.

BUILDING SUPERMODELS: A REVIEW OF EMERGING COMPUTATIONAL AVATARS FOR PRECISION MEDICINE

Sherry-Ann Brown

Mayo Clinic, Department of Medicine, Rochester, MN

Precision medicine tailors prevention, prognosis, diagnosis, and therapeutics for each individual patient. A subset of precision medicine is systems medicine, which leverages systems biology for clinical application. This review catalogues several biomathematical models that have been developed in systems biology and can be or have been translated to clinical medicine. Many of these computational models can simulate biomedical properties of cancer, heart, and brain cells, personalized for each patient. Therapeutic strategies and predictions for safer and more efficacious drug delivery are studied and influence cancer and heart care. The models are created by groups in medical science academia and industry, as well as the International Business Machines Corporation (IBM). A large proportion of the in silico exemplars focus on delivery, efficacy, and adverse effects of chemotherapeutics, such as doxorubicin, which is used widely to treat diverse cancers. Adverse drug effects often lead to the withdrawal of various pharmacologics from cancer care in response to toxic effects on the heart or other organs. As an example to counteract this, a mathematical model predicts the optimal mode of doxorubicin delivery – bolus, continuous infusion, or liposomes – to maximize antitumor efficacy and minimize adverse effects. Consistent with model predictions, liposomal delivery has subsequently been studied in a number of clinical trials, which have shown superior toxicity profiles compared to standard non-liposomal delivery. A consortium, OncoTrack, synthesizes such mathematical modeling, omics, and other systems biology approaches, to produce a personalized molecular imprint of a patient's cancer and its microenvironment, aimed at predicting the right diagnostics and prescriptions. Further, IBM has harnessed natural-language processing and cognitive computing to create clinical decision support tools, e.g., the preliminary MD Anderson Oncology Expert Advisor, combining guidelines, recommendations, and primary scientific literature with a virtual advisor trained by medical experts to guide individualized care. The emergence of these multidisciplinary models, consortia, and platforms heralds an era in which comprehensive computational avatars ('SuperModels') will optimize patient care in precision medicine. Building these SuperModels will foster global collaboration among clinicians, scientists, technology industry specialists, institutes, and patient groups, with central repositories for vast amounts of appropriately channeled systems medicine data. Solidifying the place of these customized and integrative SuperModels in Precision Medicine will facilitate the right care for the right patient at the right time.

ABERRANT ASTROCYTE MATURATION CONTRIBUTES TO RETT SYNDROME PATHOGENESIS

Natasha L Pacheco[1,3], Leanne M Holt[2,3], Michelle L Olsen[1,2,3]

[1]University of Alabama at Birmingham, Genetics, Genomics and Bioinformatics Theme, Birmingham, AL, [2]University of Alabama at Birmingham, Neuroscience Theme, Birmingham, AL, [3]University of Alabama at Birmingham, Cell, Developmental, and Integrative Biology, Birmingham, AL

Rett syndrome (RTT) is an X-linked neurodevelopmental disorder caused by mutations in the transcriptional regulator MeCP2 which affects 1:10,000 females worldwide annually. RTT is characterized by having apparently normal development until 6-18 months, when a progressive decline in motor and language functions begins and breathing abnormalities and seizures present. Astrocytes, the most abundant cell type in the CNS, have recently been shown to express MeCP2. Importantly, postnatal re-expression of MeCP2 in astrocytes in globally *Mecp2*-deficient mice ameliorated many RTT disease symptoms, indicating that deficiencies in astrocytic function contribute to the pathophysiology of RTT. However, the causative mechanisms are currently unknown. Preliminary data demonstrates significant differences in expression of several key astrocytic genes associated with astrocyte development and maturation, leading us to conclude that astrocyte maturation is abnormal in the Rett brain. We hypothesize that the loss of MeCP2 in astrocytes results in aberrant astrocyte developmental gene expression and maturation. Given the broad transcriptional regulatory role of MeCP2, we predict that many astrocytic genes are dysregulated during astrocyte development. To test this prediction, we have utilized RNA-Seq analysis to examine global gene expression changes in whole cortical tissue from symptomatic *Mecp2*-deficient mice compared to wild-type (WT) littermate controls. We have identified more than 600 genes as significantly, differentially expressed. Forty of these genes have been identified as astrocyte-enriched compared to other CNS cell types. Future directions will examine gene and protein expression changes in enriched cortical astrocytes isolated from symptomatic *Mecp2*-deficient mice and WT littermates and across developmental time points. Through the identification of key groups of astrocytic genes, proteins and pathways, we can begin to tease apart the exact mechanisms in which astrocytes contribute to RTT pathogenesis and identify new and much needed therapeutic targets for RTT patients.

INTEGRATED GENOME MAPPING IN NANOCHANNEL ARRAYS AND SEQUENCING FOR BETTER HUMAN GENOME ASSEMBLY AND STRUCTURAL VARIATION DETECTION

Andy Wing Chun Pang, Alex Hastie, Palak Sheth, Thomas Anantharaman, Zeljko Dzakula, Han Cao

BioNano Genomics, Computational Biology, San Diego, CA

De novo genome assemblies using purely short sequence reads are generally fragmented due to the complexity such as repeats and duplications found in most genomes. These characteristics can hinder short-read assemblies and alignments, and that in turn limits our ability to study the genomes.
The BioNano Genomics Irys System linearizes extremely long DNA molecules, thus yielding single-molecules containing long-range information. These hundreds of kilobases molecules can capture distal and structural information that may be missed by other sequencing platforms. The assembled genome maps from these molecules can then scaffold sequencing contigs to validate the accuracy of the sequences, and to anchor the adjacent sequences into the proper order and orientation. The long-range hybrid scaffolds can identify novel chromosomal rearrangements recalcitrant to short-read alignment or reference-guided assembly approaches.
We present a comprehensive analysis of a human genome by combining single molecule genome mapping with one of the most annotated sequence assemblies, the HuRef assembly. Overall, we found that the assemblies of two technologies correspond well, and the resulting hybrid scaffolds are highly contiguous, with a N50 of >35Mb, a value typically unachievable by short-read sequencing. In addition, we compared the structural variation with calls previously detected in the HuRef assembly, and found multiple novel variants spanning over hundreds of kilobases in size. Some of these variants reside in areas where the sequence assembly was poorly covered or was highly fragmented; yet these variants encompass numerous genes, and can be of functional importance. Finally, we identified genome maps that span over the remaining gaps in the reference, as well as maps that resolve and measure long tandem repeats.

THE IDENTIFICATION OF GENETIC MARKERS FOR EXTRATHYROIDAL EXTENSION IN PAPILLARY THYROID CANCER

Ji Yeon Park[1], Jin Wook Yi[2], Chan Hee Park[1], Younggyun Lim[1], Kye Hwa Lee[1], Kyu Eun Lee[2], Ju Han Kim[1]

[1]Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Division of Biomedical Informatics, Seoul, South Korea, [2]Seoul National University Hospital, Department of Surgery, Seoul, South Korea

With advance of diagnosis technology, the early detection of thyroid cancer is greatly increasing, and the most common type is papillary thyroid carcinoma (PTC). PTCs generally show good prognosis, but a few cases have aggressive potential. Extrathyroidal extension (ETE), an indicator for invasiveness of the primary tumor into adjacent tissue, is an important adverse prognostic factor especially for the patients with advanced-stage PTC. The clinical parameter is currently determined by cytological and/or histopathological examination. However, microscopic ETE is not easily detectable before surgery and its impact on clinical outcome remains controversial. Here, to enhance the molecular classification of the microscopic ETE, we aimed to identify genes and pathways involved in ETE. Using both clinical and genomic data of 486 PTC patients in TCGA (The Cancer Genome Atlas), we compared DNA mutation and mRNA expression profiles depending on ETE. We found that BRAF V600E mutation and RAS mutations are positively and negatively associated with ETE, respectively (odds ratio: 2.5 and 0.3, Fisher's exact test $p<0.01$). In GO analysis of differentially expressed genes in ETE, significantly upregulated genes are enriched in "extracellular matrix organization" (Fisher's exact test, 5x10-10), "cell-cell signaling" (4x10-6), "vasculature development" (4x10-6), "cytokine activity" (8x10-6), and downregulated genes are enriched in "ion transport" (2x10-7). Interestingly, the regulation of both WNT signaling and tumor necrosis factor (TNF) signaling is also significant (both, 6x10-4). Notably, WNT4 has reduced expression in ETE while SFRP2 and SFRP4 have enhanced expression, suggesting that the altered antagonistic interaction in WNT signaling is correlated to the invasive phenotype of PTCs. We suggest that the dysregulation of WNT and TNF signaling disrupt cell-cell interaction and increase the motility of PTC cells. However, the gene expression is not well responsive to ETE in the presence of BRAF V600E or RAS mutation, so that our selection of genes may have limitation to represent the invasive property of ETE with genetic defect on BRAF or RAS (58% of investigated patients). Our genome-wide study elucidated genetic components underlying the pathophysiology of ETE. They will be used to evaluate a variety of microscopic ETE by their molecular features, and develop therapeutic targets to inhibit cancer progression.

MULTIPLEX EVALUATION OF PROGRAMMABLE CRISPR/CAS9 TRANSCRIPTION FACTORS USING COMPETITIVE GROWTH ASSAYS IN YEAST.

Justin D Smith[1], Ulrich Schlecht[2], Sundari Suresh[2], Cosimo Jann[3], Hsueh-Lui Ho[4], Ken Haynes[4], Lars M Steinmetz[1,2,3], Ronald W Davis[1,2], Leopold Parts[1,3], Robert P St.Onge[2]

[1]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [2]Stanford University, Stanford Genome Technology Center, Stanford, CA, [3]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany, [4]University of Exeter, Biosciences, Exeter, United Kingdom

Bacterial type II CRISPR-Cas9 systems have been widely adapted for RNA-guided genome editing and transcription regulation in eukaryotic cells. Here, we apply these methods to modulate yeast gene expression levels. As a proof of concept, we activated and repressed several known drug targets in *Saccharomyces cerevisiae*, leading to increased and decreased growth, respectively, in the restrictive condition. We then designed guide RNAs to tile upstream and downstream the transcription start site of different gene sets, array-synthesized oligos to clone complex guide RNA libraries, and used competitive growth assays with sequencing readout to quantify the relative fitness effects of individual guides. First, we tested whether guide positioning had an effect on activity, and observed optimal performance in a 200bp window around the transcription start site. Next, we compared the dCas9 activity using 18bp and 20bp specificity sequence guides, and assessed their relative ability to distinguish between perfect and imperfect target sequences. We confirmed that mutations near the PAM sequence have a large effect compared to distal ones, and found surprisingly good concordance of effects for the 18bp and 20bp versions. We then screened large guide RNA libraries in multiple drug and stress conditions to test whether previously identified QTLs act via changes in gene dose, and to find new modulators of small molecule effects. Using this approach, we identified a novel resistance mechanism to the antifungal drug Fluconazole. Finally, we are extending these methods to *Candida glabrata*, a pathogenic yeast responsible for ~26% of human candidemias and candidiasis, to further characterize the Fluconazole resistance. The high-throughput phenotype-by-sequencing methods we are developing can be readily applied to answer a range of questions on gene function.

Y10K - A POWERFUL YEAST MAPPING POPULATION OF 10,000 FULL GENOME SEQUENCED AND DENSELY PHENOTYPED DIPLOID INDIVIDUALS.

Johan Hallin[1], Kaspar Martens[2], Martin Zackrisson[3], Francisco Salinas[1], Anders Bergstrom[1], Jonas Warringer[3], Leopold Parts[4,5], Gianni Liti[1]

[1]University of Nice, IRCAN, Nice, France, [2]University of Tartu, Department of Computer Science, Tartu, Estonia, [3]University of Gothenburg, Department of Chemistry and Molecular Biology, Gothenburg, Sweden, [4]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany, [5]Stanford University School of Medicine, Department of Genetics, Stanford, CA

While additive genetic effects have been successfully mapped in cohorts of ever-increasing sizes, identification of non-additive signals due to genetic interactions has proven difficult. Previously, we studied the cross between West African (WA) and North American (NA) yeast strains to map genetic effects in haploid strains using novel and traditional linkage methods. Here, we constructed an all-against-all cross of 96 sequenced F12 WAxNA progeny and four additional strains of both mating types, yielding the "Y10k" mapping population of 10,000 strains with complete genome information.

The Y10k population has several features that make it well-suited for analyses of genetic interaction effects. Its size of 10,000 individuals is larger than for any classical linkage population, and is comparable to the contemporary GWAS studies that only have genotype information at marker loci. The 200 parents originate from a two-founder cross, thus all alleles are at high frequency, and due to the all-against-all structure, they are also in perfect Hardy-Weinberg equilibrium. This enables better powered scans of dominance and epistasis due to adherence to the linear model assumptions, and large number of phenotype observations from each of the possible genotype combinations. Further, the diploid yeast population is a natural representation of the reproductive state, avoiding confounding from ploidy-trait interactions observed previously for growth phenotypes.

As a proof of principle, we analysed growth traits for 1152 individuals measured under three stress conditions in quadruplicate. We estimated the phenotype narrow sense heritabilities to range from 64 to 92% using linear mixed models, and mapped QTLs. Using genetic prediction models selected by 5-fold cross-validation, we could achieve prediction errors near the theoretical limit imposed by heritability. We identified significant non-additive effects influencing all the strongest QTLs, observing both dominance of a beneficial allele, and modulators on other chromosomes. We are in the process of expanding these experiments to the full set of sequenced individuals for highly powered genetic mapping, analysis of interactions and heterosis, and genomic prediction.

THE EUROPEAN GENOME-PHENOME ARCHIVE: A MULTI-SITE DATABASE SERVICE FOR CONTROLLED-ACCESS DATA ARCHIVING OF INDIVIDUAL LEVEL –OMICS DATA

Justin E Paschall[1], Jordi Rambla[2], Oscar Martinez-Llobet[2], Marc Sitges-Puy[2], Mario Alberich[2], Sabela Torre[2], Lappalainen Ilkka[1], Jeff Almeida-King[1], Alexander Senf[1], John D Spalding[1], Saif Ur Rehman[1], Paul Flicek[1], Arcadi Navarro[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom, [2]Centre for Genomic Regulation, Barcelona, Spain

The European Genome-phenome Archive (EGA), a joint-service managed by the Center for Genomic regulation (CRG) and the EMBL European Bioinformatics Institute (EMBL-EBI), is a long term archive for potentially identifiable genetic and phenotypic data. The EGA provides world-wide secure data access to bona fide researchers who meet usage criteria consistent with patient consent in a study.

Our collections include, but are not limited to, major reference data for rare and common diseases, including data derived from the UK10K project, RD-Connect IRDirC projects, Wellcome Trust Case Control Consortium (WTCCC), International Cancer Genome Consortium (ICGC) as well as additional datasets useful as controls.

As of February 2015, the EGA securely stores ~1.7 petabytes of data derived from over 600K unique samples and spread across ~1400 distributable datasets.

To accommodate increasing rates of submission, now at more than 100 new studies totaling more than 1 PB per year, the EGA have implemented changes to facilitate programmatic integration with sequencing centers and easier submission processes for researchers, clinicians and consortia, through a new Submitter Portal.

Through a new download matrix 'Data Mart' webpage, users are able to sort, filter and select files based on 'Study', 'Dataset' or 'Sample' centric criteria. This interactive filtering step enables grouping and drilling down to the files of interest, before initiating a download. The new portal enables the user to query all EGA public metadata through a REST API, as well as downloading metadata directly through the EGA website. Similarly, user-experience driven changes to the submission process will allow earlier QC and feedback as a pre-submission step. The collaboration between the EMBL-EBI and the CRG has motivated changes to both submission and data distribution to make these systems fully distributed and federated, and provided a test-bed and additional resources to explore alternative interfaces to the jointly held EGA dataset.

Future developmental plans include further integration with ELIXIR infrastructures and services, and integration with the Global Alliance for Genomics and Health recommendations, APIs and suggested standards.

The EGA is currently available at ega.crg.eu and www.ebi.ac.uk/ega.

IDENTIFICATION OF PATHOGEN-SPECIFIC RESPONSE PATHWAYS IN ACTIVATED IMMUNE CELLS USING A SYSTEMS BIOLOGY APPROACH

Ashwini Patil, Kenta Nakai

University of Tokyo, Institute of Medical Science, Tokyo, Japan

The innate immune response is the first level of protection in organisms against invading pathogens. It is primarily mediated by the Toll-like receptors functioning through the Myd88-dependent and TRIF-dependent pathways. Despite being widely studied, it is not yet completely understood and systems-level analyses have been lacking.
In this study, we identified high-probability networks of genes activated during the innate immune response on exposure to various pathogenic components. We used a network flow optimization approach[1] to analyze time course gene expression profiles of activated immune cells in the context of a large gene regulatory and protein-protein interaction network. We compared the regulatory networks responsible for the distinct immune outcomes produced from different pathogens to identify unique regulatory genes associated with each pathogenic component.

References

1. Patil, A., et al. (2013), PLoS Computational Biology, 9, e1003323.

eMERGE PHENOME-WIDE ASSOCIATION STUDY (PheWAS) IDENTIFIES CLINICAL ASSOCIATIONS AND PLEIOTROPY FOR FUNCTIONAL VARIANTS

A. Verma[1], S.S. Verma[1], S.A. Pendergrass[1], D.C. Crawford[2], D.R. Crosslin[3], H.K. Kuivaniemi[4], W.S. Bush[2], Y Bradford[1], I Kullo[5], S.J. Bielinski[5], R Li[6], J.C. Denny[7], P Peissig[8], S Hebbring[8], E Pugh[9], M De Andrade[5], M.D. Ritchie[1,4], G Tromp[4]

[1]The Pennsylvania State University, Center for Systems Genomics, University Park, PA, [2]Case Western Reserve University, Institute for Computational Biology, Cleveland, OH, [3]University of Washington, Department of Medicine, Seattle, WA, [4]Gesinger Health System, Danville, PA, [5]Mayo Clinic, Rochester, MN, [6]National Human Genome Research Institute, Bethesda, MD, [7]Vanderbilt University, Nashville, TN, [8]Marshfield Clinic, Marshfield, WI, [9]John Hopkins University, Baltimore, MD

We performed a phenome-wide association study (PheWAS) exploring the association between functional stop-gain genetic variants, variants resulting in a truncated protein or lack of transcription, and a comprehensive group of phenotypes, to identify novel associations and uncover potential pleiotropy. We selected 25 stop-gain variants: 20 functionally predicted variants using multiple SNP annotation tools, and 5 known disease associated null variants. We performed analyses using data from the Electronic Medical Records and Genomics (eMERGE) Network across 7 sites with a total of 41,057 unrelated patients above 19 years of age. To create discovery and replication datasets, we divided samples into two datasets by equal proportion by eMERGE site, sex, race, and genotyping platform. We calculated comprehensive associations between these variants and case-control status for 3,835 ICD-9 diagnosis codes (requiring ≥ 3 visits per individual to identify case status, ≥10 case subjects per ICD-9 code). Associations were adjusted for sex, site, genotyping platform and the first 3 principal components. We found 9 replicating associations at p-value<0.01. The top result was between SNP rs328 mapped to *LPL* and "pure hyperglyceridemia" ($p_{discovery}$ =6.50x10-06, $p_{replication}$ =3.21x10-4). We also performed an ancestry-stratified analysis. We identified previously known associations, such as variants in *LPL* associated with hyperglyceridemia and variant in *DARC* associated with neutropenia and leukopenia in the African American set. In addition, we also identified new associations such as *MADD* variants with hypertension, hyperlipidemia, and heart failure. In conclusion, our PheWAS shows stop-gained variants may have important pleiotropic effects, and that PheWAS are a powerful strategy to mine the full potential of the electronic heatlh data for genome-phenome associations. In future, we will be conducting this analysis on ~20,000 pediatric and infant samples in eMERGE network to identify associations with diseases in early ages of development.

LOW FREQUENCY VARIANT PheWAS ANALYSIS FOR LIPID GENES

S Pendergrass[1], S Verma[2], A Verma[2], J Wallace[2], A Okula[2], S Mukherjee[3], J Overton[3], J Reid[3], A Baras[3], F Dewey[3], D Carey[1], D Ledbetter[1], M Ritchie[1]

[1]Geisinger Health System, Sigfriend and Janet Weis Center for Research, Danville, PA, [2]The Pennsylvania State University, Center for Systems Genomics, University Park, PA, [3]Regeneron Genetics Center, Tarrytown, NY

Low frequency single nucleotide variants (SNVs) may prove important for new discovery elucidating the architecture of complex traits and for development of new drug targets. Phenome-Wide Association Studies (PheWAS) provide a way to explore associations between genetic variation and a wide range of phenotypic measures, providing insight into dynamic networks that exist between the genome, gene products, signaling pathways, intermediate phenotypes, and outcome traits, as well as novel associations for further research and drug discovery. As a proof of principle, using whole exome sequencing data for 7403 subjects from the Geisinger MyCode™ biorepository we have performed the first PheWAS focused on low frequency variants within 6 lipid-associated genes (*PCSK9*, *APOB*, *LPL*, *LDLR*, *MIR6886*, *APOE*). We developed a high-throughput analysis pipeline for rare-variant binning and subsequent statistical analyses of these binned variants for association with 474 3-digit and 1,166 5-digit international-classification of disease version 9 (ICD-9) diagnoses. We required ≥ 3 visits per individual with the same ICD-9 diagnosis to identify case status. For SNVs ($0.0015 < MAF < 0.05$), we used the automated bioinformatics tool BioBin to collapse SNVs based on gene boundaries, and used a Madsen and Browning-style variant weighting scheme to statistically evaluate binned variants for association with outcomes, with covariates for sex, year of birth, and BMI. For 3-digit ICD-9 codes we found 25 results with $p < 0.01$, and for 5-digit codes we found 35 results with $p < 0.01$. Results include expected associations between *APOB* and *LDL* associated with ICD-9 272 “disorders of lipoid metabolism”. These analyses illustrate feasibility and validate an approach for unmasking known and novel phenotype associations for low-frequency variants. Our future directions include performing the first comprehensive low-frequency variant PheWAS across all genes using additional samples. We will also bin variants by other biological features such as pathways, and perform a PheWAS identifying pathways and other biological features enriched for SNVs associated with phenotypes. We will also complete a PheWAS analysis of the common variants sequenced within this study, for further discovery. The unprecedented comprehensive phenotypic and genotypic data of these studies will allow us to both test our novel rare-variant PheWAS data analysis pipeline and embark on new PheWAS-based discovery.

SINGLE CELL GENE EXPRESSION RESPONSE TO GLUCOCORTICOIDS

Roger Pique-Regi, Adnan Alazizi, Cynthia Kalita, Gregory Moyerbrailean, Francesca Luca

Wayne State University, Genetics, Detroit, MI

Individual cellular response to environmental changes remains largely unexplored; yet recent results suggest high degree of heterogeneity in baseline gene expression and also in transcriptional response of individual cells to bacterial infection. Glucocorticoids (GC) are steroid hormones that bind the GC receptor (GR) and are a central part of a feedback mechanism in reducing inflammation, a key process in immune response and cellular damage. Previous studies have shown that treatment with Dexamethasone (a synthetic glucocorticoid) induces transcriptional changes in thousands of genes in EBV transformed B-cells (lymphoblastoid cell lines, LCLs). Here, 96 single cells were sorted from a population of LCLs derived from a single individual (GM18507). 50% of the cells were treated with Dexamethasone for 6 hours, and the other 50% were treated with vehicle control. Single cell isolation and RNA-seq libraries preparation were performed on the Fluidigm C1 system following the manufacturer's recommendations. After quantifying and normalizing gene expression levels for each individual cell, a principal component analysis (PCA) reveals two distinct cell populations separated on the first PC which is not informative of treatment. The second PC captures the differences in gene expression induced by the treatment. Together, the two PCs suggest that there are two major cell subpopulations with distinct gene expression profiles following independent trajectories upon treatment. Our results confirm that the average response measured in a population of cells is generated by more complex single cell behaviours. This complexity seems to be influenced by perturbations to the cellular environment and the underlying heterogeneity of the cells. Further investigations of single-cell transcriptional response will be able to shed light on the underlying mechanisms and relevance for organismal phenotypes.

INVESTIGATING THE MOLECULAR UNDERPINNINGS OF HUMAN HIPPOCAMPAL NEUROGENESIS AND THE EFFECTS OF ANTIDEPRESSANTS

Timothy R Powell[1], Tytus Murphy[2], Simone de Jong[1], Jack Price[2], Sandrine Thuret[2], Gerome Breen[1]

[1]King's College London, SGDP Centre, London, United Kingdom, [2]King's College London, Department of Basic and Clinical Neuroscience, London, United Kingdom

Background
Hippocampal neurogenesis (HN) is the process by which new neurons are formed from differentiating neural progenitor cells in the hippocampus. HN is hypothesized to be important in cognition and mood and is affected by diet, inflammation, age and stress. Lower rates of HN have been reported in animal models of depression and antidepressant therapy increases HN, and may represent the cellular mechanism of antidepressant action. The molecular underpinnings of HN and antidepressant-induced HN in human cells remains poorly understood.
Methods
Here we use a human hippocampal progenitor cell line which has been genetically engineered to proliferate in co-culture with tamoxifen and growth factors, and differentiate in their absence. We investigate the genome-wide expression (Illumina HT12.v4) changes associated after 7 days of differentiation, in which hippocampal progenitor cells become neuroblasts. We additionally use the in vitro model to assess the effects of two antidepressants, escitalopram and nortriptyline, whilst cells are proliferating as part of a 48-hour treatment protocol, and whilst cells are differentiating as part of a 9-day treatment protocol. We assess changes to cells using immunohistochemistry; assessing changes to proliferation (BrdU, Ki67), differentiation (Dcx, Map2), gliogenesis (S100B), and cell death (CC3). We additionally characterize genome-wide expression changes after treatment and identify pathways activated by each antidepressant using WEB-based Gene SeT AnaLysis Toolkit.
Results
We found 6713 Bonferonni significant expression changes associated with a 7-day hippocampal progenitor differentiation protocol. Expression changes were enriched for cell cycle pathways. With antidepressant treatment, we found that only when differentiating cells were treated with escitalopram did we find increased neurogenesis; i.e. increases in BrdU, S100B and Dcx staining. This was paired with changes to gene expression profiles enriched for axon formation and ribosomal biogenesis.
Discussion
Results will help in our understanding of the regulation of hippocampal neurogenesis and may aid in the discovery of novel antidepressants.

FUNCTIONAL IMPACT AND EVOLUTION OF A NOVEL HUMAN POLYMORPHIC INVERSION THAT DISRUPTS A GENE AND CREATES A FUSION TRANSCRIPT

Marta Puig[1], David Castellano[1], Lorena Pantano[1], Carla Giner-Delgado[1,2], David Izquierdo[1], Magdalena Gayà-Vidal[1], José Ignacio Lucas-Lledó[1], Tõnu Esko[3,4], Chikashi Terao[5], Fumihiko Matsuda[5], Mario Cáceres[1,6]

[1]Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain, [2]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain, [3]Estonian Biobank, Estonian Genome Center, University of Tartu, Tartu, Estonia, [4]Boston Children's Hospital, Harvard Medical School, Broad Institute of MIT and Harvard University, Boston, MA, [5]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan, [6]Institució Catalana de Recerca i Estudis Avançats, (ICREA), Barcelona, Spain

Despite many years of study of inversions, very little is known about their functional consequences, especially in humans. A common hypothesis is that the selective value of inversions stems in part from their effects on nearby genes, although evidences of this in natural populations are almost nonexistent. Here we present a global analysis of a new 415-kb polymorphic inversion that is among the longest ones found in humans and is the first with clear position effects. This inversion is located in chromosome 19 and has been generated by non-homologous end joining between blocks of transposable elements with low identity. PCR genotyping in 534 individuals from seven different human populations allowed the detection of tag SNPs and inversion genotyping in multiple worldwide populations, showing that the inverted allele is mainly found in East-Asia with an average frequency of 4.7%. Interestingly, one of the breakpoints disrupts a gene coding a zinc-finger transcription factor, causing a significant reduction in the total expression level of this gene in lymphoblastoid cell lines. RNA-Seq analysis of the effects of this expression change in standard homozygotes and inversion carrier individuals revealed distinct expression patterns that were validated by quantitative RT-PCR. Moreover, we have found a new fusion transcript that is generated exclusively from inverted chromosomes around one of the breakpoints. Finally, by the analysis of the associated nucleotide variation, we have estimated that the inversion was generated approximately 43,450 years ago and, while a neutral evolution cannot be ruled out, its current frequencies are more consistent with those expected for a deleterious variant, although no significant association with phenotypic traits has been found so far.

REAL-TIME MONITORING OF DISEASE PROGRESSION BY LONGITUDINAL ANALYSIS OF TUMOR SUBCLONE STRUCTURE IN REFRACTORY BREAST CANCER PATIENTS

Yi Qiao[1], Sam W Brady[2], Andrea Bild[2], Gabor T Marth[1]

[1]University of Utah School of Medicine, Human Genetics, Salt Lake City, UT, [2]University of Utah School of Medicine, College of Pharmacy, Salt Lake City, UT

Metastatic breast cancer is heterogeneous, as a typical tumor contains multiple cell subpopulations, or subclones, with distinct sets of somatic mutations. When chemotherapeutic agents are administered, some of these subclones may gain a selective advantage and develop drug-resistance, resulting in refractory disease. Thus it is imperative to identify resistant subclones and their evolution across treatment. To study these questions, we have collected a highly unique set of metastatic tumor samples from women before, during, and after treatments, often across multiple courses of chemotherapy.

We utilized deep DNA sequencing to find genomic aberrations at each time point, and applied computational methods both to identify the subclones, and to follow their evolution in response to chemotherapy. We substantially extended SubcloneSeeker, our tumor subclone analysis method, for the analysis of an arbitrarily large number of longitudinal tumor biopsies collected from a cancer patient. Our methods include (1) joint clustering of somatic mutations in a multi-dimensional cellular prevalence (CP) space, where CP is derived from somatic allele frequencies in all samples; (2) enumeration of all subclone evolution trajectories across all time points consistent with the CP observations; and (3) the construction of a “consensus subclone structure” i.e. the the evolutionary trajectory shared across all or a majority of computationally derived alternative subclone structures.

We applied our method to a breast cancer patient who had received three regimens of thermotherapy, and had tumor biopsies taken at 5 consecutive time points. Our analysis revealed that the tumor reacted to treatment in two distinct patterns. In the first pattern (observed in Taxol+Herceptin treatment), although a significant reduction in tumor burden was achieved, the major pre-treatment subclones survived, and subsequently expanded post-treatment, indicates that the treatment was effective, but terminated prematurely. In the second pattern (observed in Lipo-Doxorubicin treatment), a single pre-treatment subclone with extra mutations (in our patient in a tumor suppressor gene related to cell differentiation/growth) survived the treatment, and became the founding clone for the tumor population post-treatment, indicating that the extra mutation(s) are most likely to be causative for chemo-resistance. These algorithms make real-time monitoring of patients through chemotherapy possible, and offer a rational basis for personalized therapeutic choices for oncologists.

DENISOVAN ANCESTRY IN EAST EURASIAN AND NATIVE AMERICAN POPULATIONS.

Pengfei Qin[1], David Reich[2], Mark Stoneking[1]

[1]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany, [2]Harvard Medical School, Genetics, Boston, MA, [3]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany

Although initial studies suggested that Denisovan ancestry was found only in modern human populations from island Southeast Asia and Oceania, more recent studies have suggested that Denisovan ancestry may be more widespread. However, the geographic extent of Denisovan ancestry has not been determined, and moreover the relationship between Denisovan ancestry in Oceania and that elsewhere has not been studied. Here we analyze genome-wide SNP data from 2493 individuals from 221 worldwide populations, and show that there is a widespread signal of a very low level of Denisovan ancestry across Eastern Eurasian and Native American (EE/NA) populations. We also verify a higher level of Denisovan ancestry in Oceania; the Denisovan ancestry in Oceania is correlated with the amount of New Guinea ancestry, but not the amount of Australian ancestry, indicating that recent gene flow from New Guinea likely accounts for signals of Denisovan ancestry across Oceania. However, Denisovan ancestry in EE/NA populations is equally correlated with their New Guinea or their Australian ancestry, suggesting a common source for the Denisovan ancestry in EE/NA and Oceanian populations. There are various scenarios that could account for this shared Denisovan ancestry; overall, our results indicate a more complex history involving East Eurasians and Oceanians than previously suspected.

ASSESSING CELL-TO-CELL DNA METHYLATION VARIABILITY ON INDIVIDUAL LONG READS

Wei Qu[1], Hideaki Yurino[2], Shin-ichi Hashimoto[2], Tatsuya Tsukahara[3], Hiroyuki Takeda[3], Shinichi Morishita[1]

[1]the University of Tokyo, Department of Computational Biology, Tokyo, Japan, [2]Kanazawa University, Graduate School of Medical Sciences, Kanazawa, Japan, [3]the University of Tokyo, Department of Biological Sciences, Tokyo, Japan

Cell-to-cell variability of cytosine methylation is essential for deep understanding inherent cellular perturbation and its molecular machinery. However, traditional methylation studies has overlooked it but focused on common mechanisms at cell population level; little has been uncovered by recent single-cell approach either because of technical limitations. Here we report a genome-wide detection of cell-to-cell DNA methylation variability by comparing methylation status of coexisting CpGs on long sequencing reads. We observed much lower methylation variability in hypomethylated regions across the entire genome, and a dynamical gradational change of methylation status on the boundaries of hypomethylated regions. This method allows a concise and comprehensive assessment of cell-to-cell DNA methylation variability.

IDENTIFICATION OF GENES INVOLVED IN FUNCTIONAL RECOVERY AFTER STROKE THROUGH EXOME SEQUENCING OF EXTREME PHENOTYPES

Raquel Rabionet[1], Marina Mola[2], Carolina Soriano[2], Caty Carrera[3], Georgia Escaramís[1], Stephan Ossowski[4], Israel Fernandez-Cadenas[3], Jordi Jimenez-Conde[2], Xavier Estivill[1]

[1]Center for Genomic Regulation (CRG), UPF and CIBERESP, Bioinformatics and Genomics, Barcelona, Spain, [2]Institut Hospital del Mar d'Investigacions Mediques (IMIM), Neurovascular Research Group (NEUVAS), Barcelona, Spain, [3]Fundació Docència i Recerca Mutuaterrassa, Terrassa, Spain, [4]Center for Genomic Regulation (CRG) and Univestitat Pompeu Fabra (UPF), Bioinformatics and Genomics, Barcelona, Spain

Cerebrovascular diseases are the second most important cause of death in Spain, and together with other neurodegenerative diseases, they are the leading cause of disability in adults. Variability in functional outcome after a stroke can be influenced by many factors. Irrespective of clinical factors such as age, stroke etiological subtype, vascular stenoses, location of the injury and the size of the affected area, inter-individual variation in capacity of neuronal recovery is considerable. A number of systems and metabolic pathways are important for response to cerebral ischemic damage, and their activity may be modulated by variation in the genes that encode their various components. We aim to identify genetic variants, genes and pathways influencing the functional recovery process.
With this aim, we have selected 81 patients with extreme phenotypes (36 bad vs 45 good outcome), suffering an anterior territorial ischemic stroke, with similar stroke severity, and matched for basal functional level, age and gender, from a cohort of over 4000 stroke cases. These patients underwent exome sequencing (Nimblegen v3 and Illumina sequencing).Downstream analysis was performed using a generalized mixed effect model association test, with variants collapsed by gene and weighted by their frequency and functional (condel) scores, with the aim of identifying genes with an accumulation of variants in either set of samples. This provided a set of 300 nominally significant genes (none passed multiple testing correction), including some relevant genes for stroke risk and stroke outcome. CNV analysis and integration with GWAS results from this and other cohorts of stroke cases is underway.

CHARACTERIZING SUBCLONAL EVOLUTION IN LYMPHOMA

Deepthi Rajagopalan, Jenny Zhang, Andrea Moffitt, Anupama Reddy, Casandra Love, Tiffany Tzeng, Sandeep Dave

Duke University, Duke Center for Genomic and Computational Biology, Durham, NC

Introduction
Lymphomas represent the fourth most common type of cancer affecting over 300,000 individuals world-wide every year. Previous studies (including ours) have identified numerous cancer genes and a striking genetic heterogeneity among lymphomas. However, the role of genetically diverse subclones driving tumor progression remains to be defined. In this study, we sought to define the role of subclonal evolution as a driver of intratumor heterogeneity in lymphomas.

Methods
We performed 100x depth exome sequencing on 250 aggressive lymphomas (and paired normal tissue) to define somatic genetic mutations that occur in these tumors. We compared sequencing read depth from tumor and normal samples to identify somatic copy number alterations. We then inferred copy number alterations in conjunction with somatic mutation allelic fractions to infer tumor purity and clonality. We further integrated the allelic fraction and copy number alterations to identify distinct tumor subpopulations.

Results
We identified significant subclonal heterogeneity involving many recurrently mutated genes. The range of measured copy number for frequently mutated genes varied up to 8-fold indicating a significant genetic instability in a number of these tumors. The 250 cases were widely variable in subpopulation structure, with some cases characterized by a dominant clone and others consisting of many smaller subclones. When these data were combined with extensive clinical data on these cases, we observed that the subclonal evolution of mutations plays an important role in the clinical outcome and treatment response. This analysis reveals the genetic evolution of the tumors and provides predictive power on their clinical outcome.

Our data indicate the importance of not only identifying the numerous genes that undergo alterations in the tumor but also the degree of subclonal evolution in these tumors. Our work underscores the need to analyze the tumor genome, not as a single entity, but rather as an ecosystem of subclones that play a role in the disease outcome.

NOVEL PROBABILISTICALLY INTERPRETABLE METHODS FOR IDENTIFYING AND LOCALIZING TARGETS OF SELECTIVE SWEEPS

Lauren A Sugden[1], Brenna M Henn[2], Sohini Ramachandran[1]

[1]Brown University, Ecology and Evolutionary Biology and Center for Computational Molecular Biology, Providence, RI, [2]Stony Brook University, Department of Ecology and Evolution, Stony Brook, NY

Many methods have been introduced to locate genomic targets of hard selective sweeps. These rely on single statistics that identify a major signature of a sweep: long-range haplotype blocks, enrichment for high- and low-frequency variants in the site frequency spectrum, or locus-specific population differentiation. Recent methods that combine multiple statistics into a composite score have shown increased power to detect sweep sites, but these methods must artificially compensate when a subset of their component statistics is undefined, as often happens with long-range haplotype statistics. These methods also yield scores that lack a clear interpretation, and their classification schemes rely on arbitrary thresholds.

We developed novel methods for identifying targets of hard selective sweeps that report the probability that a site in question is the site of an adaptive mutation and inherently account for undefined statistics. One method is a Naive Bayes classifier that learns the distributions of multiple component statistics through simulations of neutral loci and sweep scenarios, given a demographic model for the populations of interest. Our method infers the posterior probability that a variant has undergone a sweep, given the set of defined component statistics at that site. Based on comparisons using simulated data, our method outperforms the Composite of Multiple Signals (Grossman et al. 2010), another composite method, particularly in two biologically interesting scenarios: when identifying completed sweeps (where the beneficial allele fixes) and when identifying recent fast sweeps. Our method recovers known sweep targets in 1000 Genomes data and identifies a potential novel adaptive variant in the ADH gene family. We extend our classification framework and further increase power by considering dependencies between component statistics. We find that controlling for observed values of FST leads to an excellent classifier, due to the correlation of FST with numerous other statistics used to identify selective sweeps from genomic data.

We apply our classifiers to exome and SNP array data from the San population in southern Africa. In addition to identifying hard sweep sites, we compare two demographic models (Gronau et al. 2011, Schaffner et al. 2005) that differ dramatically in parameters like divergence times and effective population sizes. This application of our methods allows us to examine the effect of demographic parameter estimates on the performance of our classifiers.

SPATIAL RESOLUTION OF RNA STRUCTURES BY PROXIMITY LIGATION.

Vijay Ramani, Ruolan Qiu, Jay Shendure

Department of Genome Sciences, University of Washington, Seattle, WA

The folding of ribonucleic acid (RNA) species into complex secondary and tertiary structures is central to RNA's catalytic, regulatory, and information-carrying roles within every cell of every living organism. Traditional approaches for resolving essential RNA and protein-RNA structures (e.g. ribosomes, spliceosomes, snRNPs) employ high-resolution structural techniques (e.g. crystallography, electron microscopy). However, these are limited in throughput, and provide only a snapshot of dynamic molecules. More recently developed methods that combine chemical probing with high-throughput sequencing (e.g DMS-seq, SHAPE-seq) powerfully probe the *extent* of participation of individual bases in secondary structures transcriptome-wide, but do not directly query *which specific pairs* of bases or regions interact to form these structures. We speculated that *in situ* proximity ligation of native RNA followed by deep sequencing would yield transcriptome-wide maps of spatial proximity. Specifically, proximity information would be captured by chimeric reads with ligation junctions in the immediate vicinity of structured RNA regions, analogous to how 3C techniques capture genome conformation. Here, we present proof-of-concept of this method, termed RNA Proximity Ligation (RPL), in both yeast and human cells. We apply RPL to generate contact probability maps for ribosomal and other abundant non-coding RNAs, including the yeast U2 spliceosomal RNA, several snoRNAs, and the human and yeast RNA subunits of the signal recognition particle. RPL measurements, which are informative of pairwise interactions between bases or regions within a given RNA, broadly agree with accepted secondary and tertiary structures where they are available, recapitulating physical proximity implied by both base-pairing relationships and available crystallographic data. Our results suggest that RPL may eventually bridge the gap between high-throughput assays and classical structural biological techniques. In ongoing work, we are further developing RPL in the following ways: 1) Examining the regulatory consequences of recurrent structural motifs in transcripts; 2) Defining secondary and tertiary structures for native long noncoding RNAs (lncRNAs) with well-defined functions, including Xist and HOTAIR; 3) Integrating RPL measurements with high-throughput methods for in vivo chemical probing (e.g. SHAPE-seq) and computational structural prediction (e.g. FARFAR) to generate empirically-rooted three-dimensional models for any RNA species of interest.

GENETIC MAPPING UNCOVERS *CIS*-REGULATORY LANDSCAPE OF RNA EDITING

Gokul Ramaswami, Jin Billy Li

Stanford University, Genetics, Stanford, CA

A-to-I RNA editing plays an essential role in normal functioning of the nervous system and is catalyzed by the ADAR family of enzymes that act on dsRNA substrates. To gain insights into the underlying *cis*-regulatory architecture of A-to-I editing, we looked at how natural genetic variation affects RNA editing levels in *Drosophila melanogaster*.

Using a microfluidics-based multiplex PCR approach developed in our lab[1], we accurately quantified the editing levels at over 800 editing sites in 131 strains of *D. mel* from the Drosophila Genetic Reference Panel (DGRP)[2]. We identified 486 cis editing-QTLs (ed-QTLs) for 358 different editing sites. Using computational RNA structure prediction methods, we discovered structural features within the primary editing duplex that affect ADAR binding affinity. Additionally, we also observed an enrichment of ed-QTLs within distant secondary RNA duplexes that have recently been shown to influence editing[3,4]. Currently, we are extending our analysis by performing in vitro editing assays coupled with experimental RNA structure mapping on a library of ADAR substrates containing many different mutations.

In this study, we identified genetic variants in *Drosophila* that influence RNA editing levels through changes in RNA structure. A combination of computational and experimental methods has allowed us to more finely elucidate the relationship between RNA structure and ADAR binding.

1. Zhang et al. *Nat Methods* 2014
2. Mackay et al. *Nature* 2012
3. Rieder et al. *Nat Commun* 2013
4. Daniel et al. *Nucleic Acids Res* 2012

GENETIC LANDSCAPE OF COMMON VARIABLE IMMUNE DEFICIENCY

Anupama Reddy, Manoj Kanagaraj, Andrea Moffitt, Jenny Zhang, Sandeep Dave

Duke University, Center for Genomic and Computational Biology, Durham, NC

Introduction:
Common variable immune deficiency (CVID) is the most common form of immune deficiency. This heterogeneous immune disorder is characterized by a deficiency of antibodies, resulting in increased susceptibility to infections. CVID patients also develop cancer at 10-100 fold higher rates. Although the disease has the hallmarks of a genetic disorder including young age of onset and a tendency to run in families, the genetic basis of CVID remains unknown. In this study, we sought to define the genetic basis of CVID through exome sequencing of several hundred CVID patients and case-parent trios.

Results:
We sequenced the exomes of 272 samples (194 CVIDs and 24 trios) to define the genetic variation that characterizes CVID. We further defined the chromatin states and expression patterns in normal human immune cells. We developed a novel statistical framework to identify genes involved in CVID that included genetic variation rates compared to controls, background mutation rates, and by comparing expression and chromatin patterns in immune cells. This ongoing work has identified ~50 candidate genes including PIK3CD, ANAPC1, EXO1 and ZP3 that are highly enriched in immune regulatory pathways. A number of these gene mutations occur in a mutually exclusive fashion indicating potential functional relationships. Analysis of the case-patient trios identified complex patterns of de novo mutations and compound heterozygous inheritance that contribute to CVID.

Conclusion:
We have characterized for the first time, the genetic landscape of CVID and have identified a number of candidate genes. Our work highlights the genetic basis of CVID and has the potential for developing novel biomarkers and patient stratification strategies in the clinic.

SYSTEMATIC IDENTIFICATION OF METHYLATION QUANTITATIVE TRAIT LOCI ACROSS THE HUMAN LIFECOURSE

Caroline L Relton[1,2], Tom R Gaunt[1], Hashem A Shihab[1], Gibran Hemani[1], Josine Min[1], Paolo Casale[3], Geoff Woodward[1], Oliver Lyttleton[4], Chris Zheng[1], Wendy L McArdle[4], Karen Ho[4], Oliver Stegle[3], Sue M Ring[1,4], David M Evans[1,5], George Davey Smith[1]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom, [2]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom, [3]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, United Kingdom, [4]ALSPAC, University of Bristol, Bristol, United Kingdom, [5]Diamantina Institute, University of Queensland, Brisbane, Australia

Understanding the genetic contribution to variation in DNA methylation will provide insight into the epigenetic mediation of traits and diseases and has important implications for identifying the environmentally responsive component of the methylome. We performed a genome-wide survey of blood methylation quantitative trait loci (mQTL) in a longitudinal analysis of ~800 children (birth, 7 years and 15-17 years old) and their ~800 mothers (pregnancy and 15-17 years later) from the Accessible Resource for Integrated Epigenomic Studies (ARIES). We identified c.1.15 million cis mQTL and c.150,000 trans mQTL at birth, with similar numbers in childhood, adolescence, pregnancy and middle age. Of these c.22,000 cis-mQTL and c.1,700 trans-mQTL were independent. A comparison of mQTL across time demonstrated that the majority (75%) of associations were stable over time replicating at $p<1x10-7$ at all timepoints). However, a subset (<1%) of associations were unique to each timepoint. The proportion of variation in methylation that is due to genetic effects reduced over time, suggesting increased importance of environmental influences through the lifecourse.

HOST CELL FACTOR 1 BINDS TO GENE PROMOTERS IN THE MOUSE LIVER CHROMATIN SHOWING DIVERSE TRANSCRIPTIONAL REGULATIONS

Leonor Rib[1,2], Dominic Villeneuve[1], Viviane Praz[1,2], Olivier Martin[2], Nouria Hernandez[1], Nicolas Guex[2], Winship Herr[1]

[1]University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, Vital-IT, Lausanne, Switzerland

The protein Host Cell Factor 1 (HCF-1) is a conserved metazoan transcription regulator that coordinates the interaction of DNA-binding transcription factors and chromatin-modifying enzymes. In mammalian cells, its activity is key for cell cycle progression and a recent genome-wide binding study of HCF-1 in human HeLa cells showed it is a common component of the majority (over 5,000) of active CpG-island promoters (Michaud et al., Genes Dev 2013). Currently, we are investigating the genome-wide binding sites of HCF-1 and the associated transcriptional profiles in the mouse liver in both quiescent liver and during its regeneration post partial hepatectomy (PH).

The liver performs a variety of important metabolic functions, while being constituted of largely (80%) one differentiated cell type, the hepatocyte. The resection of two thirds of the liver during PH induces a rapid and synchronized entry of hepatocytes into the cell-division cycle until the liver restores its original mass. Being highly synchronous, PH-induced liver regeneration is an attractive system to investigate the genome-wide HCF-1 occupancy during cell proliferation of differentiated cells in a natural stress response.

We show here that, as in HeLa cells, HCF-1 binding sites are largely limited to transcriptional start sites (TSSs) but that in the quiescent liver the number of HCF-1 bound TSSs is considerably more limited, being roughly 200. Post PH, the number of HCF-1 binding sites increases rapidly, such that within one hour over 1,000 TSSs are bound and yet they remain enriched in CpG-island promoters that are co-occupied by the RNA polymerase II. HCF-1 bound TSSs belong to genes involved in the regulation of cell-cycle progression and liver metabolic functions. We thus show the diversity of transcriptional regulation by HCF-1 in both cell-proliferation and differentiated-cell processes.

GENOME ANALYSIS OF A PHYLUM: INITIAL HIGHLIGHTS FROM THE I5K PILOT AT THE BAYLOR COLLEGE OF MEDICINE HUMAN GENOME SEQUENCING CENTER.

Stephen Richards, Daniel Hughes, Shwetha C Murali, Shannon Dugan, Kim C Worley, Richard A Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

The i5K is an initiative to sequence the genomes of 5,000 arthropods of medical, agricultural and scientific importance. As a pilot project to identify potential problems and challenges in species selection, identification and acquisition, DNA isolation, sequencing strategies and assembly, automated and manual annotation, analysis and publication, the BCM-HGSC is sequencing ~30 arthropods, selected by the i5K species selection committee.
We present our progress, working with 30 groups of collaborators on the project, the lessons from our attempts to obtain DNA, sequence and assemble multiple genomes in a factory environment. We also present an automated annotation pipeline for arthropods that enables the automated annotation of high quality gene models for 1,000's of insect genomes. The annotation pipeline is based on the Maker pipeline, but uses extensive metazoan protein alignment evidence, RNAseq evidence from the species being sequenced, as well as automatic training of the ab-initio Augustus and SNAP gene predictors. Approximately 30 evidence tracks are additionally generated for manual annotators using the web-Apollo tool, enabling manual annotators to see all input evidence into a gene model. To date 24 species have been put through the i5K pilot annotation pipeline, and are being hosted by the National Agricultural Library providing access to search, community manual annotation and browser functionality
Finally, we provide initial highlights from the pilot, with results showing the whole genome duplication in the spider lineage from the house spider genome, expanded gene numbers and the connection between insect olfactory and crustacean gustatory as seen in the copepod genome, extensive lateral gene transfer in the bedbug genome, and interactions with the host and pesticide resistance in the sheep blow fly.

THE HISTORY AND WEAPONRY OF AN EXISTENTIAL BATTLE BETWEEN A GALL FORMING PARASITE AND ITS PLANT HOST AS TOLD THROUGH THE GENOME SEQUENCE OF MAYETIOLA DESTRUCTOR.

Stephen Richards[1], Chaoyang Zhao[2], Robert M Waterhouse[3], Ming-Shun Chen[4], Susan J Brown[4], Jeffery J Stuart[2]

[1]Baylor College of Medicine, Human Genome Sequencing Canter, Houston, TX, [2]Purdue University, Department of Entomology, West Lafayette, IN, [3]University of Geneva Medical School, Department of Genetic Medicine and Development,, Geneva, Switzerland, [4]Kansas State University, Division of Biology, Manhattan, KS

Mayetiola destructor or the hessian fly is a obligate dipteran wheat parasite that during its extremely short adult lifespan (~24 hours) lays its eggs on budding wheat tissue. Emerging larvae 'spit' salivary fluid on growing wheat tissue tricking the host wheat plant into feeding and protecting it's larvae with gall tissue at the expense of stunting it's own growth. Failure to form a gall by *M. destructor* is fatal for the individual, which generates intense selection pressures on genes encoding gall formation and maintenance. There is an ongoing battle between wheat breeders identifying resistant wheat lines based on plant recognition of foreign proteins - effector triggered immunity (ETI) - and *M. destructor* overcoming wheat resistance lines within seven years per line.
We applied genome sequencing, proteomic analysis, and genetic mapping of *M. destructor* lines showing virulence in previously resistant plant lines, to identify effector molecules overcoming host resistance by effector triggered immunity. We find over 1,000 putative effector proteins undergoing extreme diversifying selection to avoid detection by host plant ETI. These genes are in large gene families with both regular and inverted conservation patterns underscoring their functional importance for gall formation and survival. Genetic mapping of newly virulent strains identified two members of the largest gene family of effector proteins with 426 members – perhaps the largest single gene family known to date. Members of this gene family are found by proteomic analysis in larval saliva. The virulence mutations in one line knocked out gene expression, and produced a stop codon in the other, enabling evasion of ETI and allowing other members of the family to initiate and maintain gall function. Thus the history of plant gene detection and evasion is shown in these massive gene families. We hypothesize that *M. destructor* populations contain null mutations of all effector genes at low allele frequency, such that a small fraction of the population always avoids resistant gene based effector triggered immunity.
Weaponry: we also investigated the function of effector proteins, and despite difficulties due to extreme diversifying selection, find the genes contain F-box and LRR domains involved in and mimic E3 ubiquitin ligases, to target plant proteins for degradation. Yeast two hybrid analyses confirmed direct interaction with SKP-1 like proteins in these plant ubiquitin ligase complexes and suggests a hijacked plant proteasome directly producing nutritive tissue for the larvae, and additionally defeating plant basal immunity and stunting plant growth.

EPISTATIC GENE-BASED INTERACTION ANALYSES FOR GLAUCOMA IN eMERGE AND NEIGHBOR CONSORTIA

Shefali S Verma[1], Jessica N Cooke Bailey[2], Anastasia Lucas[1], Yuki Bradford[1], Jim Linneman[3], Peggy Peissig[3], Murray Brilliant[3], Catherine A McCarty[4], Tamara R Vrabec[5], Mariza de Andrade[6], Gerard Tromp[5], Janey L Wiggs[7], Jonathan L Haines[2], Marylyn D Ritchie[1,5]

[1]The Pennsylvania State University, University Park, PA, [2]Case Western Reserve University, Cleveland, OH, [3]Marshfield Clinic, Marhfield, WI, [4]Essentia Institute of Rural Health, Duluth, MN, [5]Geisinger Health System, Danville, PA, [6]Mayo Clinic, Rochester, MN, [7]Harvard Medical School, Boston, MA

Primary open angle glaucoma (POAG) is a complex disease and one of the major leading causes of blindness worldwide. Genome-wide association studies (GWAS) have successfully identified various common genetic variants associated with glaucoma. However, most of these variants only explain a small proportion of genetic risk. Since biology is complex, it is unlikely that individual loci account for all genetic risk; therefore, it is believed that gene-gene interactions can explain part of the missing heritability. We first performed a GWAS on glaucoma samples obtained from electronic medical records (EMR) in the eMERGE network, to show the power of EMR data in detection of non-spurious and relevant associations. Our findings from GWAS suggest consistent evidence with most of the previously known associations in POAG. Next, we performed an interaction analysis for variants that are marginally associated with glaucoma (main effect p-value <0.01) and observed interesting findings in the eMERGE data (N=5090). Loci from the top epistatic interactions from eMERGE (Likelihood Ratio Test i.e. LRT p-value <1e-05) were then replicated in NEIGHBOR dataset (N=4422). To address the challenge of heterogeneity, we performed a gene-based SNP-SNP interaction analysis and observed significant interactions among genes found in both eMERGE and NEIGHBOR datasets. This resulted in 13 unique replicating gene-gene models consisting of 25 genes. Out of 25 genes, 21 are expressed in the eye. Among the top associations are the interaction among protein coding genes DNAH11 and ZNF521 (LRT pval 3.18E-06 and 7.63E-05 for eMERGE and NEIGHBOR respectively). Another interesting association discovered from this analysis is an interaction between PTRF and SLC7A1 (LRT pval 5.74E-05 and 4.58E-04 for eMERGE and NEIGHBOR respectively) that may possibly interact in TGFB signaling pathway via other genes. PTRF encodes cavin molecules that are involved in many processes such as signal transduction. This analysis represents a gene based interaction search in glaucoma and can provide major insights in exploring genetic architecture of POAG.

CATALOG OF FUSION GENES EXPRESSED IN THE CANCER CELL LINE ENCYCLOPEDIA

Heather Geiger, Nicolas Robine

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

Fusion genes resulting from genomic rearrangements are important biomarkers and potential drug targets in cancer. In order to help the annotation of novel fusion genes, we established a catalog of fusions from the Cancer Cell Lines Encyclopedia (CCLE).
We used the fusion discovery software FusionCatcher to search for fusion genes in the RNA-Seq libraries sequenced from the 929 cell lines of the CCLE project, assembled by Novartis and the Broad Institute. We identified 14853 unique fusion gene candidates, of which 9899 are assigned a high confidence description and 3928 are predicted to be “in-frame” fusions. 2630 fusion genes (and 558 “in-frame” fusions) were identified in more than one sample. We identified fusion genes private to the tissue of origin of the cell lines (such as BCR-ABL1 in leukemia cell lines), as well as fusions present in cell lines corresponding to multiple cell types.
We compared our results with existing catalogs of fusion genes obtained from the TCGA project and from the Genentech cell line bank. We also ran several fusion gene discovery softwares on a subset of the CCLE dataset and compared the list of candidate fusions.
We believe the catalog of fusion genes in Cancer Cell Line is a valuable resource for the cancer genomics community and will help researchers to annotate novel fusion genes and select appropriate cell lines to further characterize of the biological effect of these fusion genes.

EPIGENOMIC ANNOTATION OF GENETIC VARIANTS USING THE ROADMAP EPIGENOME BROWSER

Nicole B Rockweiler[1], Xin Zhou[1,2], Daofeng Li[1], Bo Zhang[1], Rebecca F Lowdon[1], Renee L Sears[1], Ting Wang[1]

[1]Washington University School of Medicine, Department of Genetics, St. Louis, MO, [2]Washington University in St. Louis, Department of Psychiatry, St. Louis, MO

Advances in next-generation sequencing technology have reshaped the landscape of functional genomic and epigenomic research as well as human genetics studies. Large consortia, such as the Roadmap Epigenomics Consortium and ENCODE, have generated tens of thousands of sequencing-based genome-wide datasets, creating a valuable resource for the scientific community. Here, we present the Roadmap EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/roadmap/) that integrates data from the Roadmap Epigenomics Consortium and ENCODE for visualization and bioinformatics analysis. The user interface and the governing data structure of the Browser are optimized to perform large dataset integration; advanced visualization such as gene set view, genome juxtaposition, and chromatin interaction display; and real-time statistical and clustering analyses. Investigators can also view their own datasets in the Browser to analyze them in the context of the Browser's complete reference epigenomes, or alternatively, clone the Browser on the Amazon Cloud to explore private data alongside the public data. The Browser enables researchers to explore and analyze the cell/tissue-specificity of epigenetic modifications; covariation of epigenomic, transcriptomic, and transcription factor binding profiles across cell/tissue types; and epigenomic annotation of genetic variants. Annotation of noncoding variants in the genome with genomic and epigenomic data using the Browser facilitates the generation of novel, testable hypotheses regarding the functional consequences of genetic variants associated with human complex traits. For example, by clustering the H3K4me1 profile near noncoding SNPs associated with multiple sclerosis and the expression of their closest gene across multiple tissues and cells types, we identified a SNP in an immune cell-specific enhancer that is potentially targeting TCF7, a T cell-specific gene downstream from the SNP. Together with the rapidly growing number of publically available epigenomic datasets, the Roadmap EpiGenome Browser will facilitate the translation of genetic signals into molecular mechanisms, leading to prognostic, diagnostic, and therapeutic advances.

WHOLE GENOME ASSEMBLY OF THE GRAY MOUSE LEMUR (MICROCEBUS MURINUS) GENOME: INTEGRATING DIVERSE PLATFORMS AND DATA TYPES

Jeffrey Rogers[1], Peter A Larsen[2], Muthuswamy Raveendran[1], Yue Liu[1], Adam English[1], Yi Han[1], Vanessa Vee[1], C R Campbell[2], Jennifer Shelton[3], Susan J Brown[3], Donna M Muzny[1], Richard A Gibbs[1], Anne D Yoder[2], Kim C Worley[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Duke Univ., Duke Lemur Center, Durham, NC, [3]Kansas State Univ., Bioinformatics Center, Manhattan, KS

Mouse lemurs (genus *Microcebus*) are a diverse radiation of strepsirrhine primates, endemic to Madagascar, with approximately 21 recognized species. The gray mouse lemur (***M. murinus***) is remarkable for its capacity to manifest both opportunistic and seasonal torpor (a state of reduced metabolic rate). Ongoing studies in captive populations also demonstrate that this species exhibits a number of traits that are significant risk factors for disease in humans, including individual variation in age-related cognitive decline and associated neurodegeneration. To facilitate future research in these and other areas, we developed a significantly improved whole genome assembly for the gray mouse lemur using the HGSC approach. Employing multiple paired-end and mate-pair libraries, we generated deep Illumina coverage and assembled the data with ALLPATHS, Atlas-Link and Atlas-GapFill generating an initial assembly with contig N50 of 56 kb. We next used 9x Pacific Biosciences long read coverage and PBJelly2.0 software to fill 59% of intra-scaffold gaps, improving the assembly to a contig N50 of 183 kb. Optical mapping using the BioNano Irys platform provided orthologonal quality assessment of the assembly and generated super-scaffolds with N50 of 7.2 Mb (longest scaffold 45.6 Mb) --- a 128-fold improvement over the original assembly. This improved assembly reveals a novel inversion on the mouse lemur X-chromosome that affects 500 kb. Whole genome sequencing of additional animals from the Duke Lemur Center provides information about SNP and indel polymorphism in this species. Deep RNA-sequencing through a collaboration with the Nonhuman Primate Reference Transcriptome Resource project has generated extensive tissue-specific transcriptome data for 9 tissues. Collectively, this multi-faceted genomic approach makes *M. murinus* an outstanding resource for studies of comparative genomics, inferences regarding ancestral primate genome content and gene complement, chromosome evolution and specific genetic mechanisms related to risk for several human diseases.

IDENTIFYING PATHOGENIC HUMAN VARIANTS: COMPUTERS VERSUS HUMANIZED YEAST

Song Sun[1,2,3], Fan Yang[1,2], Guihong Tan[1], Michael Costanzo[1], Rose Oughtred[4], Jodi Hirschman[4], Chandra Theesfeld[4], Pritpal Bansal[1,2], Nidhi Sahni[5], Song Yi[5], Analyn Yu[1,2], Tanya Tyagi[1,2], David E Hill[5], Marc Vidal[5,6], Brenda J Andrews[1], Charles Boone[1], Kara Dolinski[4], Frederick P Roth[1,2,5]

[1]University of Toronto, Donnelly Centre, Toronto, Canada, [2]Mt Sinai Hospital, Lunenfeld-Tanenbaum Research Institute, Toronto, Canada, [3]Uppsala University, Medical Biochemistry and Microbiology, Uppsala, Sweden, [4]Princeton University, Lewis-Sigler Institute, Princeton, NJ, [5]Dana-Farber Cancer Institute, Center for Cancer Systems Biology (CCSB), Boston, MA, [6]Harvard Medical School, Department of Genetics, Boston, MA

Each individual human carries 100-400 rare variants, each with a potentially major impact on health and disease. Technology identifying deleterious variants can identify causal genes and variants, estimate disease risk and inform therapy. Diverse computational and experimental methods exist to infer pathogenicity for rare human coding variants. While computational methods have limited predictive power, experimental assessment of variant function by human-cell-based phenotyping is also hampered, for example, by inefficient allele replacement technology and by paralogs with overlapping function. Thus, complementation testing in 'humanized' model organisms is an attractive alternative. However, a systematic comparison of computational and experimental methods has been lacking. Using a reference panel of 179 variants in 22 human disease genes, we find that functional complementation testing in *S. cerevisiae* outperforms current computational methods in predicting pathogenic variants. Success in this one billion-year-diverged model organism argues more generally for human variant functional assessment within model organisms.

EVOLUTIONARY ANALYSIS OF ENDOGENOUS RETROVIRUSES IN PRIMATES

Andrei Rozanski*, Fabio P Navarro*, Ana Paula S Urlass, Paola A Carpinetti, Anamaria A Camargo, Pedro A Galante

Hospital Sirio-Libanes, Molecular Oncology Center - IEP, Sao Paulo, Brazil

Endogenous retroviruses (ERVs) sequences account for ~8% of the human genome. These transposable elements are remnants of ancient viral infections and are spread throughout our genome and other primates. One of the main forms of ERV are the Solo-LTRs, a product of homologous recombination of complete ERV. In one hand, ERVs have been related to the gain of new physiologic capabilities, on the other, studies associating ERVs to diseases like cancer, multiple sclerosis among others, have becoming more frequent. Here, we conducted a genome-wide analysis of ERVs in primates genomes, focusing the evolution of Solo-LTRs and the identification of human-specific events. Our approach included human and other five primate genomes (marmoset, rhesus, orangutan, gorilla and chimpanzee). Briefly, using 5782 LTR events identified by RepeatMasker, we developed a pipeline to identify syntenic regions for each event and search for the conserved ones. We were able to detect ~200 human specific events adding more than 150 events to previously identified human specific Solo-LTRs. These data and future analysis may contribute to understanding the structural and functional impact of endogenous retroviruses on human genome.

USING CANCER TO INVESTIGATE THE INTERACTION BETWEEN CODON USAGE AND tRNA ABUNDANCE

Konrad LM Rudolph[1], Bianca M Schmitt[2], Claudia Kutter[2], Duncan T Odom[2], John C Marioni[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom, [2]Cancer Research UK, Cambridge Institute, Cambridge, United Kingdom

Codons encode the sequence of amino acids in a molecule of mRNA. The codon usage of a cell is thus dictated by the transcriptome of the cell at a given time point. The interactions at the codon–anticodon interface during protein synthesis are formalised in the genetic code. We have previous demonstrated that the codon usage and its interaction with tRNA anticodons remains remarkably stable, even across the variability of the transcriptome during mammalian development (Schmitt, Rudolph *et al.*, Genome Res., 2014).

However, it has also been hypothesised that well-defined subsets of cell-type-specific protein-coding genes do in fact possess variable codon usage, and that changes in the tRNA abundance lead to an anticodon pool highly adapted to these protein-coding genes, thus ensuring their efficient translation. In particular, this has been reported for proliferation-driving genes, which are highly expressed in cancer.

Cancer is characterised by uncontrolled cellular expansion, driven by distinctive gene expression programmes. Here, we report our efforts to reconcile the homoeostasis of the tRNA transcriptome we observed previously with the dynamic changes of the tRNA gene expression in cancer. We predict that high gene expression of proliferation might be coupled with corresponding changes in the tRNA transcriptome.

IMPROVING COMPUTATIONAL PREDICTION OF MISSENSE VARIANTS PATHOGENICITY FOR CLINICALLY RELEVANT GENES

Anna Rychkova[1], MyMy C Buu[2], Curt Scharfe[3], Martina Lefterova[3], Justin Odegaard[3], Carlos Milla[2], Iris Schrijver[2,3], Carlos D Bustamante[1]

[1]Stanford University, Genetics Department, Stanford, CA, [2]Stanford University, Pediatrics Department, Stanford, CA, [3]Stanford University, Pathology Department, Stanford, CA

Rapid, accurate, and inexpensive genome sequencing promises to transform medical care. A critical hurdle to enabling personalized genomic medicine is predicting the functional impact of novel genome variation. This is a particularly pressing problem at “clinically relevant genes” where some mutations are already known to impact phenotype (such as the BRCA1/2 cancer susceptibility genes), but for which we have an incomplete map of how genotype impacts clinical phenotype. Differentiating “benign” from “pathogenic” genetic variants is currently challenging and oftentimes physicians are left with the unsatisfying and inconclusive result that a patient carries a “Variant of Unknown Significance” (VUS). Existing computational approaches to variant classification all suffer from low overall accuracy rates. Their poor performance limits the general utility of these tools in the determination of whether a novel genetic variant at a clinically relevant gene is actually related to the disease of interest and limits accurate assessment of whether incidental findings ought to be disclosed, and of how significant these are.

We aim to develop a gene-specific variant pathogenicity assessment method of increased accuracy. Our main strategy is the use of a meta-predictor, which combines information from the most promising available tools supplemented with protein structural and stability features, as well as relevant individual level phenotypic data. Here we are presenting the results of this approach as designed for missense variants in the CFTR gene. The variant data used for training and testing were derived from a variety of sources, including public datasets (The Clinical and Functional TRanslation of CFTR (CFTR2), The database of Genotypes and Phenotypes (dbGaP), The Exome Aggregation Consortium (ExAC)) and internal datasets (The Stanford Cystic Fibrosis Center, The Stanford Molecular Pathology Laboratory). We used several machine learning algorithms to train our model and found increased overall performance when compared with separate predictors. In addition, the use of individual sweat chloride concentrations, helps to increase the performance even further, showing the importance of considering additional disease-specific features. We are planning to apply the current strategy for other clinically relevant genes as part of the Clinical Genome Resource project.

PREPARING COHORTS OF WHOLE GENOMES FOR COMMUNITY ANALYSES

William J Salerno[1], Matthew N Bainbridge[1], Adam C English[1], Mike Dahdouli[1], Simon White[1], Xiaoming Liu[2], Naryanan Veeraraghavan[1], Shalini N Jhangiani[1], Donna M Muzny[1], Eric Boerwinkle[1,2], Richard A Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]University of Texas Health Science Center at Houston, Human Genetics Center, Houston, TX

To improve high-throughput whole-genome sequencing analysis we developed Mercury2, a whole-genome, petabase-scale upgrade to our Mercury infrastructure that calls and prioritizes small and structural variants using consensus methods, whole-genome annotation sources, and comparison to large control cohorts and robustly characterized personal genomes. Mercury2 comprehensive genomes provide the appropriate biological, epidemiologic and medical contexts, with special attention to previously under-analyzed genomic features such as non-coding regions and large, complex variants.

Mercury2 is optimized for the petabases of whole-genome data generated by the Illumina X Ten platform. Mapping, quality control, and variant calling have a raw-data-to-annotated-variants turnaround time of less than a week. Mercury2 SNVs and Indels are called with ATLAS3 and annotated with whole-genome genomic features including regulatory features from Encode, GTEx, promoter and enhancer sites from FANTOM5 and deleteriousness scores from CADD, RegulomeDB, and Funseq2. Structural variants (SVs) are identified with Parliament, a consensus SV tool that automates assembly-based force calling. Mercury2 re-analysis options include unaligned read remapping to non-reference sources, integrated RNA-seq and miRNA-seq, teleomere assessment, and priority region refinement (read-stitching, local assembly, deep SV calling). Variants are stored in a data warehouse containing results from a high-confidence gold genome and large cohorts that can be used for comparison to ethnic-matched case groups. All variants are prioritized based on annotation and commonality to large cohorts and the gold genome, which is especially relevant to SV analysis given the relative dearth of available annotation compared to that of smaller variants.

To validate Mercury2, we analyzed 10 deep whole-genome trios. ATLAS3 and Parliament identified ~500 putative de novo events per family. Of these, less than 10 per variant type (SNV, indel, SV) intersected protein-coding regions. Annotation summaries of all variants are also provided. Finally, these findings are placed in the context of similar findings within the ARIC cohort and Mercury2 titrations of the HS1011 genomic data, which include 100 bp and 150 bp (PCR-free) short reads and ~10 kbp long reads.

Deployed on the cloud-based service DNAnexus, Mercury2 delivers to users "BAM-free" project-ready summary files, visualization tools, and an automated re-analysis toolbox.

STRUCTURAL VARIATION AMONG RHESUS MACAQUES IDENTIFIED USING THE PARLIAMENT SOFTWARE

Shruthi Ambreth[1], William Salerno[1], Adam English[1], Muthuswamy Raveendran[1], David R Deiros[1], Laura Cox[2], Betsy Ferguson[3], Eric Vallender[4], Michael Kubisch[5], Sree Kanthaswamy[6], David G Smith[6], Kim C Worley[1], Donna M Muzny[1], Richard A Gibbs[1], Jeffrey Rogers[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Texas Biomedical Res. Inst., Southwest National Primate Res. Center, San Antonio, TX, [3]Oregon Health Sciences Univ., Oregon Nat. Primate Res. Ctr., Beaverton, OR, [4]Harvard Medical School, New England Primate Res. Ctr., Southborough, MA, [5]Tulane Univ., Tulane Nat. Primate Res. Ctr., Covington, LA, [6]Univ. of California, California Primate Res. Ctr., Davis, CA

Recent advances in human genomics have clearly demonstrated that structural variation is common in the human genome and has significant functional effects. The number of diseases shown to result directly from structural variation, either de novo events or segregating variants, is growing rapidly. However, very little is known about structural variation in nonhuman primate genomes. The rhesus macaque (*Macaca mulatta*) is the most widely used nonhuman primate in biomedical research, but has not been investigated for structural variation (SV). Identification and characterization of SV in the rhesus genome will provide new information for understanding genome structure and dynamics in laboratory primate models and is likely to identify structural variation that will have phenotypic consequences relevant to macaque models of human disease. Parliament is an integrated pipeline that consolidates results from several SV detection programs (Pindel, CNVnator, Delly and Breakdancer) to evaluate an individual genome for insertions, deletions and other rearrangements of varying sizes. Using deep short-read whole genome sequence data for 10 rhesus macaques (9 Indian-origin, 1 Chinese-origin), we identified 370,040 distinct putative SVs between 100 bp and 100 kbp (186,067 deletions, 95,519 insertions, and 88,454 uncategorized breakpoints) in at least one individual with at least one program. The average number of events per individual is 95,237. There were 22,711 deletions and 2,491 insertions detected by more than two programs in any individual, and 13,325 deletions and 811 insertions were found by multiple programs across all individuals. These are the initial results from an on-going broader survey of SV events in this species, but demonstrate both the utility of Parliament for analysis of nonhuman mammalian genomes and constitute the first estimates of the genome-wide prevalence of SV events in this important primate model organism.

SELENOPROTEIN EXTINCTION IN *DROSOPHILA* OCCURRED CONCOMITANTLY TO GENOME CATASTROPHES.

Didac Santesmasses[1], Marco Mariotti[1], Salvador Capella-Gutierrez[1], Silvia Perez[1], Andrea Mateo[2], Montserrat Corominas[2], Toni Gabaldón[1], Roderic Guigó[1]

[1]Centre for Genomic Regulation, Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat de Barcelona, Departament de Genètica and Institut de Biomedicina, Barcelona, Spain

Selenoproteins contain Selenocysteine (Sec), the 21st amino acid. Incorporation of Sec occurs at redefined UGA codons, normally a stop signal, by a complex molecular mechanism that involves a dedicated set of factors. Selenoproteins are present in the three domains of life: Archaea, Bacteria and Eukarya.
The human genome encodes 25 selenoprotein genes, with many of them being conserved as selenoproteins across animals. However, insects show reduced selenoproteomes and, together with nematodes, are the only metazoan taxa where selenoprotein losses have been reported. In insects lacking selenoproteins, all Sec-encoding genes were either lost or converted to Cys homologues. The cell machinery for Sec production and insertion degenerated concomitantly. The distribution of these species in the phylogenetic tree implies that several independent Sec extinction events occurred in parallel insect lineages.
To investigate the mechanism of Sec loss, we focused in the most recent, and hopefully most insightful, Sec extinction known, the one described in *Drosophila willistoni*. We sequenced the full genome of 8 species in the *saltans* group, phylogenetically sister to the *willistoni* group. The *saltans* group contains both species with and without selenoproteins, and we mapped at least 3 Sec extinction events. Although independent, these events happened in a single *Drosophila* lineage. Interestingly, a few genomic features set the saltans/willistoni lineage apart from the rest of drosophilas: a lower GC content, and a lower codon bias (favoring AT nucleotides). Moreover, we compared the transcriptome of several species in different Drosophila lineages, and the clustering based on gene expression does not recapitulate the phylogenetic tree, suggesting a general shift in the cellular transcriptional program.
We formulated a model for the Sec extinction process in drosophilas. First, the functions of essential Sec-encoding genes must be transferred to a non-selenoprotein gene (Cys conversion) or become unimportant (allowing gene loss). Then, mutations disrupting the Sec coding ability can be tolerated, and the Sec machinery quickly degenerates. The selenoprotein SPS2, necessary for Sec production, is always the last selenoprotein to be lost, since it is needed for expression of the other selenoproteins.

INTRA- AND INTERHOST EVOLUTION OF LASSA AND EBOLA VIRUSES FROM WHOLE GENOME SEQUENCING

Kristian G Andersen[1,2], Jesse Shapiro[1,2,3], Christian B Matranga[2], Rachel Sealfon[2,4], Stephen F Schaffner[1,2], Andreas Gnirke[2], Joshua Z Levin[2], Christian T Happi[5,6], Robert F Garry[7], Pardis C Sabeti[1,2]

[1]Harvard University, Organismic and Evolutionary Biology, Cambridge, MA, [2]Broad Institute, Infectious Disease Initiative, Cambridge, MA, [3]University of Montreal, Biological Sciences, Montreal, Canada, [4]MIT, CSAIL, Cambridge, MA, [5]Irrua Specialist Teaching Hospital, ILFRC, Irrua, Nigeria, [6]Redeemer's Univesity, Tulane Health Sciences Center, Lagos, Nigeria, [7]Tulane University, Tulane Health Sciences Center, New Orleans, LA

Lassa virus and Ebola virus are both biosafety level-4 pathogens currently causing severe hemorrhagic fevers with high case fatality rates in West Africa. Their epidemiological and transmission dynamics are quite different, however, since Lassa fever, unlike Ebola, is a long-standing endemic disease in the region and involves frequent transmission events from the animal reservoir.

We have used deep viral genomic sequencing of hundreds of patient samples for each virus to compare the epidemiological and evolutionary characteristics of the two viruses. As expected of an endemic virus, Lassa displays an order of magnitude higher genetic diversity at the population level than Ebola. More surprisingly, Lassa also has considerable diversity within individual hosts, and shows evidence for intra-host selection for immune escape. The longer evolutionary history of Lassa virus is reflected in functional differences between viral clades, including different translational efficiency, viral load, case fatality rate and different degrees of codon adaptation. Ebola virus data, on the other hand, reflect a rapid and rapidly changing outbreak from a single source, with little within-host evolution.

DE NOVO ASSEMBLY AND STRUCTURAL VARIATION ANALYSIS OF RICE USING PACBIO LONG READ SEQUENCING: THE RETURN OF REFERENCE QUALITY GENOMES

Michael C Schatz[1], James Gurtowski[1], Sara Goodwin[1], Lyza Maron[2], Maria Nattestad[1], Hayan Lee[1], Eric Antoniou[1], Panchu Deshpande[1], Susan McCouch[2], W. Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cornell University, Department of Plant Breeding and Genetics, Ithaca, NY

Rice (Oryza sativa) is one of the most important crops in the world. It is the predominant staple food for a large fraction of the world's population, especially in Asia, and provides more than one fifth of the calories consumed by humans worldwide. In 2005, the International Rice Genome Sequencing Project published the first rice genome of the Nipponbare variety using a high quality but expensive BAC-by-BAC approach. This sequence, along with a few other lower quality shotgun assemblies, has become an essential resource as the backbone for SNP analysis, RNA-seq, and other mapping-based assays of rice. However, these mapping-based approaches are challenged to properly analyze structural variations between the varieties, including of the hundreds of genes that differ between the major subpopulations.

To explore the true genomic complexities, we sequenced the Indica variety IR64 to more than 100x coverage using PacBio long read sequencing and also with Illumina short reads using the "Allpaths-recipe" with fragment, short-jump and long-jump libraries. After error correcting the PacBio reads using HGAP, more than 22x coverage was available in reads over 10kbp including many reads over 50kbp. We then assembled the PacBio reads using the Celera Assembler to produce a true reference quality assembly: the contig sizes approaches that of the BAC-by-BAC Nipponbare assembly, 4.0Mbp contig N50 versus 5.1Mbp respectively, compared to only 20kbp for the Illumina-only assembly. The reference quality PacBio assembly, with contigs spanning nearly entire chromosome arms, gives us significantly greater power to analyze gene content, regulatory regions, and synteny across large genomic spans compared to mapping or short read assembly. From this we have isolated thousands of regions specific to Indica not present in Nipponbare spanning more than 20 megabases of sequence that was previously unresolved from the short read assembly. Many of the most significant differences contain genes and other loci associated with agriculturally important traits including hybrid sterility, submergence, and drought tolerance.

PGRN-SEQ V.2: A SECOND-GENERATION CAPTURE-SEQUENCING REAGENT FOR PROSPECTIVE SEQUENCING OF CLINICALLY RELEVANT PHARMACOGENETIC LOCI.

Steven Scherer[1], Robert Fulton[2], Nicole Leahy[3], Daniel Burgess[3], Deborah Nickerson[4], Elaine Mardis[2], Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Washington University, The Genome Institute, St. Louis, MO, [3]Roche NimbleGen, Research and Development, Madison, WI, [4]University of Washington, Genome Sciences, Seattle, WA

Using lessons learned using the first version of this reagent, the Pharmacogenomics Research Network's Deep Sequencing Resource (PGRN-DSR) worked together with network members, community experts and NimbleGen to design a second generation capture-sequencing probe set that builds on successes realized with PGRN-seq v.1, including its adoption by the eMERGE Network's PGx Project. The primary aim was to modify the target list to reflect advances in the field of pharmacogenomics, address loci left unaddressed in the original and ensure that the output was more clinically relevant. To this end, targets now include all of the current Clinical Pharmacogenetics Implementation Consortium (CPIC) drug-gene pairs as well as other sources while maintaining a the original's low cost. We will present data outlining the performance and validation of this new addition to the genomics translation toolkit.

MEASURING THE RATE AND HERITABILITY OF AGING IN SARDINIANS USING PATTERN RECOGNITION

David Schlessinger[1], Eric D Sun[1], Yong Qian[1], Gonçalo R Abecasis[2], Francesco Cucca[3], Jun Ding[1], Ilya Goldberg[1]

[1]National Institute on Aging/NIH, Laboratory of Genetics and Genomics, Baltimore, MD, [2]University of Michigan, Department of Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, [3]National Research Council, Monserrato, Institute of Genetic and Biomedical Research, Cagliari, Italy

It is widely accepted that individuals age at different rates. A method that measures physiological age independently of chronological age could therefore be a first step to reveal contributing mechanisms; but searches for individual biochemical markers of physiological age have had limited success. In this study, we rather assessed the extent to which an individual's chronological age could be determined as a composite score inferred from a broad range of biochemical and physiological data. Data were collected in a longitudinal population study in Sardinia. The study (the "SardiNIA" project at https://sardinia.irp.nia.nih.gov/) includes measures of environmental factors and family structures to facilitate both epidemiological and genetic analyses. We used pattern recognition and machine learning strategies on data for the 6,000 participants in the study, who range in age from 12 to 81. The best predictive models were determined from multiple combinations of dimensionality reduction, classification, and regression algorithms. They reached very strong correlations ($R > 0.9$) between predicted and actual ages, and showed relative stability in successive visits of the same individuals ($R>0.5$). We then defined an Effective Rate of Aging (ERA) for each participant, a continuous trait measured as the ratio of an individual's predicted value to his/her chronological age. The inference that individuals have a characteristic rate of aging is supported by findings that in the entire cohort, the inferred values of ERA showed genetic heritability of 40%. This has been sufficient to initiate genome-wide association studies of the effect of variation in individual genes on the rate of aging.

ANALYSIS OF LARGE STRUCTURAL VARIANTS IN 2,200 WHOLE-GENOME SEQUENCED MYOCARDIAL INFARCTION CASES AND CONTROLS

Ellen M Schmidt[1], Jin Chen[2], Oddgeir L Holmen[3], Kristian Hveem[3], Ryan E Mills[1], Cristen J Willer[1,2]

[1]University of Michigan, Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Internal Medicine, Division of Cardiovascular Medicine, Ann Arbor, MI, [3]Norwegian University of Science and Technology, HUNT Research Centre, Public Health and General Practice, Levanger, Norway

Myocardial infarction (MI) is a major cause of death throughout the world, with both common and rare single genetic mutations contributing to early-onset risk. Discovery of large structural variants (SVs) will give a more complete understanding of the genetic etiology of this complex disease. We perform whole-genome sequencing at 5x coverage of 2,200 individuals from Nord-Trøndelag, Norway as part of the HUNT study. The cohort includes cases with early-onset MI, and age- and sex- matched healthy controls. We hypothesize that there are different frequencies of large structural variants in MI cases compared to controls and apply established and complementary SV detection algorithms to identify deletions, duplications, inversions, and translocations. Using the non-redundant union set of calls, we test for differences in both the prevalence and distribution of large structural variation among individuals with MI compared to their matched controls. In addition to analyzing association of genotypes with MI status, we examine the relationship between SVs and single nucleotide variants associated with coronary artery disease and related traits identified by genome wide association studies. We also investigate novel or Norwegian-specific SVs not present in the 1000 Genomes Project. Structural variants identified here will be imputed into a larger HUNT cohort for further study. This research provides further insight into the genetic architecture underlying myocardial infarction and other heart disease-related clinical phenotypes.

TAKING ADVANTAGE OF AN EVOLVING HUMAN REFERENCE GENOME ASSEMBLY

Valerie A Schneider[1], Tina Graves-Lindsay[2], Paul Flicek[3], Richard Durbin[4]

[1]NIH, NCBI, Bethesda, MD, [2]Washington University School of Medicine, Genome Institute, St. Louis, MO, [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom, [4]Wellcome Trust Sanger Institute, Genome Informatics, Hinxton, Cambridge, United Kingdom

GRCh38, the current major release of the human reference genome assembly, continues the evolution of a resource critical to both basic and clinical research. Since 2009, the Genome Reference Consortium (GRC), the group responsible for curation and updates to the reference assembly, has used a non-haploid assembly model in which alternate loci scaffold sequences provide additional sequence representations for genomic regions exhibiting high diversity within the human population. While only a handful of alternate scaffolds were included in GRCh37, there are more than 200 such scaffolds representing over 175 genomic regions in GRCh38. Although a reference containing multiple sequence representations is recognized to improve NGS read mapping and variant calling, adopted use of alternate loci has been slow due to a limited number of analysis tools that can handle the allelic duplication they introduce into the reference assembly and a lack of support for this more complex data structure in commonly used file formats. We will discuss the need for resources that support the current assembly model, considering both its relationship to various efforts to develop graph-based variant representations and the needs of a diverse community of researchers that utilize genomic sequence data in a variety of ways. We will also highlight features that make GRCh38 superior to previous reference assembly versions and present ongoing GRC efforts to improve the reference assembly. Newer sequencing, mapping and assembly technologies, as well as new genomic resources, are making it possible to correct and represent genomic regions that were previously considered intractable and to provide additional alternate sequence representations in the reference. The GRC provides genome updates on a quarterly basis in the form of patch releases. Like the alternate loci, the GRC fix and novel patches are scaffold sequences aligned to the chromosomes. We will provide examples of recent patches and discuss the implications and usage of the two patch types with respect to various analyses and resource development. The work presented is a collaborative effort of the GRC member institutions: NCBI, EMBL-EBI, The Genome Institute at Washington University and The Wellcome Trust Sanger Institute.

SELECTION AND ASSORTATIVE MATING SHAPE THE GENOMES OF HYBRID SWORDTAIL FISH

Molly Schumer[1], Mattie Squire[2], Gil Rosenthal[2], Peter Andolfatto[1,3]

[1]Princeton University, Ecology and Evolutionary Biology, Princeton, NJ, [2]Texas A&M University, Biology, College Station, TX, [3]Princeton University, Lewis Sigler Institute for Integrative Genomics, Princeton, NJ

Hybrid populations are a window into the mechanisms of reproductive isolation between species. Though hybrid zones have been used to study the genetic barriers between species, few studies have investigated how these populations evolve. Hybrid populations between the swordtail fish *Xiphophorus birchmanni* and *X. malinche* formed approximately 40 generations ago as a result of anthropogenic environmental disturbance and have experienced strong selection due to many genetic incompatibilities distinguishing the parental species (Schumer et al. 2014). In the present study, we analyze ~500,000 ancestry informative markers throughout the genome to investigate how hybrid populations have evolved over the last 20 generations. We show by simulation that incompatibility selection has played a major role in the current genomic structure of hybrid populations. Interestingly, in one hybrid population, two hybrid genotype clusters have emerged, one biased toward each parent. We show that these clusters have persisted over many generations, and are sustained by nearly 100% assortative mating between the two hybrid types. Our results highlight the joint effects of selection and behavior in structuring genomic variation in hybrid populations.

COMPARATIVE ANALYSIS OF THE DNA METHYLOME WITHIN INCLUDED AND EXCLUDED ALTERNATIVELY SPLICED EXONS

Renee L Sears, Ting Wang

Washington University School of Medicine, Genetics, St. Louis, MO

Observed in over 90% of multi-exon human genes, alternative pre-mRNA splicing (AS) leads to the diversification of protein products in both a tissue- and species- specific manner. Interestingly, AS frequency decreases as a function of evolutionary distance from primates. This provides a partial explanation for the phenotypic complexity of our lineage without additional genetic coding material. In the brain, AS is more frequent than in other tissues and the "percent spliced in"(PSI) of alternatively spliced exons (ASEs) is more conserved when compared to AS in different tissues and across species respectively.

Evolutionarily, AS is thought to have arisen through three methods: exon shuffling, repurposing of intronic sequences as exons, and the transformation of a conserved constitutive exon into an alternative exon primarily through relaxation of the 5' splice site. Slowing the elongation rate of RNA polymerase II-mediated transcription increases the inclusion of ASEs by allowing splicing machinery to recognize weak or hidden splice sites. Importantly, DNA methylation has been correlated with a decrease in elongation rate. Further, genetic and chemical inhibition of DNA methylation has been demonstrated to cause aberrant splicing.

Here we explore the relationship between CpG methylation and AS in the human brain. Using stranded RNA-sequencing and bisulfite sequencing data from four normal human middle frontal gyri, we show that included ASEs, as defined by PSI, are more methylated than their excluded counterparts. As expected, exons with relatively weak 5' splice sites exhibit a stronger association between PSI and average CpG methylation. These findings substantiate *in vitro* work in human cells and *in vivo* studies in insects that demonstrated an enrichment of DNA methylation in included ASEs. These observations raise the question of whether abnormal CpG methylation patterns in cancer and other diseases cause aberrant AS and the role CpG methylation plays in tissue- or species- specific AS.

COMPREHENSIVE ANALYSIS OF *DE NOVO* STRUCTURAL VARIATION IN AUTISM BY WHOLE GENOME SEQUENCING.

Jonathan Sebat[1], William Brandler[1], Danny Antaki[1], Madhusudan Gujral[1], Amina Noor[1], Christina Corsello[2], Guan Ning Lin[1], Lilia Iakoucheva[1], Suzanne Leal[3], Timothy Chapman[1]

[1]UC San Diego, Institute for Genomic Medicine, Department of Psychiatry, La Jolla, CA, [2]Rady Children's Hospital, Autism Discovery Institute, San Diego, CA, [3]Baylor College of Medicine, Houston, TX

Whole Genome Sequencing (WGS) provides a more complete ascertainment of *de novo* mutation compared to microarrays or targeted sequencing. We investigated the contributions of *de novo* structural variants (SVs) copy number variants to autism spectrum disorder (ASD) by WGS and custom mutation detection in 235 subjects, including 71 with ASD, 26 sibling controls and their parents. Deletions, Duplications, Inversions and complex rearrangements were detected using a random forest classifier *forests*, and de novo mutations were detected using a Gaussian Mixture Model-based *SVgenotyper*. We demonstrate a high rate of structural mutation in controls (15% of subjects), The overall rate of de novo SV was not elevated in ASD; however, mutations in cases differed significantly with respect to their sizes and distribution among functional categories of genes and were enriched at loci that are mutated in independent studies of neurodevelopmental disorders. We estimate that *de novo* missense, LOF and CNVs combined contribute to risk in 30% of cases.

THE IMPACT OF HIGHLY POLYMORPHIC REGIONS ON HTS RELATED STUDIES.

Fritz J Sedlazeck[1,2], Naoki Osada[3], Aya Takahashi[4], Michael Schatz[1], Arndt von Haeseler[2]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna, Vienna, Austria, [3]National Institute of Genetics, Division of Evolutionary Genetics, Shizuoka, Japan, [4]Tokyo Metropolitan University, Department of Biological Sciences, Tokyo, Japan

The advent of high throughput sequencing (HTS) has boosted the variety of sequencing projects related to molecular biology and medicine. Mapping reads to a reference genome is one of the fundamental steps in HTS related analysis, including for variant identification, transcript abundance estimation, and many others. However, mapping reads from a heterozygous sample to a reference genome can lead to biased results and inability to identify the alternate alleles i.e. “reference mapping bias”. We studied this bias in several of the most widely used read mapping algorithms including BWA/BWA-MEM, Bowtie/Bowtie2, TopHat/TopHat2, and our own NextGenMap, in an F1 cross from inbred lines of D. melanogaster Mel 6 x D. melanogaster RAL774 where a precise catalog of heterozygous positions could be determined from the parental reference genomes. By examining regions with different rates of heterozygosity, we show that both SNP calling and transcript abundance analysis were highly skewed against the alternate alleles proportional to the frequency of heterozygosity, including completely mis-analyzing certain regions as having allele-specific expression. In contrast, we show that the highly sensitive mapper like NextGenMap are less affected by reference bias and thus more suitable for reliable analyses of HTS data for polymorphic samples.

DOG DIVERSITY IS SHAPED BY A CENTRAL ASIAN ORIGIN FOLLOWED BY GEOGRAPHICAL ISOLATION AND ADMIXTURE.

Laura M Shannon[1], Ryan Boyko[2], Marta Castelhano[3], Liz Corey[3], Jessica J Hayward[1], Michelle White[1], Carlos D Bustamante[4], Rory Todhunter[3], Robert K Wayne[5], Adam R Boyko[1]

[1]Cornell University School of Veterinary Medicine, Biomedical Sciences, Ithaca, NY, [2]Yale University School of Public Health, Epidemiology of Microbial Diseases, New Haven, CT, [3]Cornell University School of Veterinary Medicine, Clinical Sciences, Ithaca, NY, [4]Stanford School of Medicine, Genetics, Stanford, CA, [5]University of California Los Angeles, Department of Ecology and Evolutionary Biology, Los Angeles, CA

Dogs were the first domesticated species, originating between 15 and 30,000 years ago. Present-day domestic dogs consist primarily of two specialized groups of animals--a highly diverse set of nearly 400 pure breeds and a far more populous group of free-ranging animals adapted to a human commensal lifestyle (village dogs). Village dog populations show more genetic diversity than purebred dogs making them vital for unraveling dog population history.

Using a semi-custom 183,000 marker genotyping array, we conducted the first large-scale survey of autosomal, mitochondrial and Y chromosome diversity in 3800 purebred dogs (2000 males) from 160 breeds and 400 village dogs (250 males) from 30 countries.

Despite many universal mitochondrial and Y chromosome haplogroups, geographic structure is readily apparent suggesting both isolation and admixture have shaped genetic diversity in village dog populations. Some extant village dog populations (notably those in the Neotropics and the South Pacific) are almost completely derived from European stock while others are clearly admixed between indigenous and European dogs. Importantly, many populations---including those of Vietnam, India and Egypt---show no evidence of European admixture. These populations exhibit a clear gradient of short-range linkage disequilibrium consistent with a Central Asian domestication origin.

Previous genetic studies have implicated Southern China, Europe or the Middle East as likely origins for dogs, but contemporary patterns of diversity are not solely the result of domestication. Haplotype diversity patterns have been affected by recent population expansion and admixture with co-localized canids as well as sampling biases, including relative sparse sampling from Central Asian populations. Unlike Southern China, there is strong evidence of ancient gray wolf populations in Central Asia around the time of dog domestication. As early humans in this region primarily subsisted on hunting large game, early dogs may have first assumed roles as hunters or guardians before becoming adapted to other working roles.

A SURVEY OF DNA METHYLATION POLYMORPHISM IN THE HUMAN GENOME IDENTIFIES ENVIRONMENTALLY RESPONSIVE CO-REGULATED NETWORKS OF EPIGENETIC VARIATION

Paras Garg, Ricky Joshi, Corey T Watson, Andrew J Sharp

Icahn School of Medicine at Mount Sinai, Genetics and Genomic Sciences, New York, NY

Understanding the causes and consequences of genomic variation is a major goal in the field of genetics. While studies such as the Hapmap and 1000 Genomes Projects have resulted in detailed maps of genetic variation, to date there are no robust maps of epigenetic variation in humans, and few insights have been made into their significance or underlying biology. Here we set out to define sites of common epigenetic variation in humans, that we term Variable Methylation Regions (VMRs). In order to avoid the confounder of cellular heterogeneity, we focused on DNA methylation data derived from normal populations assayed with the Illumina 450k array (n=58 to 111 individuals), representing five purified cell types: T-cells, B-cells, fibroblasts, neurons and glia. Using a robust approach we identify hundreds of VMRs in each cell type that show common variability in DNA methylation levels. We find that VMRs occur preferentially at enhancers and in 3' UTR regions, consistent with a functional role in regulating gene expression. We observed that at the majority of VMRs methylation is highly heritable, indicating that many are associated with underlying genetic variation. However, we also observed a subset of VMRs distributed across the genome that show highly correlated variation in trans. In contrast to most other epigenetically variable regions, VMRs that form co-regulated networks tend to have low heritability, differ between cell types and are enriched for specific biological pathways of direct functional relevance to each tissue. For example, in T-cells we defined a network of 61 co-regulated VMRs enriched for genes that form the T-cell receptor complex and play roles in T-cell activation; in fibroblasts a network of 21 co-regulated VMRs comprising all four *HOX* gene clusters that is highly enriched for control of tissue growth; and in glia a network of 66 VMRs enriched for roles in postsynaptic membrane organization. These VMR networks share common transcription factor binding sites that are significantly enriched within each network, indicating that the epigenetic state of these VMR networks is responsive to molecular signaling induced by environmental cues. By culturing fibroblasts under varying conditions of nutrient deprivation and cell density, we experimentally demonstrate that methylation of the *HOX* gene cluster network is responsive to environmental conditions, with methylation levels at these loci changing in a coordinated fashion in trans dependent on cellular growth. Our study provides the first detailed map of common epigenetic variation in the human genome, showing that both genetic and environmental causes underlie this variation.

INTRA-INDIVIDUAL VARIATION AND MEDIUM-TERM METHYLATION-EXPRESSION ASSOCIATION STUDY IN MONOCYTE FROM HEALTHY INDIVIDUAL.

Ryohei Furukawa[1], Tsuyoshi Hachiya[1], Hideki Ohmomo[1], Yuh Shiwa[1], Kanako Ono[1], Sadafumi Suzuki[2], Mamoru Satoh[1], Jiro Hitomi[1], Kenji Sobue[1], Atsushi Shimizu[1]

[1]Iwate Medical University, Iwate Tohoku Medical Megabank Organization, Iwate, Japan, [2]Keio University, School of Medicine, Tokyo, Japan

DNA methylation state (DNAm) and gene expressions different from the types of tissues and cells. As gene expressions are controlled by DNAm, it could define that the DNAm set the environment of gene expressions. It is also known that, while DNAm remains stable at steady state in each cell, they can be affected by environmental change, extrinsic stimuli or drug administration.

As a pilot study to use these omics data in healthy cohort studies, we measured the DNAm fluctuations within individuals over several months. First, we collected blood from two healthy male volunteers aged thirties for six times over a period spanning a month, obtained $CD14^{high}CD16^{low}$ monocytes using cell sorter (SH800, Sony). Methylation levels of approximately 480,000 CpG probes were measured by Illumina Infinium HumanMethylation450 BeadChip microarrays, and were evaluated by β-values, which are the ratio of the methylated probe intensity and the total signal intensity (ranged from 0 to 100 %). For >99.5% of CpG probes, the maximum differences of DNAm levels between the six times were less than 20%.

Next, we examined the medium-term dynamics and stability of gene expression and DNAm of monocytes from one volunteer for 24 times in three months. We conducted RNA sequencing, and then detected methylation-expression associations (MEA) using a linear regression model and the ANOVA test to investigate whether these variations of DNAm level affect expression of neighboring genes. Of 341,233 CpG positions located on the promoter or gene body regions, 380 (0.11%) were associated with the expression of neighboring genes with a significance level of 0.001. In these MEA CpG positions, the maximum differences of DNAm levels exceeded 20% at six CpG probes, and in approximately 92% of the genes detected by MEA analysis, the maximum difference of the $\log_{10}$[FPKM+1] were less than 0.3.

From these results, the drastic variation of DNAm does not seem to contribute much to the dynamic change of gene expression. However, a possibility that small changes of DNAm may affect the gene expression is required to be explored.

PRIVACY LEAKS FROM BEACONS

Suyash S Shringarpure, Carlos D Bustamante

Stanford University, Genetics Department, Stanford, CA

In the coming decade, a great deal of human genomic data will be collected in the context of patient care along with linked (phenotypically rich) Electronic Health Records (EHRs). A major goal of the human genomics community is to enable efficient sharing, aggregation, and analysis of these data to understand the genetic contributors of health and disease.

The Beacon Project allows international sites to share genetic data in the simplest of all technical contexts. Beacons are servers installed by institutions that users can query for information about genomic data available at the institution. Queries are of the form ``Do you have a genome that has a specific nucleotide (e.g.,`A') at a specific genomic position (e.g., position 11272 on chromosome 1)" and the server can answer "Yes" or "No". Nearly 20 beacons have been set up at various sites at this time.

We hypothesize that beacons are susceptible to re-identification attacks, i.e., given a query individual's genome, we can predict with high confidence whether or not this individual was among the set of genomes used to create the beacon. This is troubling since beacons typically summarize data at a single institution, often characterized by having a particular phenotype or disease of interest. For instance, identifying that a genome is present in the beacon at a cancer research institute suggests that the subject is likely to have cancer. Thus, beacons can leak not only membership information but also phenotype information.

We have developed a likelihood ratio test to predict whether a given individual genome is present in the beacon database. We show that for beacons with ~100 individuals, it is sufficient to query ~1000 unlinked SNPs from an individual genome to detect membership in the beacon with 99% power and 1% false positive rate. For empirical evaluation, we set up a private beacon using 65 CEU individuals from Phase 1 of the 1000 Genomes project. Using 250 SNPs, membership within the beacon can be detected with 95% power and 5% false positive rate.

We also extend our framework to show that relatives can be detected in a beacon (although with either a reduction in power or an increase in the number of SNPs queried). Preliminary results suggest that while combining datasets from different populations reduces the possibility of detection by increasing the number of SNPs required, this may not be sufficient if an attacker has access to whole genomes.

Our results show that beacons can leak unintended phenotypic information about subjects if enough care is not taken to ensure privacy. They have implications about necessary precautions that must be taken when proposing new data sharing mechanisms.

MISSING HERITABILITY IN DIVERSITY OUTBRED MOUSE POPULATION

Petr Simecek, Gary A Churchill

The Jackson Laboratory, Center for Genome Dynamics, Bar Harbor, ME

Although last two decades provided a valuable insight into the genetic basis of many human diseases, the polymorphisms identified by genome wide association studies typically capture only a small proportion of estimated heritability. This problem has been referred as "missing" or "hidden" heritability. Various explanations have been proposed including epistasis, rare casual variants and a combination of large number of loci with small effects.

We have observed a phenomenon similar to hidden heritability for a number of physiological measurements in Diversity Outbred (DO) population, a heterogeneous stock derived from eight inbred founder strains. Unlike fully inbred mouse strains, DO mice are genetically unique with the normal level heterozygosity and genetic diversity, recapitulating that of humans. The advantage of DO in comparison to human populations is our ability to control and manage the environmental factors and breeding schema, thereby avoiding the possible confounding.

Analogically to human studies, while the estimated heritability is significantly positive for the most of the observed traits, a percentage of heritability explained by the identified loci is relatively small. We analyze the power of mixed-effects association methods, modelling hidden heritability as a random animal effect, not only for DO but also other classical mouse populations like F2 intercross and backcross. Further, we examine how the heritability estimate depends on a definition of genetic relationship matrix (GRM) and show that for DO a haplotype level GRM predicts the observed traits better than a SNP level GRM.

Finally, we show how to use protein expression data to narrow the identified loci to gene candidates. As an example, we mapped a quantitative trait locus for blood triglyceride concentration to Chr 5 and identified *Sept11* as a casual gene.

STRUCTURAL VARIATION ON THE Y-CHROMOSOME IN THE DANISH POPULATION

Laurits Skov[1], Mikkel H Schierup[1,2], Simon Rasmussen[3], Siyang Liu[4], Palle Villesen[1,2]

[1]Aarhus University, Bioinformatics Research Center (BiRC), Aarhus, Denmark, [2]Aarhus University, Centre for Integrative Sequencing (iSEQ), Aarhus, Denmark, [3]Technical University of Denmark, Center for Biological Sequence Analysis,Department of Systems Biology, Kgs. Lyngby, Denmark, [4]University of Copenhagen, Department of Biology, Copenhagen, Denmark

The Y chromosome has often been used as a phylogenetic marker due to its lack of recombination and direct inheritance from father to son. The discovery of new haplotypes and resolving structural variation in the Y chromosome is still a work in progress.
In the Danish Pan genome study 50 Danish trios were sequenced with high coverage (80x) and multiple insert size libraries up to 20 kb. The reads were De novo assembled into scaffolds with lengths of multiple Mb with around 3.1% gaps. 18 of the 50 trios had a father-son-pair meaning that the de novo assemblies could be compared and validated (in cases of no de novo mutations).
Here I present an extensive catalogue of unbiased structural variation on the Y chromosome including variations that are not detectable using traditional variation calling methods due to the lengths of the scaffolds.

The Danish Pangenome Consortium: Anders D. Børglum, Anders Krogh, Arcadio Rubio-García, Christian N. S. Pedersen, David Flores, David Westergaard, Ditte Demontis, Emil Rydza, Esben Nørgaard Flindt, Francesco Lescai, Hans Eiberg, Hao Liu, Jacob Malte Jensen, Jakob Grove, Jette Bork-Jensen, Jihua Sun, John Damm Sørensen, José M. G. Izarzugaza, Jun Wang, Junhua Rao, Laurits Skov, Karsten Kristiansen, Kirstine Belling, Kristoffer Rapacki, Lars Bolund, Mikkel H. Schierup, Ning Li, Ole Lund, Oluf Pedersen, Ou Wang, Palle Villesen, Piotr Chmura, Piotr Dworzynski, Rachita Yadav, Ramneek Gupta, Ruiqi Xu, Rune M. Friborg, Shengting Li, Shujia Huang, Simon Rasmussen, Siyang Liu, Søren Besenbacher, Søren Brunak, Thomas D. Als, Thomas Mailund, Thorkild I. A. Sørensen, Torben Hansen, Weijian Ye, Xiaofang Cheng, Xun Xu

INTEGRATIVE ANALYSIS OF AUTISM SPECTRUM DISORDERS

Jingjing Li[1], Minyi Shi[1], Zhihai Ma[1], Alexander Urban[1,2], Joachim Hallmayer[2], Michael Snyder[1]

[1]Stanford Center for Genomics and Personalized Medicine, Genetics, Stanford University, Stanford, CA, [2]Dept. of Psychiatry & Behavioral Sciences, Stanford School of Medicine, Stanford, CA

Genetic research in the past decade has implicated hundreds of loci in autism spectrum disorders (ASDs), and its extreme locus heterogeneity has prevented us from portraying a complete picture for its molecular etiology. To unravel the molecular basis of the disease, we built an integrative framework to identify molecular pathways significantly involved in autism by incorporating data from autistic genomes, brain transcriptomes and the human protein interactome. In a global scale, we deconstructed the human protein interactome, and identified a highly interacting module significantly affected by genetic alterations associated with ASD. Expression of this module was dichotomized with a ubiquitously expressed subcomponent and another subcomponent preferentially expressed in the corpus callosum. Our immunochemical analysis showed that the human corpus callosum is predominantly populated by oligodendrocyte cells, and our functional genomic analyses further revealed the significant involvement of this ASD-associated module in regulating the development of oligodendrocyte cells. In a finer scale, we mapped the ASD-associated mutations onto an entire collection of human protein complexes, and identified a specific set of complexes strongly associated with ASD, including the BAF complexes in neural progenitors or in neurons. This analysis allowed us to identify histone deacetylase as important regulators in ASD, whose perturbation led to considerable mis-expression of known ASD gene candidate orthologs in the embryonic mouse brain. Overall, our study at the systems level depicts the underlying molecular architecture of this disease, reveals novel ASD-associated gene candidates, and provides novel insights into the molecular basis of this disease.

LIGHTER AND RCORRECTOR: A SUITE FOR NEXT GENERATION SEQUENCING ERROR CORRECTION

Li Song[1,2], Ben Langmead[1,2], Liliana Florea[1,2]

[1]Johns Hopkins School of Medicine, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD

Next generation sequencing (NGS) has enabled fast and cost-efficient genome and transcriptome analyses. While abundant, reads produced by NGS platforms often contain low-rate sequencing errors. Error correction of sequences has been shown to significantly improve the quality of genome assembly and read alignment, which are critical for all downstream structural and functional analyses. We developed two software tools, Lighter and Rcorrector, to correct sequencing errors in short DNA and RNA reads, respectively, produced with Illumina instruments.

Both tools employ a k-mer spectrum approach, converting low frequency k-mers into high frequency ones for each read, taking into account characteristics of the data. For DNA sequences, reads are expected to cover the genome uniformly and follow a Poisson distribution. Lighter then uses a two-stage sampling method implemented with two Bloom filters to determine the set of all *solid* k-mers, i.e. high-frequency k-mers that are trusted to be error-free. Although not used explicitly in Lighter, the frequency threshold is defined genome-wide. Due to k-mer sampling and by avoiding to store weak k-mers, Lighter is very fast and considerably more space-efficient than previous error correctors, while maintaining comparable accuracy.

In contrast to DNA sequencing, read coverage for RNA sequencing is not uniform, due to the different expression levels of genes and transcripts in a sample. Additionally, alternative splicing may cause *bona fide* low frequency k-mers from a low expression isoform to be interspersed within a stretch of high-frequency k-mers. To address these challenge, we developed Rcorrector, an efficient tool to correct sequencing errors when the read coverage is non-uniform. Unlike Lighter, it takes into account the local gene context to classify a solid k-mer. Rcorrector uses a de Bruijn representation of all k-mers in the read set. At correction time it locates the path with the minimum number of changes to the read as the likely transcript of origin, based on which it determines the 'corrected' sequence. Compared with similar tools, including the only other RNA-seq error corrector – SEECER (Le et al. 2013), Rcorrector is faster and considerably more memory-efficient, at comparable or higher accuracy. It also produces better or comparable alignments and transcripts assemblies. Lighter and Rcorrector can be downloaded from https://github.com/mourisl/{Lighter,Rcorrector}.

Acknowledgments: NSF IIS-1349906 to B.L., and NSF IOS-1339134 and ABI-1356078 to L.F..

THE EUROPEAN VARIATION ARCHIVE

John D Spalding, Gary Saunders, Ignacio Medina, Cristina Y Gonzalez, Jag Kandasamy, Francisco J Lopez, Ilkka Lappalainen, Jacobo Coll, Jose M Mut, Tom Smith, Justin Paschall

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom

The European Variation Archive (EVA) is a genetic variation archive hosted at EMBL-EBI set up to support full access for all types of variation for all species. Uniquely, EVA is aiming to provide, within the same resource, an integrated view of all variation irrespective of size or type in any species. This allows single queries across both SNV and structural variants to ascertain the full range of genetic variation seen at a given locus. Both clinically relevant data from disease cohorts as well as control data from healthy populations are submitted to EVA. EVA is closely linked with ClinVar, both exposing ClinVar data and brokering submissions of new datasets. The EVA web interface and REST API allows full online queries of ClinVar data within the context of the EVA archive. Further to this, EVA is working with other groups at EBI to integrate relevant EVA data with other resources and present clinically relevant variation within a stable clinically relevant geneset in the LRG project.

For data submitted to EVA, value-added annotation of data is performed using a variety of methods, including Ensembl's VEP, for a variety of Gencode genesets.

With collaboration and data distribution crucial for ongoing research, all data submitted to the EVA is distributed as widely as possible: novel data is submitted to dbSNP / ClinVar as appropriate, while structural variation uses the DGVa pipeline to distribute data to dbVar and DGV. EVA has imported all dbSNP human variation, and is continuing to load other species. For EVA submitted data, EVA has both submitter defined and dbSNP annotation. EVA is an active partner in GA4GH standardising genomic data access over the Internet, with an operational Beacon and a fully compliant variation API, and is working closely with the CTTV drug target validation project hosted at EMBL-EBI to supply all clinically relevant data.

Users can mine the data using the filters on the website to construct both study-centric and global queries, filtering on any combination of species, methodology, variant type, phenotype, consequence or statistic. Results from these queries can be downloaded in a variety of formats. Additionally, EVA provides a comprehensive RESTful web- service, to allow online programmatic access to EVA data, and hence the integration of these data with that of other resources, such as Ensembl and Uniprot. All EVA software is released as open-source. Please contact eva-helpesk@ebi.ac.uk for data submissions and other queries.

SVVIZ: A READ VISUALIZER FOR STRUCTURAL VARIANTS

Noah Spies[1,2,3], Justin M Zook[3], Marc Salit[3], Arend Sidow[1,2]

[1]Stanford University, Dept of Genetics, Stanford, CA, [2]Stanford University, Dept of Pathology, Stanford, CA, [3]National Institute of Standards and Technology, Genome Scale Measurements Group, Stanford, CA

We present svviz, a sequencing read visualizer for structural variants that sorts and displays only reads relevant to a candidate SV. This is accomplished by searching input bam(s) for potentially relevant reads, realigning them against the inferred sequence of the putative variant allele as well as the reference allele, and identifying reads that match one allele better than the other. Reads are assigned to the proper allele based on alignment score, read pair orientation and insert size. Separate views of the two alleles are then displayed in a scrollable web browser view, enabling a more intuitive visualization of each allele, rather than a single-reference, genome-based view common to most current read browsers. The web view facilitates examining the evidence for or against a putative variant, estimating zygosity, visualizing affected genomic annotations, and manual refinement of breakpoints. An optional command-line-only batch validation interface allows summary statistics and graphics to be exported directly to standard graphics file formats. svviz is open source and freely available from github, and requires as input only structural variant coordinates (called using any other software package), reads in bam format, and a reference genome. Reads from any high-throughput sequencing platform are supported, including illumina short-read, mate-pair, Moleculo (assembled), Pacific Biosciences, and Oxford Nanopore.

SINGLE TUBE, WHOLE GENOME PHASING USING BEAD-BASED INDEX PARTITIONING.

Frank J Steemers[1], Fan Zhang[1], Lena Christiansen[1], Mostafa Ronaghi[1], Ros Jackson[2], Natalie Morrell [2], Niall Gormley[2], Andrew Adey[4], Jay Shendure[3], Kevin L Gunderson[1]

[1]Illumina, Advanced Research Department, San Diego, CA, [2]Illumina, Technology Development Department, Little Chesterford, United Kingdom, [3]University of Washington, Department of Genome Sciences, Seattle, WA, [4]Oregon Health & Science University, Department of molecular and medical genetics, Portland, OR

We recently introduced an unprecedented fast (3 hours) and simple 3-step (transposition-dilution-PCR) sequencing library preparation method, CPT-seq (Contiquity-Preserving Transposition), for accurate and complete whole genome phasing and *de novo* assembly [1,2]. CPT-Seq effectively generates accurate long-strobed reads that are on average ~30-50 kb in length with ~10% coverage. The method relies on the transposition of adapters and index sequences into long DNA molecules at high frequency while preserving the contiguity information of DNA. The long indexed DNA contiguity complexes are partitioned into physical wells to ensure parental copy separation and subsequent extraction of phasing and assembly information. We have demonstrated the feasibility of CPT-Seq using a two-tier combinatorial indexing scheme at both the transposition- and PCR level that effectively generates 9216 "virtual" compartments with only 192 physical wells. As such, we have significantly improved the dilution haplotyping workflow reducing both the required physical wells and number of individual indexed library preparations of fractions of the genome.
In current work, we show for the first time that index partitioning of parental DNA copies can be performed in a single tube without the need of many physical partitions. The key concept of the approach is based on the macroscopic transfer of clonal indices attached to a single bead to a long DNA molecule. As DNA molecules diffuse and bind to the immobilized transposomes on a single bead, indices transfer through tagmentation to the DNA molecule. During this process individual libraries from a single DNA molecule are kept in proximity to the bead. Since intra-tagmentation reactions are significantly faster than intermolecular reactions between beads, individual DNA molecules are clonally labeled, while simultaneously labeling many long DNA molecules in the population with different i.e., 96-10,000 clonal indexed beads in a single tube.
As a proof-of-concept, we apply this method to phase over 95% of heterozygous variants from a HapMap sample into long, accurate haplotype blocks. Additionally, we demonstrate utility of CPT-Seq in assembly applications, greatly improving scaffold contig lengths [2] and ability to detect gene-fusions and re-arrangements. These technologies provide a rapid, scalable and highly automatable route towards accurate haplotype-resolved sequencing of the genome.

[1] Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat. Genet. 46, 1343–1349 (2014).
[2] Adey, A. et al. *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. Genome Res. 24, 2041 (2014).

NON-DIPLOID INDEL DISCOVERY VIA *DE NOVO* ASSEMBLY.

Nicholas Stoler[1], Boris Rebolledo-Jaramillo[1], Marcia Shu-Wei Su[2], Kateryna D Makova[2], Anton Nekrutenko[1]

[1]Penn State University, Department of Biochemistry and Molecular Biology, State College, PA, [2]Penn State University, Department of Biology, State College, PA

Extremely high depth of coverage can now be achieved for a moderate cost. This allows identification of very low frequency variants in resequencing studies dealing with complex non-diploid mixtures represented by viral, bacterial, and organellar genomes, as well as genetically abnormal samples such as genomic DNA isolated from tumors. The major obstacle in scoring low frequency variant alleles is high sequencing error rate exhibited by NGS technologies. While several approaches have been developed for the detection of low frequency substitutions, identification of indels remains extremely challenging particularly in low complexity regions. Here we implemented a new indel discovery method using *de novo* assembly. In our pipeline, each sample is assembled individually to produce a custom reference to serve as the mapping target. The original reads are then mapped to this high-quality assembly, and carefully designed quality filters are applied to separate the final indels from sequencing artifacts. To validate our approach we recovered all detected indels from a recently published Ebola resequencing dataset. Next, we applied this pipeline to mitochondrial sequence data from a sample population of 39 healthy mother-child pairs for which we were able to identify 10 common intra-sample indels, or heteroplasmies, with 54 of the individuals possessing at least one indel. The pipeline is available in Galaxy (usegalaxy.org) and can be used for processing of large multi-sample re-sequencing datasets.

A STATISTICAL MODEL FOR SIGNAL DETECTION AND BIAS CORRECTION IN CHIP-SEQ DATA.

Alexander Engelhardt*[1], Georg Stricker*[2]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany, [2]Gene Center Munich, Ludwig Maximilians University, Munich, Germany

*Authors contributed equally

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) is a widely used method for studying interactions between proteins and DNA to better understand processes such as gene expression. ChIP-Seq data, however, has certain biases: Due to differences in factors like chromatin accessibility, mappability, and GC content, control experiments are performed that are used to correct for these biases. We developed a method based on Generalized Additive Models (GAMs), that smoothly models the coverage tracks as piecewise polynomials and is able to extract the signal component from the data. The framework allows for flexible inclusion of additional covariates
such as GC content, and elegantly handles multiple proteins and replicates. Furthermore, using piecewise polynomials for the smoothed coverage tracks leads to a simple way to find peaks based on derivatives. Our method also yields corrected coverage tracks with confidence bands for further analyses.

Benchmarks against existing methods on yeast data look promising. Furthermore the higher fraction of correct peaks out of all peaks (precision) suggests that correctly fit smooth functions are able to detect the summit of a peak more
accurate than moving-average-based methods.

FUNCTIONAL GENETIC VARIANTS IN THE CHROMOGRANIN A GENE PROMOTER GOVERN PLASMA PROTEIN LEVELS BY DIFFERENTIAL TRANSCRIPTION REGULATION

Lakshmi Subramanian[1], Prasanna K R Allu[1], Bhavani S Sahu[1], Abrar A Khan[1], Malapaka Kiranmayi[1], Ajit S Mullasari[2], Nitish R Mahapatra[1]

[1]Indian Institute of Technology Madras, Department of Biotechnology, Chennai, India, [2]Madras Medical Mission, Institute of Cardiovascular Diseases, Chennai, India

Chromogranin A (CHGA), a protein ubiquitously expressed in secretory cells of the neuroendocrine system, plays an important role in the biogenesis of catecholamine secretory vesicles. CHGA is co-stored and co-secreted with catecholamines and also acts as a prohormone giving rise to several bioactive peptides with potent cardiovascular functions. Rodent models of hypertension displayed over-expression of CHGA. Elevated levels of plasma CHGA have also been observed in human essential hypertension and cardiovascular disease states. To understand the mechanistic basis behind the elevated protein levels and its plausible association with genetic variants in the regulatory regions of the *CHGA* gene, re-sequencing of 1.2kb promoter region of *CHGA* was carried out in hypertensive and normotensive individuals in an Indian population (n=594). Nine single nucleotide polymorphisms (SNPs) in the *CHGA* proximal promoter were identified: G-1106A, A-1018T, T-1014C, T-988G, G-513A, G-462A, T-415C, C-89A and C-57T. While the G-513A SNP was a novel one in the present study, the other 8 SNPs were detected previously in other ethnic populations. While the SNPs at -1014, -988, -462 and -89 bp positions were shown to be in linkage disequilibrium (LD) in earlier studies, this study identified an additional strong LD between the A-1018T and C-57T SNPs. Haplotype analysis across the 1.2kb promoter region of *CHGA* yielded five major haplotypes. Promoter haplotype-reporter constructs displayed differential activities in neuronal cell lines; specifically, the promoter haplotype containing variant alleles at -1018, -415 and -57bp positions showed the highest activity. To assess the contribution of these SNPs to the increased promoter activity, transcription factor binding sites across these SNPs were predicted computationally. Consistent with computational predictions, the transcription factor c-Rel activated the haplotype harboring the variant alleles (-57T and -1018T)whereas the haplotype containing major alleles did not show significant promoter activation. Further, the human plasma CHGA levels corroborated with the promoter activity profile of the haplotypes establishing the role of the functional promoter polymorphisms *in vivo* in an Indian population.

DE NOVO METAGENOME ASSEMBLY USING PACBIO LONG READS

Yoshihiko Suzuki, Junko Taniguchi, Jun Yoshimura, Kenshiro Oshima, Masahira Hattori, Shinichi Morishita

The University of Tokyo, Department of Computational Biology, Kashiwa, Japan

Whole genome sequencing of metagenomic samples is expected to be able to elucidate more information on the entire genomes of individual samples than amplification sequencing of specific genes such as 16S rDNA. Metagenome assembly, however, is thought to be much more difficult than single genome assembly because similar genomic regions shared among different species in addition to repeats in a single species make it hard to output unambiguous assembly. To cope with this problem, we took an approach of using PacBio's long reads of length more than 10 Kbp that can contain long repeats (for example, multi-copy rDNA operon of length about from 5 Kbp to 7 Kbp) as substrings, and developed a new method of filtering erroneous overlaps of long reads that stem from repeats in a single species or from similar regions in distinct species, so as to avoid a major source of problems in metagenome assembly. In this study, we *de novo* assembled human gut microbial data with no binning into groups of reads from individual species but filtering overlaps that had extremely different coverage between both ends and obtained about 30 contigs of more than 500 Kbp in size that were much longer than contigs by using only Illumina's short reads. Moreover, the new filtering method reduced the possibility of mis-assemblies. These results show that assembly of metagenomic data is dramatically improved by using PacBio long reads instead of short reads. Furthermore, without preforming read binning, we can assemble long reads from metagenomes in line with single genome assembly if we can properly trim putative overlaps in similar regions between different species or strains.

CHARACTERIZING DNA METHYLATION OF LIVING LINE/L1 TRANSPOSONS IN HUMAN GENOMES USING LONG SMRT READS

Yuta Suzuki[1], Shoji Tsuji[2], Shinichi Morishita[1]

[1]The University of Tokyo, Department of Computational Biology, Graduate School of Frontier Sciences, Kashiwa, Chiba, Japan, [2]The University of Tokyo, Neuroscience, Faculty of Medicine, Graduate School of Medicine, Bunkyo, Tokyo, Japan

Long Interspersed Elements (LINEs) make up more than 20% of the human genomes, and its younger subfamily, LINE/L1 transposon, is known to be still active in various tissue types including germ-line, brain, and cancer cells. While human genomes contain the hundreds of thousands of L1 elements, only a handful of copies (<150 copies) have been found to be autonomous, that is, retaining two ORFs intact and possibly mobile. Though the activity of L1 transposon was reported to be regulated by repressive epigenetic modifications like CpG methylation, these living transposons share highly homologous DNA sequence over its full-length of 6 kb, making it extremely difficult to determine the methylation states of individual occurrences of the living transposons because bisulfite-treated short reads hardly map to their originating positions unambiguously.
We have developped a computational method to detect CpG-hypomethylated regions with high accuracy using long SMRT (Single-Molecule Real Time) reads with polymerase kinetics data obtained from PacBio RSII system. Long reads, with its N50 longer than 15 kbp, were able to address these living transposons separately. Based on the latest data from ongoing PacBio sequencing of several Japanese genomes, we will present a comprehensive methylation analysis on these living transposons on the reference genome, and will discuss the relation between recent transposon activity and DNA methylation. Since long reads can detect SVs (structural variants) like novel insertions of L1 elements that do not appear in the reference genome, our method would be able to explore the possible personal variation of SV events and its methylation status simultaneously.

NANOPORE SEQUENCING FOR GENOTYPING PATHOGENS OF TROPICAL DISEASES

Yutaka Suzuki[1], Arthur E Mongan[2], Josef Tuda[2], Junya Yamagishi[3]

[1]University of Tokyo, Department of Medical Genome Sciences, Kashiwa, Japan, [2]Sam Rutulalgi University, Department of Medicine, Manado, Indonesia, [3]Hokkaido University, Center for Protozoan Diseases, Sapporo, Japan

Nanopore sequencer, MinION, has enabled sequencing analysis without pre-installation of expensive conventional sequencers or pre-requisite of specific skills in biological experiments. Even electric supply is not always necessary, by connecting MinION to a laptop PC. These features of MinION have opened the opportunity to enable precise genotyping of pathogens in tropical diseases in a developing country even in its filed areas. In this study, we attempted genotyping Dengue viruses regarding their serotypes (types 1-4). We directly used serum samples of ten Indonesian Dengue patients, from which viral genomes were directly amplified by the reverse-transcription-LAMP method in an isothermal reaction condition. We directly used the amplified templates for MinION sequencing allocating one flow cell per sample. We found, although the overall sequencing quality was low (70% sequence identify to the reference genome and the quality value of QV= 5 on average, when evaluated based on the LAST alignment), thereby obtained sequence data could discriminate different serotypes of the viruses, whose genome sequences were diverged with the sequence similarity of 70%, with the overall accuracy of >98%. To further examine whether MinION sequencing can be also applied for detecting SNVs, we conducted genotyping of presumed drug resistance-causing SNVs in malaria parasites, Plasmodium falciparum. We similarly subjected ten PCR amplicon-mixes, covering these SNVs which were generated from ten Indonesian malaria patients to the MinION sequencing. Again, in spite that the sequence alignments generated by LAST showed that the average sequence identity was 65%, we found that the mutations at a particular position could be called by the accuracy of >85%, when all the reads covering the corresponding positions were collectively evaluated. Taken together, we provide the first simple experimental and analytical MinION sequencing procedure and applied them for effective genotype for pathogens of tropical diseases in filed.

ABERRANT PRE-mRNA SPLICING DUE TO MUTATIONS IN *RNU4ATAC*, A MINOR SPLICEOSOMAL snRNA, RESULTS IN SEVERE DEVELOPMENTAL PHENOTYPES IN NEW MOUSE MODELS

David E Symer[1,2], Dandan He*[1,2], Jingfeng Li*[2], Katherine A Yates[1], Keiko Akagi[1], Christopher J Hlynialuk[3], Zhengqiu Zhou[2], Xiaomei Meng[1], Yanqiang Wang[1], Chelsea A Moherman[1], Tanvi V Joshi[1], Huiling He[1], Albert de la Chapelle[1], Brad N Bolon[4], Richard A Padgett[5]

[1]OSU, MVIMG, Columbus, OH, [2]OSU, Internal Medicine, Columbus, OH, [3]Univ Minn, Neurology, Minneapolis, MN, [4]OSU, Veterinary Pathology, Columbus, OH, [5]Cleveland Clinic, Molecular Genetics, Cleveland, OH

Human developmental diseases often can be faithfully modeled by mutations in orthologous mouse genes, but phenotypic differences also have been described. We recently identified the molecular basis for a severe human developmental disorder, microcephalic osteodysplastic primordial dwarfism, type I (MOPD I) (1). It results from biallelic mutations in a small nuclear RNA (snRNA) gene, *RNU4ATAC*, which plays a crucial role in minor spliceosomal function. Here we show that orthologous mutations in the mouse gene, *RNU4atac*, also result in developmental abnormalities. We engineered an allelic series of mutations including the nullizygous allele to study their impacts on pre-mRNA splicing and mouse development. In numerous instances we observed similarities between the mouse and human phenotypes. Certain combinations of biallelic mutations resulted in growth defects and abnormal bone development. In addition, mutant mice displayed tissue-specific defects including abnormal thymus, spleen and lymph node architecture, hepatic vacuolization, markedly reduced brown fat, ventricular septal defects and hypomorphic gonads, which have not been described in human MOPD I. Conversely, in contrast with human MOPD I, we observed no defects in brain architecture in the mouse models. We confirmed that overall splicing of minor class introns is abnormal in the mutant mice. These results indicate that pre-mRNA splicing plays a critical role in both human and mouse development. We speculate that particular differences in tissue-specific phenotypes may reflect subtle differences in gene-specific splicing defects between the two species. Our new mouse models have facilitated a detailed investigation of the highly conserved downstream target genes whose aberrant pre-mRNA splicing drives these and other species- and tissue-specific developmental phenotypes.

(1) He, H. et al., Mutations in U4atac snRNA cause microcephalic osteodysplastic primordial dwarfism type I by disrupting minor spliceosome function. Science 332: 238-240, 2011.

APOBEC3 MUTATIONAL SIGNATURES ARE ENRICHED IN HUMAN PAPILLOMAVIRUS-POSITIVE ORAL CANCERS

David E Symer[1,2], Keiko Akagi[1], Kevin R Coombes[4], Jingfeng Li[2], Weihong Xiao[2], Tatevik R Broutian[2], Bo Jiang[2], Robert Pickard[2], Amit Agrawal[3], Maura L Gillison[2]

[1]OSU, MVIMG, Columbus, OH, [2]OSU, Internal Medicine, Columbus, OH, [3]OSU, Otolaryngology, Columbus, OH, [4]OSU, Biomedical Informatics, Columbus, OH

Human papillomavirus is a major contributor to the pathogenesis of an increasing number of oral cancers. Although the viral oncoproteins E6 and E7 are necessary for cancer development, they are insufficient, and the secondary genetic events that promote cancer progression are largely unknown. To characterize somatic mutational processes in oral cancers with and without HPV, we conducted whole genome sequencing (WGS) on a panel of 59 pairs of tumors and matched normal samples. Here we describe large differences in the patterns of mutations observed in 34 HPV-positive oral cancers when compared with 25 HPV-negative cancers. Although overall somatic mutation rates were comparable in the HPV+ vs. HPV- cancers, APOBEC-mediated C>T and C>G mutations were disproportionately increased in HPV-positive samples, both genome-wide in the host DNA as well as in the viral genomes themselves. The APOBEC signature accounted for approximately 55% of all MutSig mutations in the HPV+ but only 8% of the HPV- cancers. By contrast, an increased proportion of APOBEC-associated mutations associated with structural variation was seen in HPV- cancers. RNA-Seq data revealed significant increases in expression of immune defense genes characteristically expressed in leukocytes. Long-read sequencing confirmed that the APOBEC pattern occurred in clusters of single-stranded mutations. These findings confirm that APOBEC activity plays a crucial role in creating secondary genetic mutations predominantly in HPV+ oral cancers.

MULTIPLE LINES OF TRANSGENIC MICE SHED NEW LIGHT ON THE MOLECULAR MECHANISMS UNDERLYING THE CALLIPYGE PHENOMENON.

Haruko Takeda[1], Dimitri Pirottin[1], Xuewen Xu[1], Fabien Ectors[2], Huijun Cheng[1], Tracy Hadfield[3], Noelle Cockett[3], Carole Charlier[1], Michel Georges[1]

[1]GIGA Research Center and Faculty of Veterinary Medicine, University of Liège, Unit of Animal Genomics, Liège, Belgium, [2]FARAH and GIGA Research Center, University of Liège, Transgenic platform, Liège, Belgium, [3]Utah State University, Department of Animal, Dairy and Veterinary sciences, Logan, UT

The callipyge phenotype is a muscular hypertrophy of sheep that is characterized by polar overdominance: only heterozygous animals inheriting the *CLPG* mutation from their sire express the phenotype. The *CLPG* mutation inactivates a silencer element that down-regulates genes from the imprinted *DLK1-GTL2* domain in *cis* in postnatal skeletal muscle. Consequently, $+^{Mat}/CLPG^{Pat}$ animals overexpress the paternally expressed DLK1 and PEG11 in muscle at the mRNA and protein levels, and this is thought to cause the phenotype. In *CLPG*/*CLPG* animals, *DLK1* and *PEG11* are presumably *trans*-inhibited by maternally expressed non-coding RNAs, accounting for their wild-type phenotype. Transgenic mice expressing ovine DLK1 in skeletal muscle have a muscular hypertrophy supporting a role for DLK1 in determining the phenotype. In *CLPG*/*CLPG* animals, *PEG11* is sliced by miRNAs processed from *anti-PEG11* confirming the non-coding-RNA-mediated *trans*-inhibition.

To further characterize the phenomenology, we generated additional transgenic mice. We knocked the *CLPG* mutation in the mouse genome. We show that the *cis*-effects of the mutation observed in sheep are robustly recapitulated upon maternal transmission (i.e. on *Gtl2*, *anti-Peg11*, *Meg8* and *Mirg*) but much less so upon paternal transmission (i.e. on *Peg11* and *Dlk1*), that the *trans*-effect is not observed for *Dlk1*, and that $+^{Mat}/CLPG^{Pat}$ mice are phenotypically normal. We then generated ovine PEG11 transgenic mice. These expressed a muscular hypertrophy of similar magnitude as the DLK1 transgenic mice. These imply a possible synergistic effect in double DLK1/PEG11 transgenic mice.

These results confirm the causality of the *CLPG* mutation, and suggest that ectopic expression of PEG11 contributes to the muscular hypertrophy of callipyge sheep. That $+^{Mat}/CLPG^{Pat}$ mice are phenotypically wild-type would be attributed to the lower expression of both *Dlk1* and *Peg11* in these mice when compared to callipyge sheep.

UNCOVERING NOVEL microRNAs IN DEVELOPING MAIZE KERNELS

Oliver H Tam[1], Katherine Petsch[2], Molly Hammell[1], Marja Timmermans[2]

[1]Cold Spring Harbor Laboratory, Genomics, Cold Spring Harbor, NY,
[2]Cold Spring Harbor Laboratory, Plant Genetics, Cold Spring Harbor, NY

Small RNAs, such as microRNAs, are important regulators of gene expression that act in a homology-dependent manner to guide transcriptional and post-transcriptional silencing mechanisms. However, in maize, knowledge of small RNA pathways and the targets that they regulate is greatly lacking. Limited available data indicates substantial divergence of the small RNA pathways between the model plant, Arabidopsis, and maize, necessitating a direct interrogation of these small RNA populations. To further elucidate the unique roles of small RNAs in maize, we sequenced embryo and endosperm tissues (12 days post pollination) from wild type and a dicer-like 1 (dcl1) mutant in the B73 inbred line. Using two independent approaches, we uncovered potentially novel miRNAs originating from unique and repetitive regions of the genome. These candidates share many characteristics with previously annotated maize microRNAs, and show significant depletion in the dcl1 mutant. Further investigation of these candidates to identify their regulatory targets, and to assess their conservation among other plant species is ongoing. The characterization of small RNA populations in maize is critical to our understanding of their roles in growth and development. Deep sequencing small RNAs and their targets will provide valuable insights into the molecular and functional diversity in maize, and aid in elucidating their evolutionary divergence across other plant species.

THE GENETIC ARCHITECTURE OF METABOLIC RESPONSE IN SKELETAL MUSCLE EXPRESSION

D Leland Taylor[1,2], Francesco P Casale[2], Stephen CJ Parker[3], Jeroen R Huyghe[4], Michael R Erdos[1], Heikki Koistinen[5], Ryan Welch[4], Heather Stringham[4], Laura J Scott[4], Brooke Wolford[1], The FUSION Study[4], Richard M Watanabe[6], Karen Mohlke[7], Jaakko Tuomilehto[5], Michael Boehnke[4], Oliver Stegle[2], Ewan Birney[2], Francis S Collins[1]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [2]EMBL, European Bioinformatics Institute, Hinxton, United Kingdom, [3]University of Michigan, Depts. of Computational Medicine & Bioinformatics and Human Genetics, Ann Arbor, MI, [4]University of Michigan, Dept. of Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, [5]National Institute of Health and Welfare, Metabolism Group, Helsinki, Finland, [6]University of Southern California Keck School of Medicine, Dept. of Preventive Medicine, Los Angeles, CA, [7]University of North Carolina, Dept. of Genetics, Chapel Hill, NC

Type 2 diabetes (T2D) is a complex disease that results from a combination of genetic and environmental factors, including physiological responses to lifestyle choices. T2D is characterized by dysfunction in insulin secretion and insulin resistance. The genetics of T2D is well studied, with >80 known loci from genome wide association studies. However it is unclear how the combination of genetics and physiological factors influence the etiology of T2D.

As part of the Finland United States Investigation of NIDDM Genetics (FUSION) Study, we analyzed 278 individuals with metabolic measures (glucose, insulin, and other blood metabolites), genotypes, and muscle biopsies (mRNA-seq). A considerable challenge in RNA-seq studies is the presence of technical (e.g. batch effects) and biological (e.g. effects of physiological state) variation. Furthermore, because technical and biological variation can have widespread effects across many genes, methods that account for these confounders struggle to distinguish technical from biological signals. We developed a novel technique that iteratively fits random effect covariance from gene expression data, while conditioning on biologically meaningful covariates.

We used this method to remove confounders but preserve metabolic signals in gene expression, and find it greatly improves power to identify genes sensitive to metabolic measurements – most notably insulin. We then characterized genetic effects of two types – genetic control of gene expression (eQTL), and a genotype dependent relationship between gene expression and metabolic measurements. In heterozygote individuals, we also explored the impact of these genetic responses using allele specific expression. Together, these methods reveal new regulatory loci and mechanistic insights into the response of skeletal muscle to T2D progression.

iCLiKVAL: INTERACTIVE COMMUNITY RESOURCE FOR THE MANUAL CURATION OF ALL SCIENTIFIC LITERATURE THROUGH THE POWER OF CROWDSOURCING

Todd D Taylor, Naveen Kumar

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

There are currently over 24 million citations to various forms of scientific literature in PubMed. Searching this vast resource does not always give desirable or complete results due to a number of factors such as: missing abstracts, unavailability of full-article text, non-English articles, lack of keywords, etc. Ideally, every citation should include a complete set of keywords or terms that describe the original article in enough detail that searches, using natural language, return more relevant results; however, this would require countless hours of manual curation. Our objective is to make manual curation 'fun' and social and self-correcting, thus enriching resources such as PubMed so that users are able to extract more valuable and relevant results.

We have developed a web-based open-access tool for manual curation of PubMed articles, and other media types, using a crowdsourcing approach which we believe the community will enjoy using. While we encourage the use of ontology terms and support them as auto-suggest keyword terms, we do not restrict users to these terms because we do not want to impose, within reason, any limitations on the types of annotation that one may provide. Non-English annotation is also supported. Through this ‘non-restrictive’ approach, we hope that communities of researchers, no matter where their location or what their language, will take advantage of this tool to work together on the manual curation of any type of literature related to their projects.

We have constructed a cross-browser and platform-independent application using the latest web technologies and a NoSQL database. Users perform searches to identify articles of interest, mark articles for later review or review them immediately or add them to a review request queue, load PDFs into the viewer, select annotations (values) within the text, and add appropriate keywords (keys) and relationship terms, if applicable. Article-specific comments can be made, key-value pairs can be edited and rated, live chats between users working on the same article can be held, etc. Users can even add annotations via Twitter.

As more annotations accumulate in the database, the more our semantic search feature will improve and the more relevant the results. Users will then be able to precisely filter the results. Rather than compete with other already-established curation projects, we wish to work with them to incorporate their valuable data and in return, via our REST API, make the iCLiKVAL annotations easily accessible to the entire research community. We hope this will become the default resource for community-based curation of all online scientific literature.

BALANCING SELECTION IN THE GREAT APES

João C Teixeira, Aida M Andrés

Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany

Balancing selection maintains advantageous genetic diversity in species and populations. Because balancing selection can be (unlike positive selection) long lasting, it is theoretically possible for the same locus to be continuously under balancing selection from the common ancestor of two species until their present-day populations. In fact, classical examples of balancing selection such as the major-histocompatibility complex (MHC) in vertebrates and the ABO locus in primates are examples of this. The extent to which balancing selection is conserved among species is unknown, but previous studies have detected a few trans-species polymorphisms maintained by long-standing balancing selection and shared between humans and closely related species. These polymorphisms unveil remarkable, extreme signatures of balancing selection and are expected to represent a minor fraction of the loci evolving under similar selective pressures in different species.

Our study aims to determine to what extent long-term balancing selection is shared among species, outside of the extreme examples of trans-species polymorphisms. We developed a statistic that has high power to detect the departures from the distribution of neutral allele frequencies (based on both simulated and empirical data) that we expect under a wide range of balancing selection scenarios. We analyzed genome-wide population samples from 11 great ape subspecies and identified candidate targets of long-term balancing selection in each population. We will present evidence for shared signatures among the different subspecies, including biologically relevant categories for which the selective pressures are shared across subspecies. Not surprisingly, we also uncovered examples of loci that show differing signatures of natural selection in different subspecies. Finally, we will also discuss the contribution of different factors (such as genomic sequence, recent common ancestry, or environment) to the observed patterns of shared loci evolving under balancing selection in the different populations.

COMPREHENSIVE TRANSCRIPTOME ANALYSIS USING SYNTHETIC LONG READ SEQUENCING REVEALS MOLECULAR CO-ASSOCIATION AND CONSERVATION OF DISTANT SPLICING EVENTS.

Hagen U Tilgner[1], Fereshteh Jahaniani[1], Tim Blauwkamp[2], Ali Moshrefi[2], Erich Jaeger[2], Feng Chen[2], Itamar Harel[1], Carlos Bustamante[1], Morten Rasmussen[1], Michael Snyder[1]

[1]Stanford University, Genetics, Stanford, CA, [2]Illumina Inc., Hayward, CA

Alternative splicing (AS) shapes mammalian transcriptomes and proteomes, through the regulated inclusion of alternative exons, which often harbor protein-coding sequence; and RNA molecules can harbor multiple distant AS-events. Comprehensive transcriptome analysis, including full-length isoform sequencing and exon co-association analysis has been difficult so far, because long-read-RNA-sequencing datasets have been lacking sufficient depth. Based on the Moleculo-protocol, we introduce a new and facile way of RNA-sequencing, synthetic-long-read-RNA-sequencing (SLR-RNA-Seq), in which long RNA-sequences are deduced from pools of cDNAs and in which a 1-to-1 relationship between the original RNA molecules and the resulting cDNA-reads is preserved for most genes. We apply this method to the analysis of the human brain and mouse brain transcriptome. We show that we can produce much larger quantities of long reads (average 1,900bp) than with other long-read-sequencing protocols. We demonstrate that the RNA-sequences we obtain are mostly close to full length and contain few insertion and deletion errors. Conservatively counting, we report many novel isoforms (human brain: ~13,800 affected genes, 14.5% of all molecules representing a novel isoform; mouse brain ~8,600 genes, 18% of all molecules). Analyzing pairs of alternative exons at a false discovery rate of 0.01, we observe 71 (out of a universe of 36,000 tested exon pairs) human intragenic molecularly associated pairs (IMAPs) and intragenic molecularly opposed pairs (IMOPs) of distant exons: Pairs of alternative exons that are separated by one or more constitutive exons in the mature RNA and whose inclusion levels are statistically dependent on each other. Adding other datasets and relaxing the false discovery rate to 0.15 yields 165 pairs with dependent splicing. Most exon pairs affect the encoded protein at distant sites. For the mouse brain, we find 16 non-randomly linked exon pairs (with less sequencing depth) and show that for nine of these mouse exon pairs their coordinated inclusion pattern is conserved in human. Our results indicate conserved mechanisms that can produce distant but phased features on transcript and protein isoforms and will thus impact both transcriptome and proteome research.

COMPARATIVE ANALYSIS OF THE Y CHROMOSOME GENOMES OF GREATER APES

Marta Tomaszkiewicz[1], Samarth Rangavittal[1], Monika Michalovová[1], Rebeca Campos Sanchez[1], Howard W Fescemyer[1], Oliver Ryder[2], Malcolm Ferguson-Smith[3], Rayan Chikhi[4], Paul Medvedev[5,6,7], Kateryna D Makova[1]

[1]Pennsylvania State University, Department of Biology, University Park, PA, [2]San Diego Zoo, Institute for Conservation Research, San Diego, CA, [3]University of Cambridge, Department of Veterinary Medicine, Cambridge, United Kingdom, [4]University of Lille 1, CNRS, Lille, France, [5]Pennsylvania State University, Department of Computer Science and Engineering, University Park, PA, [6]Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA, [7]Pennsylvania State University, Genome Sciences Institute of the Huck, University Park, PA

The female genomes of four greater apes — human, chimpanzee, gorilla, and orangutan — have diverged from each other by less than 3%. Yet, the human and chimpanzee male-specific regions of the Y chromosomes (MSYs), the only two sequenced greater ape MSYs, are highly divergent with more than 30% of non-homologous sequences. Moreover, we have previously demonstrated that Y-chromosomal X-degenerate genes are better conserved between human and gorilla than between human and chimpanzee. In this study, we are focusing on the comparative analysis of the male-specific regions (MSYs) of the great ape Y chromosomes.

The gorilla and orangutan Y chromosome sequences have been the missing pieces for a thorough investigation of the four greater ape Y chromosomes. Here, we sequence the whole genome amplified flow-sorted gorilla and orangutan Y chromosome DNA with both short-read (Illumina) and long-read (PacBio) technologies. Combining existing tools with new methods for extracting and assembling Y-chromosome specific sequences established in our lab, we were able to generate the first draft assemblies of gorilla and orangutan Y chromosomes. As a result of our analysis, we estimate the divergence level, gene content, and detect rearrangements among greater ape Y chromosomes. Additionally, we study the polymorphism of the Y chromosome in greater ape populations by analyzing male-specific microsatellites and copy number variations of ampliconic genes. These insights are important for conservation genetics.

PREDICTING CENTROMERIC HIGHER ORDER REPEATS IN HUMAN GENOMES WITH PACBIO LONG READS

Shingo Tomioka, Shinichi Morishita

The University of Tokyo, Department of Computational Biology, Graduate School of Frontier Sciences, Tokyo, Japan

Centromere is the locus responsible for proper chromosome segregation during meiosis and mitosis. DNA sequences in human centromeric regions are constituted with approximately 171 bp tandem repeats referred to as alpha satellites. In the majority of centromeric regions, a multimer of alpha satellites organizes a unit which itself iterates tandemly and makes several hundreds of highly identical copies called higher-order repeats (HORs, for short). Each chromosome has its own HORs that span several hundred kb to several Mb. Previous computational methods of predicting HORs are mainly based on 500 bp WGS Sanger reads that can contain less than three monomers, so they need clustering of identical monomers from different reads which may cause confusion of different HORs having similar monomers. Also these methods require certain amount of HOR copies for their detection so minor variants of HORs may still remain undiscovered. PacBio long reads can contain several copies of HOR sequence unit, offering the possibility of further description of HOR sequences and structures. We developed a method for predicting HOR structures directly from each long read. This method enables comprehensive characterization of centromeric sequences in individual genomes and comparative analysis among individuals, which would help better understanding of the centromeres in the human genome.

NANOCONFINEMENT SYSTEMS FOR GENOME ANALYSIS

Gene Tsvid[1], Kristy Kounovsky-Shafer[1,2], Juan Hernandez-Ortiz[3], Konstantinos Potamousis[1], Theo Odijk[4], Juan de Pablo[5], David C Schwartz[1]

[1]University of Wisconsin-Madison, Chemistry, Laboratory of Genetics, and the UW-Biotechnology Center, Madison, WI, [2]University of Nebraska-Kearney, Chemistry, Kearney, NE, [3]Universidad Nacional de Colombia, Materiales, Sede Medellín, Colombia, [4]University of Leiden, Institute for Theoretical Physics, Leiden, Netherlands, [5]University of Chicago, Institute for Molecular Engineering, Chicago, IL

High throughput sequencing has transformed biology and medicine, but at a cost of "completeness". Short read lengths attenuate coverage within difficult portions of a genome and incompletely characterize structural variants as copy number changes. Towards the goal of retaining and analyzing long range sequence information and completeness our group develops systems that manipulate and analyze very large DNA molecules. The main components of our systems are technologies for high-throughput presentation of very large DNA molecules for subsequent analysis. Here we present a DNA presentation approach based on nanoslits capable of achieving DNA stretching above 90%. This high degree of stretch is enabled by DNA confinement within nanoslits to the dimension smaller than the persistence length of the molecule, increase of which is aided in turn by low ionic strength of the buffer. A molecular model we have developed describes the conformation and dynamics of the DNA molecules within the nanoslits; a Langevin description of the polymer dynamics is adopted in which hydrodynamic effects are included through a Green's function formalism. Our simulations reveal that a delicate balance between electrostatic and hydrodynamic interactions is responsible for the observed molecular conformations. We demonstrate and further confirm that the "Odijk regime" does indeed start when the confinement dimensions are of the same order of magnitude as the persistence length of the molecule.

PRINCIPLES OF LONG NONCODING RNA EVOLUTION DERIVED FROM DIRECT COMPARISON OF TRANSCRIPTOMES IN 17 SPECIES

Hadas Hezroni[1], David Koppstein[2,3], Mathew G Schwartz[4], David P Bartel[2,3], Igor Ulitsky[1]

[1]Weizmann Institute of Science, Department of Biological Regulation, Rehovot, Israel, [2]Whitehead Institute for Biomedical Research, -, Cambridge, MA, [3]Massachusetts Institute of Technology, Howard Hughes Medical Institute and Department of Biology, Cambridge, MA, [4]Harvard Medical School, Department of Genetics, Cambridge, MA

Vertebrate genomes are pervasively transcribed and encode thousands of long noncoding RNAs (lncRNAs) that are dispersed throughout the genome and typically expressed at low expression levels and in a tissue-specific manner. Comparative genomics can help address a multitude of questions about lncRNAs, but it is currently impossible to predict lncRNAs from genomic sequences and it has been difficult to effectively compare them across species. Further, fast evolution of lncRNA transcription invalidates the use of genome alignments for studying their conservation.

To address these challenges, we developed a rigorous approach for transcriptome assembly, annotation and lncRNA discovery, collected RNA-seq and 3P-seq data from 17 species and identified thousands of novel lncRNA genes. Then, using stringent methodology for identifying sequence-conserved and syntenic lncRNAs, we explored functional features of lncRNAs that have been conserved during vertebrate evolution. We find that lncRNAs evolve rapidly, with >70% of lncRNAs in each species having no sequence-similar orthologs in species separated by >50 million years of parallel evolution. Fewer than 100 lncRNAs can be traced to the last common ancestor of tetrapods and teleost fish, but several hundred were likely already present in the common ancestor of birds, reptiles, and mammals. For the conserved lncRNAs, tissue specificity is conserved at levels comparable to protein-coding genes, suggesting control by conserved regulatory programs. In addition, we find that thousands of lncRNAs appear in conserved genomic positions without sequence conservation, including a particularly interesting group of lncRNAs that show sequence conservation only in mammals but have syntenic counterparts throughout vertebrates. Further, dozens of the lncRNAs that do have conserved sequences throughout vertebrates have syntenic homologs in sea urchin with similar genomic features, providing the first examples of human lncRNAs conserved beyond vertebrates.

We also studied how conserved lncRNA loci evolved and find that lncRNAs from distant species share short islands of sequence conservation, typically spanning 1–2 exons and having a significant 5' bias. Transposable elements have extensively rewired the architecture of conserved lncRNA loci, particularly in mammals, and lncRNAs show only marginal avoidance of transposon insertions in their exons. These events often altered lncRNA gene architecture, most evidently at 3' ends. Despite the rapid evolution of individual lncRNA genes, the general motif content of lncRNAs is stable between species, suggesting common underlying rules that nevertheless allow for rapid sequence evolution.

Thus, thousands of lncRNAs likely have conserved functions in mammals, and hundreds beyond mammals, yet these depend on only short patches of specific sequences and can tolerate major changes in gene architecture.

BASiCS: BAYESIAN ANALYSIS OF SINGLE-CELL SEQUENCING DATA

Catalina A Vallejos[1,2], John C Marioni[2], Sylvia Richardson[1]

[1]MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom, [2]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

Single-cell mRNA sequencing can uncover novel cell-to-cell heterogeneity in gene expression levels in seemingly homogeneous populations of cells. However, these experiments are prone to high levels of unexplained technical noise, creating new challenges for identifying genes that show genuine heterogeneous expression within the population of cells under study. BASiCS (Bayesian Analysis of Single-Cell Sequencing data) is an integrated Bayesian hierarchical model where: (i) cell-specific normalisation constants are estimated as part of the model parameters, (ii) technical variability is quantified based on spike-in genes that are artificially introduced to each analysed cells lysate and (iii) the total variability of the expression counts is decomposed into technical and biological components. BASiCS also provides an intuitive detection criterion for highly (or lowly) variable genes within the population of cells under study. This is formalised by means of tail posterior probabilities associated to high (or low) biological cell-to-cell variance contributions, quantities that can be easily interpreted by applied users. We demonstrate our method using gene expression measurements from mouse embryonic stem cells. Meaningful enrichment of gene ontology categories within genes classified as highly (or lowly) variable supports the efficacy of our approach.

ANALYSIS OF GENETIC HISTORY OF SIBERIAN AND NORTHEASTERN EUROPEAN POPULATIONS

Emily Wong[1], Andrey Khrunin[2], Larissa Nichols[2], Dmitry Pushkarev[4], Denis Khokhrin[2], Dmitry Verbenko[2], Oleg Evgrafov[5], James Knowles[2], John Novembre[2], Svetlana Limborska[2], Anton Valouev[1]

[1]University of Southern California, Division of Bioinformatics, Dept. of Preventive Medicine, Los Angeles, CA, [2]Russian Academy of Sciences, Department of Molecular Bases of Human Genetics, Institute of Molecular Genetics, Moscow, Russia, [3]University of Chicago, Human Genetics, Chicago, IL, [4]Stanford University, Computer Science, Palo Alto, CA, [5]Keck School of Medicine, Zilkha Neurogenetic Institute, University of Southern California, Department of Psychiatry and Behavioral Sciences, Los Angeles, CA

Siberia and Western Russia are home to over 40 culturally and linguistically diverse indigenous ethnic groups. Yet, genetic variation of people from this region is largely uncharacterized. We present whole-genome sequencing data from 28 individuals belonging to 14 distinct indigenous populations from Siberia and Eastern Europe as well as genotyping data from 82 individuals. We combine these datasets with additional 32 modern-day and 15 ancient human genomes to build and compare autosomal, Y-DNA and mtDNA trees. Our results provide new links between modern and ancient inhabitants of Eurasia. Eastern Siberians are related to East Asians and share 38% of ancestry with descendants of the 45,000-year-old Ust'-Ishim people, previously believed to be a dead-end population branch. Western Siberians are related to Europeans and trace 57% of their ancestry to 24,000-year-old Siberian Mal'ta people. In addition, Siberian admixtures are present in lineages represented by the 5,000-year-old hunter-gatherer excavated in Sweden, the 2,900-year-old Iron-age man excavated in Hungary, and modern-day northeastern Europeans. These results represent new evidence of ancient gene flows from Siberia into Europe.

HTLV-1/BLV ANTISENSE-RNA DEPENDENT CIS-PERTURBATION OF CANCER DRIVERS IN PRE-LEUKEMIC AND LEUKEMIC CLONES

Nicolas Rosewick[1,2], Durkin Keith[2], Artesi Maria[2], Hahaut Vincent[2], Marçais Ambroise[3], Hermine Olivier[3], Michel Georges[2], Anne Van den Broeke[1,2]

[1]Institut Jules Bordet, Université Libre de Bruxelles, Experimental Hematology, Brussels, Belgium, [2]GIGA-R, Université de Liège, Animal Genomics, Liège, Belgium, [3]Hôpital Universitaire Necker, Université René Descartes, Institut Imagine, Service d'Hématologie adultes, INSERM U1163 CNRS ERL 8254, Paris, France

It is generally assumed that the development of leukemia by HTLV-1 and its model BLV is independent of proviral integration site, which appears quasi random. Using RNA-seq to map integrations and simultaneously determine local patterns of transcription in fourty HTLV-1 dependent and fourty-seven BLV dependent primary leukemias, we upset this dogma by demonstrating the non-randomness of proviral integration sites and their preferential co-localization with known cancer drivers. We systematically observed local rewiring of the host transcriptome as a result of proviral integration despite the absence of 5' LTR-driven viral mRNA transcription from the positive strand. More specifically, we demonstrate a critical role for HTLV-1/BLV antisense-RNA dependent perturbation of upstream target genes involving the induction of virus-host fusion transcripts either alone or in combination with other virus-dependent mechanisms including promoter insertion and viral poly (A)-mediated premature transcription termination of host transcripts. We found that host genes affected by these mechanisms were significantly biased towards genes with a strong connection to cancer. Interestingly, deregulations revealed by RNA-seq could not necessarily be predicted by proviral integration site analysis. We identified antisense-affected host genes that were neither direct integration targets nor the nearest annotated gene. In addition, we found chimeric transcripts perturbing distant genes (up to 1 Mb relative to provirus) or affecting multiple host genes positioned adjacently in the same or opposite orientation. The same pattern was found for pre-leukemic as for leukemic clones, suggesting that provirus-dependent cis-perturbation of cancer driver genes underlies initial amplification of the corresponding clones, requiring additional genetic and/or epigenetic changes to develop full blown leukemia.

A FULLY AUTOMATED COMPUTATIONAL INFRASTRUCTURE FOR NGS ANALYSIS IN THE X TEN ERA

Francesco Vezzi

National Genomics Infrastructure, SciLifeLab, Stockholm, Sweden

As Next-Generation Sequencing (NGS) continues to revolutionize scientific fields from personalized medicine to population-scale genetics, one of the largest challenges facing researchers and core sequencing centers has become how to analyze the enormous quantities of data being produced: Illumina's newest sequencing platforms (HiSeq X) can produce up to two terabases of information per run which must be organized and run through various computationally-intensive steps. The practical challenges of organizing and tracking these quantities of information are formidable, and bioinformaticians are often forced to spend as much or more time wrangling and processing their data as they spend interpreting results.

The National Genomics Infrastructure ("NGI") in Sweden is host to 17 HiSeq 2500 and 10 HiSeq X (one "X Ten" system), for a theoretical peak sequencing production of more than 450 simultaneous Whole Human Genomes (30X coverage). In addition, NGI offers its users a wide range of sequencing applications (DNA-seq, RNA-seq, RAD-seq, etc.) and bioinformatic analyses (variant calling, de novo assembly, mRNA expression levels, etc.). NGI is geographically distributed across two separate sites, 100 km apart. To handle the massive quantities of data produced as well as the wide variety of applications and analyses offered, NGI has developed and put into production a cohesive set of software tools that automatically organize and process sequencing data as they come off the sequencers, producing meaningful analysis results without any human intervention. These tools are designed to work in a modular way, allowing bioinformaticians to direct the data through the bioinformatic pipelines of their choice. Selected result metrics can be tracked using the supplied database tools and the status of every process is made available via a REST-capable API and accompanying web frontend. Automatic report generation, data delivery, and a robust logging/notification system allow bioinformaticians to become involved only when required, allowing them to spend less time dealing with files and more time answering biological questions.

This software suite, dubbed the NGI Pipeline, is currently used and developed at the Science for Life Laboratory in Stockholm and Uppsala, Sweden. As a fully public and open-source project, we welcome both additional users and developers who are interested in automating their analysis workflows.

SEQUENCING HAPLOID DRONES FROM ROYAL JELLY AND HONEY BEE POPULATIONS FOR DETECTION OF DIFFERENTIATION AND SELECTIVE SWEEPS.

David Wragg[1], Benjamin Basso[2], Yves Le Conte[3], Jean-Pierre Bidanel[4], Alain Vignal[1]

[1]UMR 1388 GenPhySE, Centre INRA de Toulouse, Castanet-Tolosan, France, [2]ITSAP-Institut de l'abeille, UMT PrADE, UR0406 Abeilles et Environnement INRA, Centre INRA PACA, Avignon, France, [3]INRA, UMT PrADE, UR0406 Abeilles et Environnement, Centre INRA PACA, Avignon, France, [4]UMR 1313 GABI, INRA, Domaine de Vilvert, Jouy-en-Josas, France

In France, until the middle of last century the endemic populations of honey bees were represented by a single subspecies, the black bee, Apis mellifera mellifera. More recently, apiculturists have demonstrated an interest in using other subspecies and hybrids between A. m. ligustica, A.m. caucasica, and A. m. mellifera, which have been found to be more efficient producers of honey and royal jelly. Most population genetic studies to date in A. mellifera have been performed using microsatellite markers, and more recently with medium density SNP microarrays. However, recent advances in high-throughput sequencing and declining costs now make population studies at the genome level feasible and two recent publications have proven its utility for worldwide surveys of bee populations. The SeqApiPop project is designed to study the structure of French bee populations by whole genome sequencing (IlluminaTM) of one drone per colony for 1000 colonies. The sequencing of haploid males rather than diploid worker individuals enables high confidence variant calling, reducing the need for greater depth of coverage and will also allow for easy determination of haplotypes. As a test study, 30 individuals from a standard population used for producing honey and 30 from another population selected for the production of royal jelly were sequenced resulting in the detection of more than 3.5 million SNPs. Analysis of the SNP data reveals several chromosomal regions of differentiation between the two populations, including loss of heterozygosity, suggestive of selective sweeps.

HEPTANUCLEOTIDE SEQUENCE CONTEXT EXPLAINS SUBSTANTIAL VARIABILITY IN NUCLEOTIDE SUBSTITUTION PROBABILITIES ACROSS THE HUMAN GENOME

Varun Aggarwala[1], Benjamin F Voight[2,3]

[1]University of Pennsylvania - Perelman School of Medicine, Genomics and Computational Biology Program, Philadelphia, PA, [2]University of Pennsylvania - Perelman School of Medicine, Department of Systems Pharmacology and Translational Therapuetics, Philadelphia, PA, [3]University of Pennsylvania - Perelman School of Medicine, Department of Genetics, Philadelphia, PA

The rate of single nucleotide polymorphism varies by ~1000 fold across the human genome and fundamentally impacts evolution and incidence of genetic disease. The identities of the single nucleotides that immediately flank a polymorphic site – or the site's trinucleotide local sequence context – substantially influence the probability that a nucleotide change will occur. In human populations, the impact of local sequence context on polymorphism rate has not been fully described and is untested beyond the trinucleotide context. To examine the boundaries of the window of local sequence that impacts the probability of polymorphism, we developed a statistical framework to compare different local sequence lengths using non-coding genomic data obtained from the 1000 Genomes Project. We demonstrate that a heptanucleotide sequence context – that is, a model that incorporates the three adjacent nucleotides located both 5′ and 3′ to a polymorphic site – accounts for up to 93% of the variability in the probability of nucleotide substitution observed genome-wide. Our study also reveals previously undocumented variability in the probability of cytosine-to-thymine transition substitutions at CpG dinucleotides. Extension of our statistical framework into coding genomic data demonstrates additional context-specific variability in the probabilities of amino acid substitutions. Based on these observations, we present two statistics, informed by our best performing sequence context model, that are relevant for clinical studies: a substitution tolerance score for genes and a novel tolerance score for amino acids.

CORRELATION OF MITOCHONDRIAL DNA HETEROPLASMY AND COPY NUMBER IN HUMAN TISSUES

Manja Wachsmuth[1], Alexander Hübner[1], Mingkun Li[1,2], Mark Stoneking[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]Fondation Mérieux, Lyon, France

The human mitochondrial genome is a circular DNA molecule consisting of about 16,500 bp and harboring genes essential for respiratory chain function. It is present in several copies in a single cell. Due to different energy requirements, the total cellular number of mitochondria and therefore mitochondrial DNA (mtDNA) strongly varies between tissues. The mitochondrial genomes of an individual can either be all identical or differ at defined positions. The state of intra-individual variability in mtDNA sequence is referred to as heteroplasmy. Heteroplasmic mutations have been shown to occur in a highly tissue- and site-specific fashion, suggesting a potential role for positive selection on somatic mtDNA mutations [1]. Furthermore, a significant age-dependent increase in minor allele frequency at specific heteroplasmic sites, and in the total number of heteroplasmic sites per individual has been reported.
Little is known about the functional impacts of heteroplasmy on mitochondrial function. As the most frequently occurring heteroplasmy sites are located in the mtDNA control region, an impact on replication regulation and hence copy number can be suggested. Furthermore, the most comprehensive data sets analyzing correlations between age and mtDNA copies indicate a slight but significant decrease of copy number with age [2]. However the impact that heteroplasmy might have on this decline has not been examined previously. Here, we investigated mtDNA copy number in four different tissues obtained at autopsy from 152 human individuals. These samples were previously analyzed in detail for heteroplasmy [1]. As the determined number of copies can vary depending on the method used, we compare mtDNA copy numbers from shot-gun sequencing with those measured by quantitative PCR. For quantitative PCR we applied the droplet digital PCR method which allows precise detection of absolute molecule numbers. Finally, we compare mtDNA copy number with the presence of heteroplasmic sites and the frequency of the minor allele in order to gain further insights into the effects of heteroplasmy on mitochondrial function.

References:

[1] Li M *et al*. (2015), PNAS, in press.
[2] Mengel-From J *et al*. (2014), Hum Genet 133:1149-59.

THE SECRETS OF GWAS ARE WRITTEN IN THE READS

Claes Wadelius[1], Marco Cavalli[1], Gang Pan[1], Helena Nord[1], Ola Wallerman[1,2], Emelie Wallen Arzt[1], Olof Berggren[3], Ingegerd Elvers[2,4], Maija-Lena Eloranta[3], Lars Rönnblom[3], Kerstin Lindblad Toh[2,4]

[1]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala, Sweden, [2]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala, Sweden, [3]Science for Life Laboratory, Department of Medical Sciences, Uppsala, Sweden, [4]Broad Institute, Cambridge, MA

Genome-wide association studies (GWAS) have identified a large number of disease associated loci, but in few cases have the functional variant and the gene it controls been identified. To systematically identify candidate regulatory variants we sequenced ENCODE cell lines and used public ChIP-seq data to look for allele-specific transcription factor (TF) binding. We found 15,644 candidate regulatory SNPs of which more than 20% were rare, allele frequency <1%, and showed evidence of larger functional effect than common SNPs. This high frequency of rare functional variants adds heterogeneity to GWA studies of traits and expression and may explain divergent GWAS results between populations and why SNPs with the highest association signal rarely are functional. The majority of allele specific variants (95%) were specific to one of the six studied cell types. By examining GWAS loci we found >600 allele-specific candidate SNPs, 184 of which were highly relevant in our cell types. Results were confirmed by luciferase assays, EMSA and analysis of expression in stimulated primary cells. Functionally validated SNPs and genes support identification of an intronic SNP in MERTK associated to risk for liver fibrosis, a SNP in SYNGR1 affecting risk for rheumatoid arthritis and primary biliary cirrhosis as well as a SNP in the last intron of COG6 affecting risk for psoriasis. The results from ChIP-seq of 20 TFs detect 92-99% of candidate regulatory SNPs in a cell even if 100 TFs had been analyzed. We show that by repeating ChIP-seq experiments of 20 selected transcription factors in three to ten people most common polymorphisms can be interrogated for allele-specific binding in a selected cell type. Furthermore, the most informative TFs to ChIP-seq can be predicted for a new cell type or tissue of interest for disease. Our strategy may help to alleviate the current bottle neck in functional annotation of the genome.

GO-PCA: AN UNSUPERVISED METHOD TO EXPLORE BIOLOGICAL HETEROGENEITY BASED ON GENE EXPRESSION AND PRIOR KNOWLEDGE

Florian Wagner[1], Sandeep Dave[2,3]

[1]PhD Program in Computational Biology and Bioinformatics, Center for Genomic and Computational Biology, Duke University, Durham, NC, [2]Center for Genomic and Computational Biology, Duke University, Durham, NC, [3]Duke Cancer Institute, Duke University, Durham, NC

Motivation
Gene expression profiling is a cost-efficient and widely used method to characterize heterogeneous populations of cells, tissues, biopsies, or other specimen. In order to gain insights from such large-scale datasets, researchers typically apply generic unsupervised methods, e.g. hierarchical clustering or factor analysis. However, generic methods fail to effectively exploit the significant amount of knowledge we have about the molecular functions of genes. We asked whether directly incorporating this knowledge into an unsupervised algorithm could produce effective representations of biological heterogeneity and yield readily interpretable results.
Results
We present GO-PCA, an unsupervised expression-based method that relies on prior knowledge about gene functions in the form of gene ontology (GO) annotations. GO-PCA combines principal component analysis (PCA) with nonparametric GO enrichment analysis, in order to generate a compact set of functionally defined expression signatures that collectively aim to describe all major axes of biological variation in the data.
We first applied GO-PCA to expression profiles of diverse cell populations from the human hematopoietic lineage, and obtained signatures with clear biological interpretations representing almost all cell lineages. We then applied our method to an expression panel of glioblastoma (GBM) tumor biopsies, and obtained multiple signatures that were associated with previously described GBM subtypes. Surprisingly, GO-PCA discovered a cell cycle-related signature that exhibited significant correlation with the Proneural subtype, but not with the prognostically favorable GBM CpG Island Methylator (G-CIMP) subtype. Previous expression-based classifications have failed to separate these subtypes, suggesting that GO-PCA can detect biological heterogeneity that is missed by other methods, and that the G-CIMP subtype is characterized in part by lower mitotic activity.
Conclusions
Our results show that GO-PCA is a powerful and flexible expression-based method that facilitates exploration of biological heterogeneity, without requiring additional types of experimental data. The resulting low-dimensional representation lends itself to interpretation, hypothesis generation, and further analysis.

WHOLE-GENOME BISULFITE SEQUENCING OF ACUTE LYMPHOBLASTIC LEUKEMIA CELLS

Per Wahlberg[1], Anders Lundmark[1], Jessica Nordlund[1], Stephan Busche[2], Erik Forestier[3], Gudmar Lönnerholm[4], Tomi Pastinen[2], Ann-Christine Syvänen[1]

[1]Uppsala University, Dept. of Medical Sciences, Uppsala, Sweden, [2]McGill University, Dept. of Human Genetics, Quebec, Canada, [3]Umeå University, Dept. of Medical Biosciences, Umeå, Sweden, [4]Uppsala University, Dept. of Children's and Women's Health, Uppsala, Sweden

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer in the developed countries. Aberrant patterns of methylation have been associated with cancer and epigenetic modifiers such as the ten eleven translocation (TET) family of enzymes and DNA methyltransferases (DNMTs) are recurrently mutated in leukemic cancers. We have previously documented large differences between ALL subtypes and normal B- and T-cells using the Illumina Human 450K BeadArray platform. To further investigate the distribution of CpG methylation in ALL cells, we generated Whole Genome Bisulfite Sequencing (WGBS) data at high coverage (20-30X) of four ALL patients with different genetic subtypes of ALL. In our study we also included low-pass (~7X) WGBS data from B- (n=8) and T-cells (n=14) and additional BCP-ALL t(12;21) patients (n=3). A global analysis show that the ALL methylome largely follows an anticipated pattern of DNA methylation for an mammalian cell-type, with low levels of methylation at promotor regions and high methylation in genic and intergenic sections of the genome. However, we did observe aberrant intermediate CpG island methylation in ALL samples, particularly in ALL samples diagnosed with the t(12;21)ETV-RUNX1 translocation, affecting approximately 4,000 CpG islands. We found that the distorted intermediate CpG island methylation can to a large extent be explained by a mixture of cell populations with locally heterogenous methylation patterns, as is evident by analyzing patterns of methylation states along individual sequence reads. A DMR analysis between ALL samples and B- and T-cell identified between 103,275 - 69,849 DMRs with an average size of 379 bp and range from 6 bp up to 43 Kb. Evolutionary conservation scores show that DMR sequences are more conserved than surrounding regions. The ALL specific DMRs and cell-type specific B- and T-cell DMRs are distributed across the genome in a similar manner with the exception of hypermethylated DMRs annotated to CpG islands that are enriched in ALL samples. DMRs are under-represented in intergenic regions and strongly enriched to regions associated with enhancer binding. The observation, by us and others, that DMRs often represent transcription factor binding sites suggest that methylation pattern can be used to identify active regulatory regions and can possible lead to further understanding of the molecular pathways activated or inactivated in ALL cells.

CHARACTERIZATION OF MAIZE B73 TRANSCRIPTOME BY HYBRID SEQUENCING

Bo Wang[1], Michael Regulski[1], Andrew Olson[1], Joshua Stein[1], Tyson Clark[2], Yinping Jiao[1], Doreen Ware[1,3]

[1]Cold Spring Harbor Laboratory, 1 Bungtwon Road, Cold Spring Harbor, NY, [2]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA, [3]USDA ARS NAA Robert W. Holley Center for Agriculture and Health, Cornell University, 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 mRNA transcripts, particularly with respect to alternatively spliced isoforms. Although second-generation RNA-seq provides a quantitative assay for transcriptional and posttranscriptional events, the accurate reconstruction of full-length mRNA isoforms is challenging with short-read technologies. By producing much longer reads, third generation sequencing offers to solve the assembly problem, but can suffer from lower read accuracy and throughput. Here, we combine these complementary technologies to define and quantify high-confidence transcript isoforms in maize. Six tissues (root, pollen, embryo, endosperm, immature ear, and immature tassel) of the B73 inbred line were used for mRNA sequencing with the Illumina Hiseq2000 PE101 platform to comprehensively quantitate gene/isoform expression. In parallel, intact cDNAs from the same samples were sequenced using the PacBio RS II platform. The latter used four size fractionated libraries (1-2kb, 2-3kb, 3kb-5kb, 5kb-10kb) and generated more than 2 million full length reads. Preliminary findings suggest that mechanisms of alternative splicing are differentially employed between different tissues. In addition, these data show promise to dramatically improve the status of maize genome annotation, with the detection of previously unidentified transcript isoforms, and uncovering previously unrecognized genes.

GENOME-WIDE CROSSOVER DISTRIBUTION IN MALE AND FEMALE OF MAIZE

Minghui Wang[1], Qi Sun[1], Pawlowski Wojtek[2]

[1]Cornell University, Bioinformatics Facility, Ithaca, NY, [2]Cornell university, Bioinformatics Facility, Ithaca, NY, [3]Cornell University, Plant Breeding and Genetics, Ithaca, NY

In most species, meiotic crossovers are not distributed randomly along chromosomes but form distinct sites (recombination hotspots) characterized by recombination rates higher than those of surrounding chromosome regions. High-resolution mapping of crossover hotspots has been accomplished in several species of plants, animal, and fungi, but mostly those with relatively small genomes. However, it is challenging in species with complex genomes that contain high proportion of repetitive DNA, as presence of repetitive DNA makes determination of unique chromosome locations difficult. We are mapping the distribution of meiotic crossovers in maize using low-coverage illumina sequencing. For accurate mapping of crossover sites in the complex genome, we developed a bioinformatic pipeline to minimize the confounding effect of repetitive DNA. While mapping crossovers in male and female meiosis, we found that while the number of crossovers per chromosome does not differ between the sexes in maize, their distribution does

PRINCIPLES OF EPIGENOME CONSERVATION

Jia Zhou[1,2], Xiaoyun Xing[1,2], Bo Zhang[1,2], Daofeng Li[1,2], Renee L Sears[1,2], Nicole B Rockweiler[1,2], Rebecca F Lowdon[1,2], Hyung Joo Lee[1,2], Ting Wang[1,2]

[1]Washington University, Genetics, St. Louis, MO, [2]Washington University, Center for Genome Sciences and Systems Biology, St. Louis, MO

Uncovering mechanisms of epigenome evolution is an essential step in understanding the evolution of different cellular phenotypes. Epigenome evolution depends on accurate maps of epigenome conservation. Here we describe our efforts to map conservation of DNA methylation genome-wide and tissue-specifically and to investigate the drivers of epigenome conservation or turnover. DNA methylation is a heritable epigenetic mark with important roles in diverse biological processes such as X chromosome inactivation, transposable element repression, genomic imprinting, and tissue-specific gene expression. Using a comparative epigenomics approach, we identified and compared the tissue-specific DNA methylation patterns of rat against that of mouse and human across three shared tissue types. We confirmed that tissue-specific differentially methylated regions are strongly associated with tissue-specific regulatory elements. Comparisons between species revealed that 11-37% of this tissue-specific DNA methylation pattern is conserved, a phenomenon that we defined as epigenetic conservation. Conserved DNA methylation is also accompanied by conservation of other epigenetic marks including histone modifications. Although a significant amount of loci-specific methylation signals are epigenetically conserved, the majority of tissue-specific DNA methylation is not conserved across the species and tissue types that we investigated. Examination of the genetic underpinning of epigenetic conservation suggests that primary sequence conservation is a driving force behind epigenetic conservation. In contrast, evolutionary dynamics of tissue-specific DNA methylation are best explained by the maintenance or turnover of binding sites for important transcription factors. Our study thus provides a framework for investigating the mechanisms of epigenetic conservation.

RAPID, COMPREHENSIVE, WHOLE-GENOME INTERROGATION OF MEDICAL SEQUENCING DATA

Barry Moore[1], Alistair Ward[1], Carson Holt[1], Shawn Rynearson[1], David Nix[2], Brett Milash[2], Aaron Quinlan[1], Mark Yandell[1], Gabor Marth[1]

[1]USTAR Center for Genetic Discovery, University of Utah School of Medicine, Salt Lake City, UT, [2]Huntsman Cancer Institute, University of Utah Health Care, Salt Lake City, UT

We present a highly innovative software pipeline developed at the USTAR Center for Genetic Discovery for comprehensive genomic interrogation of massive patient collections in medical research projects and for rapid disease gene/variant identification for personalized clinical care.

1) Consensus variant identification using a graph approach. Multiple state-of-the-art variant identification tools (e.g. GATK, FREEBAYES) are used to identify a large and highly sensitive pool of candidate variants before undergoing a variant adjudication procedure. Sequencing reads are mapped to a Variant Graph constructed from all candidates using a graph Smith-Waterman algorithm. Candidates are either confirmed or discarded based on mapping to branches in the graph, resulting in substantial improvements in sensitivity and specificity.
2) Integration of structural variants. Current sequence analysis workflows focus on short variants (SNPs and INDELs) since integration of SV detection tools is difficult, and computationally expensive. We have integrated a collection of SV detection methods (LUMPY, TANGRAM, WHAM, DELLY) into the pipeline, whose variants are curated using the Variant Graph approach.
3) De novo variant detection. A highly sensitive and accurate, k-mer based, reference-free detection tool, RUFUS, is used to detect de novo variants in, for example, trio samples.
4) Variant prioritization. To be clinically relevant, variant prioritization needs to be automated in the pipeline. VAAST, pVAAST and PHEVOR make explicit use of all curated variants and phenotype information to rank the variants according to their likelihood of being causative. Additionally, PHEVOR is able to integrate information from RNA-Seq data for prioritizing regulatory variants.
5) Visually driven analysis and manual data inspection. The results of automated analyses are often insufficient to reveal the genetic causes underlying disease, so expert review is necessary. We have integrated web-based data analysis and inspection apps (OPAL, IOBIO), allowing all scientists and clinicians to query and inspect all results. These are then loaded into GEMINI, so they can be queried across multiple patient and research cohorts and existing databases.
6) Massively parallel execution. Our analysis pipeline is integrated in the PARALLEL ARCHITECT execution framework enabling parallelization of the analysis to be performed rapidly for a single patient in a clinical setting, and efficiently for massive sample collections in a research setting.

GRAMENE: A RESOURCE FOR COMPARATIVE PLANT GENOMICS AND PATHWAYS

Marcela K Monaco[1], Kapeel Chougule[1], Yinping Jiao[1], Sunita Kumari[1], Joe Mulvaney[1], Andrew Olson[1], Joshua Stein[1], Bo Wang[1], Sharon Wei[1], Vindhya Amarasinghe[2], Justin Elser[2], Sushma Naithani[2], Justin Preece[2], Peter D'Eustachio[3], Robert Petryszak[4], Paul Kersey[4], Pankaj Jaiswal[2], Doreen Ware[1,5]

[1]CSHL, Cold Spring Harbor, NY, [2]OSU, Corvallis, OR, [3]NYU, SOM, New York, NY, [4]EMBL-EBI, Cambridge, United Kingdom, [5]USDA ARS NAA, Ithaca, NY

Agriculture in the modern world faces multiple global challenges including food security and adaptation to climate change. A fast growing number of plant genomes, representing the diversity of species in the plant kingdom, is now available through bioinformatics resources that process and integrate vast amounts of data to allow plant researchers and breeders to query and visualize it for specific purposes (e.g., genomic regions with domestication signatures or lines with desirable traits), thus making it possible to better understand the complexity, structure, and evolution of plants.
Gramene (www.gramene.org) is one of such instrumental framework in plant research. A curated resource for comparative functional genomics in crops and model plant species, Gramene includes components produced in collaboration with the plants division of Ensembl Genomes and Reactome. Its strength derives from the application of a phylogenetic framework for genome comparison, and integration of genome annotation and functional data using ontologies. The current release (build 44) includes two major components of datasets. (1) The plant genome portal includes 39 complete reference genomes, with strong representation of monocots and dicots, as well as lower plants. Species added within the last year include cocoa, peach, wild mustard, wild grasses (including five Oryza species), an unicellular green algae (name), and the basal angiosperm Amborella, in addition to a new assembly for bread wheat. For each reference genome, we incorporate community annotation from primary sources and enrich this information with a series of standardized analyses. These include functional annotation by InterProScan and classification using controlled vocabularies, Gene Ontology and Plant Ontology. Evolutionary histories are provided by Compara phylogenetic gene trees and complemented by analyses of whole genome alignments. In recent years, Gramene has positioned itself as a major resource for genetic variation data, from maize, rice, Arabidopsis, barley, sorghum, wheat, grape, tomato, and Brachypodium. (2) The Plant Reactome, a plant metabolic and signaling pathways database provides reference rice pathways and orthology-based projections for 33 other plant species, along with pathway visualization and analyses tools. In collaboration with the EBI Gene Expression ATLAS project, we have analyzed numerous publicly available Arabidopsis and rice gene expression datasets that we link from the genome and pathway portals. Gramene is supported by an NSF award (IOS-1127112) and is produced in collaboration with the EBI-EMBL, the OICR, and the ASPB.

THE EFFECT OF NATURAL GENETIC VARIATION ON TRANSCRIPTION FACTOR BINDING AND ENHANCER ACTIVITY IN PRIMARY BLOOD CELLS

Stephen Watt[1], Louella Vasquez[1], Lu Chen[1,2], Joost Martens[3], Willem Ouwehand[2], Henk Stunnenberg[3], Tomi Pastinen[4], Kate Downes[2], Nicole Soranzo[1,2], BLUEPRINT EpiVar Working Group[1,2,3,4]

[1]Wellcome Trust Sanger Institute, Cambridge, United Kingdom, [2]University of Cambridge, Department of Haematology, Cambridge, United Kingdom, [3]Radboud University, NCMLS, Nijmegen, Netherlands, [4]McGill University, Montreal, Canada

Neutrophils are the most abundant white blood cell in the circulation and are the first cells to respond to infection and inflammation. These cells are short lived and approximately 70^{10} are released from the bone marrow each day. Homeostatic regulation is tightly controlled within a healthy individual; however the large variation of neutrophil count in the population indicates genetic factors alter regulation.

Our aim is to better understand how genetic variants influence transcriptional programs which result in altered blood phenotypes. To achieve this we have collected from 200 healthy blood donors' three cell types' monocytes, neutrophils and T-cells. We have performed whole genome sequencing and variant identification on these individuals, along with functional genomics data for; gene expression, chromatin state and methylation analysis.

Here I present results from epigenetic state and transcription factor location component of this project. We employed an AHT-ChIP-seq system to profile chromatin state for the modified histones; H3K4me3, H3K27ac, H3K4me1 all indicative of active regions and H3K27me3 for repressed. In addition we are assaying the transcription factors PU.1, the key pioneer factor expressed at the branch point during haematopoietic differentiation towards the myeloid lineage. The second transcription factor studied is from the CEBP CCAAT/enhancer-binding protein family which are crucial for neutrophil development, of these CEBPβ is primarily expressed in terminally differentiated mature neutrophils. To date we have generated over 1200 ChIP-seq data sets. With this we are exploring how genetically driven variation in transcription factor binding from key cellular proteins influences phenotypic variation in a healthy cohort.

EXTREME RECOMBINATION RATES SHAPE GENOME VARIATION AND EVOLUTION IN HONEYBEES

Andreas Wallberg[1], Sylvain Glémin[2], Matthew T Webster[1]

[1]Uppsala University, Medical Biochemistry and Microbiology, Uppsala, Sweden, [2]University of Montpellier, Evolutionary Sciences, Montpellier, France

We used population-scale genome sequencing to uncover population history, the genetic basis of local adaptation, and the causes and consequences of recombination rate variation in the honeybee Apis mellifera. We sequenced the genomes of 140 honeybees from a worldwide sample of 14 populations, identifying a total of 8.3 million SNPs. Levels of genetic diversity in honeybee populations are surprisingly high, and indicate effective population sizes in the range of 2 – 5 x 105. Patterns of genetic variation reflect large historical fluctuations in population size caused by past oscillations in the earth's climate. We do not find evidence for an African origin of A. mellifera, and date the emergence of extant populations ~300,000 years ago. We identify genomic signatures of adaptation to temperate and tropical climates, which are enriched in unmethylated genes that are mainly expressed in worker bees. These include genes involved in morphology, innate immunity, metabolism that might underlie geographic variation in reproduction, dispersal and disease resistance. We also find evidence for selection on sperm motility–related genes, which could potentially contribute to high reproductive success of African bees. These genetic changes form the basis for understanding the genetics of adaptation to pathogens and climate in honeybees.

Honeybees and other social insects have extremely high recombination rates (>20 cM/Mb), but the reason for this is unclear. We constructed a fine-scale map of recombination rate variation in the A. mellifera genome using our population-scale sequencing data. We do not find evidence for the existence of recombination hotspots, which contrasts with vertebrate genomes, but is similar to other invertebrates, suggesting common mechanisms for initiation of recombination. Genes that are not methylated in the germline, which tend to be expressed mainly in the two female castes, tend to have increased crossover rates, suggesting that DNA methylation is a major factor modulating recombination rate. The site frequency spectrum in honeybee populations is strongly skewed from neutral expectations. Rare variants are enriched for AT alleles whereas GC alleles tend to segregate at higher frequencies. This indicates a dominant effect of GC-biased gene conversion towards transmitting GC alleles during meiosis. We estimate that this bias is 5 – 50 times stronger than in humans. We uncover further evidence that gBGC specifically affects transitions and favours fixation of CpG sites, a finding that may explain the large excess of CpG sites in the honeybee genome. Recombination, via gBGC, therefore appears to have profound effects on genetic diversity in honeybees and interferes with the process of natural selection. This has important implications for genome evolution in honeybees and other social insects.

OGE.GRAMENE: A COMPREHENSIVE PLATFORM FOR STUDYING ORYZA GENOME EVOLUTION

Sharon Wei[1], Joshua C Stein[1], Kapeel Chougule[1], Yu Yeisoo[2], Dario Copetti[2], David Kudrna[2], Jianwei Zhang[2], Jose L Goicoechea[2], Xiang Song[2], Manyuan Long[3], Michael Sanderson[4], Carlos A Machado[5], Scott Jackson[6], Mingsheng Chen[7], Rod A Wing[2], Doreen Ware[1,8]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]University of Arizona, Arizona Genomics Institute, Tuscon, AZ, [3]University of Chicago, Department of Ecology and Evolution, Chicago, IL, [4]University of Arizona, Department of Ecology and Evolution, Tuscon, AZ, [5]University of Maryland, Department of Biology, College Park, MD, [6]University of Georgia, Center for Applied Genetic Technologies, Athens, GA, [7]Chinese Academy of Sciences, Institute of Genetics and Developmental Biology, Beijing, China, [8]USDA, ARS, Ithaca, NY

In collaboration with the NSF-funded Oryza Genome Evolution (OGE) Project and I-OMAP, Gramene has developed a dedicated genus-level resource: oge.gramene.org. Built upon the Ensembl infrastructure, oge.gramene.org serves 11 whole genomes and 4 chr3 short arms. Whole-genomes include 10 AA genomes, 1 BB, 1 FF and 1 outgroup. Chr3 short arms include one of BBCC, CC, GG genomes. All these genomes have evidence based MAKER-P genes consistently annotated by AGI on Texas Advanced Computing Center (TACC) computers, minimizing algorithm/platform specific biases. Other sequence-level annotations include non-coding RNA gene predictions, Oryza-specific repeat features, Oryza EST feature mappings, and protein feature annotations with InterPro domains, cross-reference to external identifiers, and GO. Variation data sets when available were loaded into Ensembl variation databases, followed by functional consequences predictions and classifications with SO terms. All these can be visualized in the context of genomic and gene structure at oge.gramene.org. At the phylogenomics level, Gramene computed and serves about 22,867 phylogenetic gene trees covering the taxonomy of ten Oryza species and four outgroups, including Leersia perrieri, Brachypodium distachyon, Sorghum bicolor, and Arabidopsis thaliana. Derived from the orthologs of these gene trees, syntenic regions were computed and used to categorize genes as syntelogs, ortholog_nc (nonSyntenic), and ortholog_sc (has non-syntenic ortholog which is itself syntenic). Ancestrally conserved regions and structural rearrangements are defined by whole-genome alignments and displayed in a number of informative ways, including a multi-species view that allows graphical stacking of browsers and interspecies navigation. In order to capture updated assembly of one of the AA whole genome O.meridionalis, a second, and presumably final release of the website is in preparation. In addition to resolving many outstanding questions in the evolutionary history of the Oryza genus, this resource will provide a basis for functional characterization of genes and the identification of targets for agronomic improvement of rice.

POPULATION GENOMIC ANALYSIS OF PLASMODIUM VIVAX FROM COLOMBIA REVEAL SUBSTANTIAL GENETIC DIVERSITY AND A SELECTIVE SWEEP ASSOCIATED WITH DRUG RESISTANCE.

David J Winter[1], Maria Pacheco Delgado[1,2], Reed A Cartwright[1], Ananias A Escalante[1,2]

[1]Arizona State University, Center for Human and Comparative Genomics, Tempe, AZ, [2]Temple Univsersity, Institute for Genomics and Evolutionary Medicine, Philadelphia, PA

Plasmodium vivax is the most common malarial parasite outside of Africa. Approximately 2.5 billion people live at risk of the disease, which has centers of transmission in South and Central America, Papua New Guinea and South East Asia. Although *P. vivax* malaria is less virulent than *P. falciparum*, the intermittent fevers caused by cycles of infection and re-infection create a substantial burden individuals and communities affected by the disease.

Genetic studies of *P. vivax* have largely focused on microsattelite loci, or genes known to be associated with disease processes or vaccine candidates. There have been relatively few whole genome studies, in part due to technical barriers created by the frequent presence of parasite lineages within a single host and difficulty with which parasite cells can be isolated from host cells.

Here were present a population genomic study of *P. vivax* from 8 patients in Tierralta (Department of Cordoba), northern Colombia. We show that each sampled patient is infected by a single parasite lineage. The proportion of sequencing reads originating from *P. vivax* varies greatly between samples, leading to some samples having low (<2x) coverage when aligned. Nevertheless, by correcting for the ascertainment bias created by this variable coverage we are able to study the demographic and population genetic history of *P. vivax* in Colombia. We show that, counter to findings of some single-gene studies, *P. vivax* is genetically diverse in Colombia (θ_w = 7.2 x 10 ^-4). A demographic reconstruction using the multiple sequentially Markovian coalescent method suggest this diversity arises in part by recent admixture between formerly isolated evolutionary lineages. Finally, a fine-scale analysis of genetic diversity across the genome reveals a recent selective sweep for a mutation associated with resistance to Sulfadoxine/pyrimethamine treatment.

These results demonstrate that whole genome studies of *P. vivax* can produce clinically important results, and illustrate the substantial genetic diversity of *P. vivax* in the Americas.

OPTIMIZING TRANS-ETHNIC TAG SNP SELECTION FOR GENOME-WIDE ASSOCIATION STUDIES

Genevieve L Wojcik[1], Christopher R Gignoux[1], Christian Fuchsberger[2], Henry R Johnston[3], Suyash Shringarpure[1], Alicia R Martin[1], Daniel Talium[2], Ryan Welch[2], Christopher S Carlson[4], Gonçalo Abecasis[2], Zhaohui S Qin[3], Kathleen C Barnes[5], Hyun M Kang[2], Michael Boehnke[2], Carlos D Bustamante[1], Eimear E Kenny[6]

[1]School of Medicine, Stanford University, Department of Genetics, Stanford, CA, [2]School of Public Health, University of Michigan, Department of Biostatistics, Ann Arbor, MI, [3]Rollins School of Public Health, Emory University, Department of Biostatistics and Bioinformatics, Atlanta, GA, [4]Fred Hutchinson Cancer Research Center, Public Health Sciences Division, Seattle, WA, [5]School of Medicine, Johns Hopkins University, Department of Medicine, Baltimore, MD, [6]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences, New York, NY

With the advent of large-scale biological repositories and multi-ethnic/multi-site consortia, there is a pressing need for new methods capturing trans-ethnic variation to enable genetic association studies. We have developed a novel algorithm to select tag SNPs prioritizing cosmopolitan variation while capturing fine resolution LD structure to maximize broad population applicability. This pipeline is designed to maximize imputation accuracy, rather than pairwise coverage. We conducted various iterations of selection criteria, focusing on minimum r^2 and minor allele frequency (MAF) thresholds, as well as examining various metrics of cross-population prioritization. Each iteration of potential scaffold sites was assessed through a leave-one-out validation approach to evaluate imputation accuracy and allow direct comparison of tag SNP performance. We applied this method recently to boost diverse population coverage for the Multi-Ethnic Genotyping Array (MEGA), a collaboration between Illumina and multiple consortia (PAGE, CAAPA, T2D Genes), by leveraging the whole genome sequence data available in the 1000 Genomes Project Phase 3 release. The final GWAS backbone consists of ~1.3 million SNPs genome-wide, enriched for low frequency variants informative across five continents. Imputation accuracy was found to be ≥90% for SNPs with MAF≥1% and ≥94% for MAF≥5% in all populations. This method is sensitive to the local ancestry context of haplotypes in mixed ancestry populations, such as African and Native American haplotypes in Hispanic/Latino and African-American populations, ensuring balanced coverage across ancestral backgrounds. This unified framework for tag SNP selection and imputation evaluation will be useful in large multi-ethnic epidemiological studies, large urban biobanks, as well as future biological repositories.

AN INTEGRATED OMICS PROFILE OF THE HUMAN BETA CELL MODEL ENDOC-BH1

Brooke N Wolford[1], Stephen CJ Parker[2], Xingwang Li[3], Emaly Piecuch[3], Asa Thibodeau[3], Eladio Marquez[3], Oscar Luo[3], Peter S Chines[1], Narisu Narisu[1], Michael R Erdos[1], John P Didion[1], D Leland Taylor[1], Duygu Ucar[3], Yijun Ruan[3], Michael L Stitzel[3], Francis S Collins[1]

[1]National Institutes of Health, National Human Genome Research Institute, Bethesda, MD, [2]University of Michigan, Departments of Computational Medicine & Bioinformatics and Human Genetics, Ann Arbor, MI, [3]The Jackson Laboratory, The Jackson Laboratory for Genetic Medicine, Farmington, CT

Type 2 diabetes (T2D) is a complex disease characterized by dysfunction of insulin-secreting pancreatic β cells and insulin resistance in peripheral tissues. T2D research involving human β cells has historically relied on the acquisition of islets from human cadavers to elucidate genomic, transcriptomic, and epigenomic (collectively "omics") signatures. In 2011, Ravassard et. al. produced the first functional human β cell line, EndoC-βH1, which secretes insulin when stimulated by glucose and therefore serves as a model for human β cell function. Here, we derive and integrate RNA-seq, ChIP-seq on modified histones, ATAC-seq and ChIA-PET data with dense genotyping analyses to describe the allelic omics profile of EndoC-βH1. Comparison of ChromHMM-derived chromatin state maps of EndoC-βH1 to those of existing cell types demonstrates that the chromatin architecture of EndoC-βH1 is strikingly similar to that of pancreatic islets. EndoC-βH1 enhancer regions significantly overlap T2D and related quantitative trait genome-wide association study (GWAS) SNPs. ChIA-PET enhancer-promoter interactions allow us to link allelic bias signatures in chromatin profiles with phase-consistent allele-specific expression in RNA-seq data. Analysis of ATAC-seq profiles using footprinting methods identifies general and cell type-specific transcription factor binding motifs. Finally, we report differential gene expression across high and low glucose conditions. Collectively, these data allow us to link T2D SNP-containing regulatory elements to target gene expression, thereby providing mechanistic insights about disease predisposition. Together, these findings help define the regulatory architecture of EndoC-βH1 and provide a foundation for future T2D and human β cell research.

DE NOVO GENOME ASSEMBLY FOR THE 21ST CENTURY

Yue Liu, Shwetha C Murlai, Daniel S T Hughes, Adam C English, Xiang Qin, Yi Han, Vanesa Vee, Min Wang, Eric Boerwinkle, Donna M Muzny, Jeffrey Rogers, Stephen Richards, Kim C Worley, Richard A Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

High quality reference genomes are the foundation for biological analyses. For many years, methods for creating high quality reference genomes relied upon expensive Sanger sequences with maps developed from libraries of large insert BAC or Fosmid clones. Advances early in this century have significantly reduced the cost per base of sequence generation, making shorter Illumina sequence plentiful. Alternative, long-read methods such as Pacific Biosciences can complement these short read data by filling in gaps and connecting contigs into larger scaffolds despite their higher error rates. Additional newer methods such as optical maps (BioNano Genomics and OpGen) and Hi-C chromatin analysis sequencing produce much more contiguous scaffolds that complement assembly methods that use other read data.

Here we report on our experience and experiments combining these data to produce better genome reference sequences. Experiments with human data, where we have a finished reference genome, have informed our current preferred methods. We are applying these methods to produce improved reference genomes for nonhuman primates and other mammals and experimenting with *Drosophila pseudoobscura* data to approach a finished quality genome. We have re-sequenced and assembled this smaller insect genome with our standard All-Paths, Atlas-Link and Atlas-Gapfill Illumina strategy, 70X Pacific Biosciences long reads with HGAP/Falcon assembly strategy, and BioNano Genomics and HiC analyses. Illumina assembly methods produced an excellent draft assembly (90 kb Contig N50, 2.3 Mb Scaffold N50) and our established PacBio with PBJelly method improves the Contig N50 more than three-fold to 295 kb. This is similar to our six released primate genome assemblies where PBJelly with only 9 to 12x long-read data improves the contig N50 by 2 to 4 fold from a range of 28-99 kb to 52-340 kb. Alternate strategies to error correct and *de novo* assemble *D. pseudoobscura* PacBio data using HGAP/Falcon produce 400-750 kb contig N50. We will present these data and our progress toward a finished quality assembly.

IMPROVING THE REFERENCE THROUGH LONG READ TECHNOLOGY - BETTER GENOMES FOR THE SHEEP AND THE COW

Kim C Worley[1], Adam C English[1], Xiang Qin[1], Shwetha C Murali[1], Daniel S T Hughes[1], Yi Han[1], Vanesa Vee[1], Timothy Smith[2], Jared E Decker[3], Brian Dalrymple[4], James Kijas[4], Noelle E Cockett[5], Jerry F Taylor[3], Juan Medrano[6], David C Schwartz[7], Shiguo Zhou[7], Donna M Muzny[1], Richard A Gibbs[1]
[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]U.S. Meat Animal Research Center, Clay Center, NE, [3]University of Missouri, Columbia, MO, [4]CSIRO Animal Food and Health Sciences, St. Lucia, QLD, Australia, [5]Utah State University, Logan, UT, [6]University of California, Davis, CA, [7]University of Wisconsin, Madison, WI

Reference genomes are the foundation of genomic analyses. As such, any flaw in the assembled sequence adversely affects research. Even "high quality" draft sequences have thousands of remaining gaps that impact the assessment of genes and other functional elements. We are applying our successful methods using Pacific Biosciences (PacBio) sequence and our PBJelly software (English, *et al.* 2012) to improve the high quality draft genome sequences for agriculturally important animals. We have produced 19x whole genome shotgun sequence using the PacBio technology from the Hereford cow and the Texel ram. The sequence reads are long (up to 10 kb average) and therefore useful for spanning gaps in a draft genome. For the sheep sequences mapped to chromosomes, the PBJelly method closed 70% of the gaps, reducing the number of contigs from 117,293 to 35,267. The assembly is more contiguous, with the contig N50 increased from 41.7 kb to 165.2 kb and almost ¼ of the contigs larger than 100kb (8,527; increased from 2,355). The PacBio data appears to improve the GC representation, increasing the G+C content slightly (0.1% of the contig bases). The fraction of the ambiguous bases (Ns) in the scaffolds decreased from 3.12% to 0.87% of the genome. For the cow reference where the starting assembly benefitted from Sanger sequencing with BAC clone data, the PBJelly method closed 46% of the gaps, reducing the number of contigs from 71,564 to 38,628. The assembly is more contiguous, with the contig N50 increased from 98.7 kb to 274.7 kb, 20% of the contigs exceed 100kb (8,240; increased from 7,605). The fraction of the ambiguous bases (Ns) in the scaffolds decreased from 0.78% to 0.23% of the genome.

We will also discuss ongoing research with other methods to approach finished quality genomes without using traditional expensive and manually intensive finishing efforts.

DETECTING AND ESTIMATING SPONTANEOUS MUTATION RATES IN *TETRAHYMENA THERMOPHILA*

Steven H Wu[1], David Winter[1], Allan Chang[2], Rebecca Zufall[2], Ricardo Azevedo[2], Reed Cartwright[1,3]

[1]Arizona State University, The Biodesign Institute, Tempe, AZ, [2]University of Houston, Department of Biology and Biochemistry, Houston, TX, [3]Arizona State University, School of Life Science, Tempe, AZ

The single-celled ciliate *Tetrahymena thermophila* provides a nearly perfect model system for the study of spontaneous mutation. Individual *T. thermophila* cells contain distinct germline and somatic genomes, During asexual reproduction the genes of the germline nucleus are not expressed, meaning germ-line mutations are not exposed to selection. When *T. thermophila* cultures are maintained in asexual growth otherwise lethal and highly deleterious mutations can accumulate in the germline genome, allowing the full range of spontaneous mutations to be studied.

In order to take advantage of this remarkable system, we are performing an experiment in which 96 lines of *T. thermophila* will accumulate mutations over 1500 generations. Accurately detecting mutations from short read sequencing produced from a pilot for this study has proved difficult. The challenges typically associated with detecting *de novo* mutations are compounded in Tetrahymena by an incomplete reference genome, and a strongly AT-biased genome (78% AT) that contains repeat elements. In order to overcome these challenges and we have developed a novel approach to detecting mutations from short read sequencing. We model base-calls for a given site in an assembly using the Dirichlet-multinomial distribution. The over-dispersion which can be captured by this approach can account for errors introduced during library preparation, sequencing and mapping to reference. We identify mutations by summing over all possible genotypes from both ancestor and descents, and all possible mutation and transmission events by using the tree pruning approach. Further more, we were able to estimate mutation rates from multiple categories using a expectation maximization algorithm.

Our model was validated on a pilot data set consisting of 12 mutation accumulation lines and a wide range of parameter combinations were tested in order to investigate the performance of the model. The model is able to accurately detect these mutation sites and they are validated by wet lab experiments. This results of this pilot study suggest *T. thermophila* has a remarkably low nucleotide mutation rate.

SYSTEMATIC CATALOGING OF THE HUMAN TISSUE SELECTOME AS A FOUNDATION FOR IDENTIFYING TARGETS OF HUMAN DISEASE

Hualin S Xi[1], Robert Y Yang[1], Jie Quan[1], John A Allen[2]

[1]Pfizer Inc, Computational Sciences CoE, Cambridge, MA, [2]Pfizer Inc, Neuroscience Research Unit, Cambridge, MA

Tissue-selective gene expression profiles represent keys to understand important regulatory mechanisms for tissue functions. To make available a comprehensive catalog of human tissue selectome, we conducted a transcriptome-wide survey of tissue-selective genes using an unprecedented collection of 1640 high-quality human RNA-seq samples from the GTEx Project, covering 31 human peripheral tissues and 13 brain subregions. Using a weighted tissue-selectivity scoring scheme that takes into account the similarity of related tissues and variability across individual samples, we identified thousands of genes selectively enriched in these tissue types, including many lower abundance genes vastly underestimated previously by microarray-based expression atlases. Functional enrichment and co-expression analyses showed numerous tissue-enriched functional modules. Remarkably, integrative analysis with genome-wide association studies (GWAS), covering a gamut of complex human diseases, demonstrated that genetic risk variants are highly enriched in tissue-selective expression profiles associated with relevant tissues. Using brain as an example, we further demonstrate the utility of systematic analysis anchored on the human tissue selectome in illuminating tissue physiology while providing a promising avenue to identify novel therapeutic target hypotheses.

CLOUD-BASED VARIANT ANALYSIS SOLUTION USING CONTROL-ACCESSED SEQUENCING DATA

Chunlin Xiao, Eugene Yaschenko, Stephen Sherry

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

Variation analysis plays an important role in elucidating the causes of various human diseases. The drastically reduced costs of genome sequencing driven by next generation sequence technologies now make it possible to analyze genetic variations with hundreds or thousands of samples simultaneously, but currently with the cost of ever increasing local storage requirements. The tera- and peta-byte scale footprint for sequence data imposes significant technical challenges for data management and analysis, including the tasks of collection, storage, transfer, sharing, and privacy protection. Currently, each analysis group facing these analysis tasks must download all the relevant sequence data into a local file system before variation analysis is initiated. This heavy-weight transaction not only slows down the pace of the analysis, but also creates financial burdens for researchers due to the cost of hardware and time required to transfer the data over typical academic internet connections. To overcome such limitations and explore the feasibility of analyzing control-accessed sequencing data in cloud environment while maintaining data privacy and security, here we introduce a cloud-based analysis framework that facilitates variation analysis using direct access to the NCBI Sequence Read Archive through NCBI sratoolkit, which allows the users to programmatically access data housed within SRA with encryption and decryption capabilities and converts it from the SRA format to the desired format for data analysis. A customized machine image (swift) with pre-configured tools (including NCBI sratoolkit) and resources essential for variant analysis has been created for instantiating an EC2 instance or instance cluster on Amazon cloud. Performance of this framework has been evaluated and compared with that from traditional analysis pipeline, and security handling in cloud environment when dealing with control-accessed sequence data has been addressed. We demonstrate that it is cost effective to make variant calls using control-accessed SRA sequence data without first transferring the entire set of aligned sequence data into a local storage environment, thereby accelerating variation discovery using control-accessed sequencing data.

EFFECT OF SEX, GENOTYPE, AND ENVIRONMENT ON GENE EXPRESSION AND ALTERNATIVE SPLICING IN INDIVIDUAL DROSOPHILA MELANOGASTER

Haiwang Yang[1], Yanzhu Lin[2], Kseniya Golovnina[1], Zhen-Xia Chen[1], Susan Harbison[2], Brian Oliver[1]

[1]National Institutes of Health, NIDDK, Bethesda, MD, [2]National Institutes of Health, NHLBI, Bethesda, MD

Like many eukaryotes, Drosophila melanogaster shows extensive alternative splicing. We explored the effect of Sex (S), genotype (G), and environment (E) on alternative splicing to garner insight into the relative importance of these factors, which are all likely to be important and under-explored factors in personalized medicine. We conducted RNA-seq on 726 individual sexed flies from 16 Drosophila Genetic Reference Panel (DGRP) inbred lines in three similar environments, and comprehensively studied alternative splicing differences. We observed 7210 G-biased, 6771 S-biased, and 3086 E-biased splicing events using a likelihood ratio based G-test of the three factors. After intersecting splicing changes with polymorphism data at the genome level, we also identified candidate causal mutations, the most highly significant ones were SNPs in splicing donor and acceptor sites. When normalized by the number of levels for each factor (e.g. 16 for G and 2 for S), sex became the dominant factor, which was also the case for gene-level expression. The vast amount of genotype- and sex-related splicing/expression in individual flies highlights the complex structure of the transcriptome, and the large task to realize personalized medicine based on gender and genotype.

MODELING REPRODUCIBILITY OF HIGH THROUGHPUT SEQUENCING DATA WITH TAIL DEPENDENCES WHEN PEARSON AND SPEARMAN CORRELATIONS FAIL

Tao Yang[1], Qunhua Li[2]

[1]Pennsylvanian State University, The Huck Institutes of the Life Sciences, University Park, PA, [2]Pennsylvanian State University, Department of Statistics, University Park, PA

The quality and reproducibility of sequencing-based experiments is essential to the reliability of downstream analysis and biological interpretation. Though Pearson and Spearman correlation coefficients are often used to assess the reproducibility of replicate sequencing experiments, they can be easily misled by highly repetitive regions or excessive amount of low count regions on the genome. Here we developed a novel reproducibility measure based on tail dependence that can overcome the drawbacks of correlation coefficients. We evaluate our methods on different sequencing experiments. Our measure is robust and can effectively distinguish experiments with different levels of reproducibility and quality. It is applicable to various sequencing based genetic and epigenetic data in which substantial amount of noise is often present. This measure will help practitioners identify suboptimal experiments and the causes of suboptimality.

LOW-FREQUENCY SEQUENCE VARIANTS INFLUENCE THE HUMAN METABOLOME

Bing Yu[1], Akram Yazdani[1], Fuli Yu[2], Alexander H Li[1], Ginger Metcalf[2], Donna M Muzny[2], Alanna C Morrison[1], Azam Yazdani[1], Richard A Gibbs[2], Eric Boerwinkle[1,2]

[1]University of Texas Health Science Center at Houston, Human Genetics Center, Houston, TX, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

The metabolome is a collection of small molecules resulting from multiple cellular and biologic processes, which can act as biomarkers of disease. Because of their molecular proximity to gene action, it is expected that effects of DNA sequence variation on the metabolome will be large. We performed whole genome sequencing (WGS), whole exome sequencing (WES) and measured 245 serum metabolites in a sample of 1,330 European Americans (EAs) and 1,361 African Americans (AAs) from the Atherosclerosis Risk in Communities (ARIC) Study. For WES in AAs, 17,150 genes with low-frequency functional variants (defined as nonsynonymous/splice/frameshift indels, MAF < 5%) were analyzed using gene-based burden tests and 17 genes with cMAF > 1% were identified to significantly affect 19 metabolites levels ($p < 1.2x10-8$). Depending on the metabolite, these sites were associated with 7-66% of the difference in metabolite levels, with an average effect of 26%. In particular, we identified nine genes harboring 18 loss-of-function (LoF, defined as splice/stop/frameshift indels) variants significantly affecting ten metabolites. We next leveraged the genes and metabolites to disease endpoints, and found the metabolites influenced disease constantly across two populations though the genetic spectrum showed substantial differentiation. For example, a LoF variant (c.481+1G>T) in SLCO1B1 was associated with high levels of hexadecanedioate (MAF = 2.6%, $p = 2.2x10-9$), a C16 dicarboxylic acid. SLCO1B1 is an organic ion transporter and the associations were replicated in an independent sample of 508 AAs ($p = 4.6x10-5$). High levels of hexadecanedioate consistently predicted the risk of incident HF beyond the effect of traditional risk factors among EAs and AAs (HR = 1.10, p = 0.003 and HR = 1.14, p = 0.0004, respectively). For WGS, sliding window designs was applied to aggregate variants within a window and analyzed them across the genome using multiple tests including T5, SKAT and a newly developed convex-concave rare variant selection (CCRS) method. Each sliding window contained 50 low frequency variants (MAF < 5%) with a skip length of 25 variants. By integrating –omic technologies into deeply phenotyped populations, these data and results are identifying new avenues of gene function, novel molecular mechanisms and potentially treatment targets for multiple disease.

DYNAMIC ENHANCER LANDSCAPES DURING PANCREATIC DIFFERENTIATION OF HUMAN ES CELLS

Feng Yue[1], Allen Wang[2], Yan Li[3], Bing Ren[3], Maike Sander[2]

[1]Penn State Schoolf of Medicine, Biochemistry and Molecular Biology, Hershey, PA, [2]UC San Diego School of Medicine, Pediatrics and Cellular & Molecular Medicine, La Jolla, CA, [3]UC San Diego School of Medicine, Ludwig Institute for Cancer Research, La Jolla, CA

One of the most fundamental and fascinating questions in biology is how different types of cells in a human body are derived from the same genome but possess distinct appearances and functions. Temporal and spatial-specific gene transcription, which is tightly controlled by cis-regulatory elements such as promoters, enhancers and insulators, is regarded as the main cause. Here we show that epigenetic priming of enhancers signifies developmental competence using during pancreatic differentiation of human ES cells. In this project, we performed RNA-Seq, GRO-Seq and ChIP-Seq for H3K4me3, H4K4me1 and H3K27Ac in each developmental stages, including hESCs, definitive endoderm (DE), primitive gut tube (GT), posterior forgut (FG), and pancreatic endoderm (PE). We observed that poised enhancer state could be used to predict the ability of developmental intermediates to respond to inductive signals. We further find that lineage-specific enhancers are first recognized by transcription factors involved in chromatin priming, while subsequent recruitment of lineage- inductive transcription factors leads to enhancer and target gene activation. Our results identify acquisition of a poised chromatin state at enhancers as a general mechanism by which progenitor cells gain the competence to rapidly activate lineage- specific genes in response to inductive signals.

DISCOVERY OF NOVEL GENETIC ELEMENTS BY METAGENOME MINING

Natalya Yutin, Sofiya Shevchenko, Eugene Koonin

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

Virophages are a unique group of double-stranded DNA viruses that rely on the replication machinery of giant viruses to reproduce within eukaryotic hosts. The conserved core of the virophage genomes consists of genes for primase-helicase, packaging ATPase, cysteine protease, and two capsid proteins. The latter four genes that are implicated in virion morphogenesis along with the protein-primed DNA polymerase that is encoded in a subset of virophages are shared with an expansive family of eukaryotic self-replicating DNA transposons, the Polintons, that are predicted to also form virus particles.

An exhaustive search of metagenomic sequence databases for homologs of conserved virophage proteins resulted in the discovery of several new families of genetic elements that include novel virophages, as well as previously unknown linear mobile elements. These elements share features with virophages and polintons and are likely to be capable of both integration into host genome and formation of infectious virions.

These results buttress the concept of a vast evolutionary network that spans the world of selfish genetic entities and might provide the "missing links" between bona fide viruses and other classes of self-replicating mobile elements.

GENETIC CONTROL OF CHROMATIN STATES AND GENE EXPRESSION IN HUMANS INVOLVES LOCAL AND DISTAL CHROMOSOMAL INTERACTIONS

Judith B Zaugg[1,2], Fabian Grubert[1], Maya Kasowski[1], Oana Ursu[1], Damek Spacek[1], Alicia Martin[1], Doug Phanstiel[1], Aleksandra Pekowska[3], Jonathan Pritchard[1,4,5], Carlos Bustamante[1], Lars M Steinmetz[1,3], Anshul Kundaje[1,6], Michael P Snyder[1]

[1]Stanford, Genetics, Stanford, CA, [2]European Molecular Biology Laboratory, Structural and Computational Biology, Heidelberg, Germany, [3]European Molecular Biology Laboratory, Genome Biology, Heidelberg, Germany, [4]Stanford, Howard Hughes Medical Institute, Stanford, CA, [5]Stanford University, Biology, Stanford, CA, [6]Stanford University, Computer Science, Stanford, CA

Uncovering how genetic variants affect gene regulation is fundamental to understanding human disease. Although gene regulation often involves long-range enhancer-promoter interactions it is unknown to what extent genetic variants in these elements act distally. Here we combine chromatin profiling for three promoter/enhancer histone marks in 76 individuals with HiC- and ChIA-PET-based physical chromatin interaction maps. We uncover interconnected networks of genetic links among regulatory elements: 10-15% have local histone quantitative trait loci (hQTLs), 15% of which affect distal elements. Physically interacting elements jointly contribute to gene expression, suggesting coordination among enhancers as well as genes. Transcription factor motif disruptions are enriched in hQTL peaks for local and distal sites and alter gene expression over long distances. Importantly, hQTLs are enriched for immune disease and cancer GWAS-SNPs. Overall, we show that genetic variation affects networks of regulatory elements and sequence variation in these elements may play an important role in mediating phenotypic variation in humans.

ANNOTATING NON-GENIC REGIONS IN ENSEMBL

Daniel R Zerbino, Nathan Johnson, Thomas Juettemann, Steven P Wilder, David Richardson, Avik Datta, Laura Clarke, Paul R Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom

Ensembl is one of the world's leading sources of information on the structure and function of the genome. It already provides an up-to-date, comprehensive and consistent database containing genome sequences, genes, non-coding RNAs, known variants, etc. However, it is very common for GWAS to return statistically significant hits that are hard to interpret because they fall outside of these annotations. We have therefore redesigned from the ground up the way we define regulatory regions along the genome.

Large projects such as ENCODE and Roadmap Epigenomics measure epigenomic marks through a wide spectrum of experimental assays across a diversity of cell types. They paved the way for more specialized endeavors such as BLUEPRINT, which focuses on hematopoietic differentiation.

Ensembl provides a summary of these public datasets by processing them through a unified pipeline and making them available through a single interface. The data is synthesized into the Regulatory Build, which defines functionally active regions across the human genome (on both the GRCh37 and GRCh38 assemblies) and mouse cell lines, assigning them a function wherever possible. We are currently collaborating with a number of data producing teams to broaden our coverage of cell types. Having defined a set of active regions along the genome, their activity levels can then be reliably determined in a new sample with a reduced set of assays.

Our goal is to progressively develop an annotation of regulatory elements in the genome, akin to the gene annotations that Ensembl already produces. To enrich this annotation we are looking at an array of technologies and assays to determine the links between enhancers and their target genes, such as eQTLs or Hi-C data. This will ultimately provide medical researchers greater resolution when prioritizing candidate variants for functional relevance.

In parallel, we are developing tools for basic research in epigenomics. For example, the WiggleTools browser allows users to remotely compute statistics on large collections of data, as produced for example by the BLUEPRINT project, without downloading data or software. Our simplified representation of epigenomes can also be used to quickly compute differences between cell types, and establish clear differentiation pathways. This opens the way for rapid identification of cell type based on epigenomic markers.

DYNAMIC DNA METHYLATION CHANGE OF TRANSPOSABLE ELEMENTS IN HUMAN CANCER

Bo Zhang[1,2], Rebecca Lowdon[1], Xiaoyun Xing[1], Daofeng Li[1], Joseph Costello[3], Ting Wang[1]

[1]Washington University School of Medicine, Department of Genetics, St. Louis, MO, [2]Washington University School of Medicine, Department of Developmental biology, St. Louis, MO, [3]University of California San Francisco, Department of Neurological Surgery, San Francisco, CA

Near half of human genome is transposable elements (TEs) sequence. Many studies have shown that TEs contributed regulatory elements including transcription factor binding sites and enhancers, and have shaped gene regulatory networks during evolution and in diseases.
Transposable elements are thought to be highly methylated in normal tissues, but become hypomethylated during carcinogenesis. How to reconcile the regulatory roles of some TE-derived enhancers and their global hypomethylation in cancer? How are TE-derived enhancers epigenetically modified in cancers, and what are the consequences of their deregulation? To answer these questions, we analyzed DNA methylomes of brain tumor (glioblastoma multiform, GBM), endometrial cancer (endomtrioid adenocarcinoma, EAC), colon cancer (CC), and several cancer cell lines in a genome-wide, unbiased fashion. We found distinct DNA methylation alteration in a large number of TEs, and most of these TE enriched active histone modification signal, which suggested their potential regulatory roles as promoters or enhancers. We observed distinct DNA methylation dynamics of these cancer-specific differentially methylated TEs (dmTEs) that specifically associated with glial and neuron cell types. Moreover, we discovered two opposite patterns of DNA methylation changes in TE-derived regulatory elements in cancer: 1) Tissue-specific active TE-derived regulatory elements became hypermethylated in cancer; 2) Tissue-specific silenced TE-derived regulatory elements became hypomethylated in cancer. These two distinct classes of dmTE-derived regulatory elements affected the expression of nearby genes, and may contribute to dysregulation of gene regulatory network.
This work revealed a novel mechanism that leads to deregulation of cancer genes – TE-derived regulatory elements specific to the cell type of origin are turned off, while those specific for other unrelated cell types are turned on. Perhaps this mechanism helps explain how cancer cells lose their identity.

GENOMIC, EPIGENOMIC, AND GENE EXPRESSION ANALYSIS REVEALS THE CONNECTION BETWEEN MUTATIONAL PATTERN AND LINEAGE OF B CELL LYMPHOMAS

Jenny Zhang[1,2], Andrea Moffitt[1,2], Cassandra Love[1,2], Sandeep Dave[1,2]

[1]Duke University, Center for Genomic and Computational Biology, Durham, NC, [2]Duke University, Cancer Institute, Durham, NC

Introduction

Exome and whole-genome sequencing have enabled definition of the genetic landscape of many cancers. Although some genes, such as TP53, are widely mutated, others are cancer type-specific. The nature of disease-specific genetic alterations lies in differences in their cell of origin's gene expression and epigenetic landscape.

Decades of work have defined the normal cell of origin of B Cell Lymphomas, including Mantle Cell Lymphoma (Naïve B Cell), Burkitt Lymphoma (Germinal Center), Germinal Center-like Diffuse large B-cell Lymphoma (GC), and Activated B-Cell like DLBCL (post-GC). The connection between the diseases and their cells of origin has been studied extensively through gene expression analysis, but only recently has it become technically possible to determine the genetic landscape of multiple lymphoma subtypes and relate it to the epigenetic context of the cells of origin.

Methods

Exome sequencing at 50-100 fold depth was applied to 200 B-Cell lymphomas including MCL, BL, GCB, and ABC DLBCL. In each disease, candidate driver genes were identified by recurrent somatic mutations in tumor-normal pairs, enrichment over background mutation rate of normal controls, and predicted mutation functional consequence.

To characterize the epigenomics of the normal cells of origin, Naïve, Germinal Center, and Memory B-Cells were subjected to chromatin immunoprecipitation and sequencing for H3K4me1, H3K4me3, H3Ac, H3K36me3, H3K27me3 and PolII. By integrating the signal from these chromatin markers, a gene-level open chromatin score was developed.

Gene expression analysis was applied to microarray data from BL and MCL, as well as normal naïve and germinal center B Cells, to determine differentially expressed genes.

Results

A significant number of genes were differentially mutated in lymphoma subtypes. For example, mutations in ATM, CCND1, and RB1 were found to be high in MCL; ID3 and MYC mutations were BL high, and PIM1 and CREBBP mutations were DLBCL high. The lineage context of this phenomenon was investigated by analyzing the chromatin state of the corresponding normal B Cells. Independent of gene expression, genes found to be more significantly more mutated in BL (Germinal Center origin) compared to MCL (Naïve origin) also exhibited a higher open chromatin score in GC B cells relative to Naïve B cells. We have thus investigated the interplay between genetic alterations, gene expression, and chromatin state of B-Cell lymphomas and their normal cells of origin, and introduced analytical methods that can be applied to other cancer types.

NOVEL IDEAS FOR DETECTING EPIGENETIC VARIATION IN MULTIPLE HUMAN CELL TYPES

Yu Zhang[1], Marta Byrska-Bishop[2], Feng Yue[2], Ross C Hardison[2]

[1]The Pennsylvania State University, Statistics, University Park, PA, [2]The Pennsylvania State University, Biochemistry and Molecular Biology, University Park, PA

Genome segmentation is a powerful tool for detecting and characterizing regulatory elements, particularly for non-coding regions that critically contribute to gene expression and phenotypes. With massive amount of epigenetic data sets generated in many cell types, we are provided with powerful means to study the dynamics of epigenomes by contrasting signals across cell types. Current methods are however inadequate for detecting epigenetic variation across many samples, because they ignore position specificity of regulatory activities. We present a Bayesian method that jointly models position specific epigenetic states and detects differential regulatory regions in many cell types in different distance scales. Using ENCODE data sets in six cell types, our method achieved superior accuracy and robustness for predicting regulatory elements than existing genome segmentation tools. We marked 13.9% of the human genome being variable in terms of regulatory patterns in at least one cell type. The variable sites are strongly associated with differential gene expression and are significantly enriched of genetic variants associated with complex phenotypes that are relevant to the specific cell types. The identified differential regulatory regions yielded stronger enrichment of disease variants than current methods using cell-type specific epigenetic states. Overall, we present a general framework for annotating regulatory elements and detecting epigenetic variation in many samples. Our results have potentially better implications on the connections between regulatory activities and diseases.

GENOMIC PREDICTION FROM WHOLE GENOME SEQUENCE DATA IN CATTLE

Ben J Hayes[1,2], Iona M Macleod[2,3], Aurelien Capitan[4], Hans D Daetwyler[1,2], Phil J Bowman[1], Mike E Goddard[1,3], Amanda J Chamberlain[1]

[1]Department of Primary Industries, Computational Biology, Melbourne, Australia, [2]Latrobe University, Science, Melbourne, Australia, [3]University of Melbourne, Veterinary and Agricultural Science, Melbourne, Australia, [4]INRA, Genetique Animale et Biologie Integrative, Jouy-en-Josas, France

Discoveries from whole genome sequencing large numbers of cattle have been rapidly translated into tests that cattle breeders use. An example is a lethal recessive mutation in the SMC2 gene in Holstein-Friesian cattle that causes embryonic loss – this mutation is now routinely screened on very large numbers of cattle, so that carrier matings are avoided. A more ambitious goal is to use whole genome sequence variant genotypes for prediction of complex traits. Genomic prediction is widely used in dairy cattle breeding programs around the world, where selection of young animals for breeding is based on genomic predictions of milk yield, fertility, mastitis resistance, longevity and temperament. While this has certainly increased genetic gain for these traits, the fact that these predictions are based on 50,000 SNP genotypes could limit the accuracy and persistency of these predictions. Using whole genome sequence data could overcome these limitations, as the causative mutations underlying the variation in the complex traits should be in the data set. We describe a new method for genomic prediction that simultaneously maps causative mutations and estimates variant effects. The value of the method is demonstrated on a dairy cattle data set of 16,000 animals with imputed whole genome sequence data, and temperament and milk production phenotypes. In some cases plausible mutations affecting these traits were identified. This information can be rapidly applied in dairy cattle breeding, by including these mutations on low density SNP arrays that are now used to test 100s of thousands of animals each year.

OPEN CHROMATIN REVEALS THE FUNCTIONAL PORTION OF THE MAIZE GENOME

Eli Rodgers-Melnick[1], Peter J Bradbury[1,2], Daniel L Vera[3], Hank W Bass[3], Edward S Buckler[1,2]

[1]Cornell University, Institute for Genomic Diversity, Ithaca, NY, [2]United States Department of Agriculture, Agricultural Research Service, Ithaca, NY, [3]Florida State University, Department of Biological Science, Tallahassee, FL

Chromatin accessibility is a highly informative feature of the eukaryotic genome. As recent human ENCODE results demonstrate, assays of open chromatin using DNaseI hypersensitivity can be used to pinpoint diverse sets of cis-regulatory elements. The discovery of such putative functional regions in crop species has the potential to illuminate the genetic architecture of quantitative traits, as recent data strongly suggests much of the underlying genetic variation resides in the regulatory, non-genic regions of the genome. In this study we use an MNase hypersensitivity assay to discover open chromatin regions within the the maize reference genome, B73. These open chromatin regions compose approximately 0.6% of the total genome, but they appear strongly tied to the functional variation important for plant breeding. We show that recombination hotspots within the maize genome correspond to up to 4-fold enrichments of open chromatin within nearly all sequence contexts. We also demonstrate that open chromatin is enriched by 2-fold in and around variants explaining diverse quantitative traits and that this enrichment persists both within gene-proximal and gene-distal regions. Finally, using a variance partitioning approach, we find that putative open chromatin regions explain 30% of the phenotypic variance - half the heritable variation - of diverse quantitative traits, with the remainder of the heritable variation explained by the 2% of the genome containing coding sequence. Together, these results suggest that assays of chromatin accessibility will be at least as useful as the transcriptome in defining the functional portions of crop genomes.

A 3D TISSUE CULTURE PLATFORM TO INVESTIGATE DRUG RESISTANCE IN MULTIPLE MYELOMA

Theodorus E de Groot, Erwin Berthier, Ashleigh B Theberge, David J Beebe

University of Madison - Wisconsin, Biomedical Engineering, Madison, WI

Multiple myeloma (MM) is the second most common bone cancer, it treatable but not curable. The impact of targeted treatment is limited by the widespread genomic heterogeneity as rare, resistant phenotypes emerge from therapy. The effectiveness of treatment is further reduced by the specific microenvironment of the bone marrow inducing a complex network of interactions, providing protection from therapy. The significance the interactions of MM with the microenvironment has been realized in recent years and is the target of several new and developing drugs. Due to the complexity of the microenvironment, current methods for predicting appropriate therapies by either genotyping or in vitro tissue culture are not well suited.

We have developed an in vitro model to investigate patient-specific interactions between MM, its microenvironment, and treatments. The model enables compartmentalized co-culture of MM cells and stromal cells in of a 3D in vitro model that replicates the MM microenvironment. The model reproduces salient aspects of the bone marrow environment, including a mineralized bone-like scaffold cultured with bone marrow stromal cells (BMSC) in proximity to a MM spheroid. A specialized hanging drop platform allows for manipulation of the soluble and physical microenvironment of the culture. This approach enables simple access for media changes, sampling, and additions without disturbing the culture. The system also enables multiple compartments of the microenvironment such as the bone marrow and tumor to be prepared separately, combined, then removed. In these experiments, the bone marrow tissue is prepared in the bone-like scaffold while MM is cultured in spheroid in the droplet. Manipulation of multiple 3D tissue compartments enables control and tracking and partitioning cells based on responses to stimuli. The approach also has the advantage of concentrating cells within the droplets, allowing for a tissues to be cultured with minimal cells, allowing for many conditions to be experiments with from starting material, like patient samples.

To investigate the role of the microenvironment in drug resistance, stromal cells are isolated from bone aspirate of MM patients. The stroma is cocultured with a multiple myeloma cell line, by keeping the MM cells constant, the role of the microenvironment in resistance to therapy can be studied through genetic and expression data to predict patient outcome.

This technology creates a unique opportunity to perform tunable in vitro studies of human tumor and bone cells in a bio-mimetic platform closely approximating the in vivo MM microenvironment. The ability of this platform to reproduce the complex microenvironmental interactions occurring the bone marrow while allowing simple access to each individual component allows more omics-based analyses and elucidate pathways and mechanisms in MM drug resistance.

LARGE-SCALE *IN VIVO* ENHANCER DELETION WITH CRISPR/CAS9

Diane E Dickel[1], Yiwen Zhu[1], Marco Osterwalder[1], Brandon Mannion[1], Veena Afzal[1], Ingrid Plajzer-Frick[1], Alan Fang[1], Catherine Pickle[1], Jennifer A Akiyama[1], Edward M Rubin[1,2], Axel Visel[1,2], Len A Pennacchio[1,2]

[1]Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, CA, [2]US Department of Energy Joint Genome Institute, Walnut Creek, CA

Enhancers are *cis*-regulatory sequences that are thought to play an important role in a spectrum of developmental and disease processes. Community efforts to characterize the noncoding functions embedded in the human genome have led to the identification of hundreds of thousands of sites that have features associated with enhancers. However, the requirement for the vast majority of these putative enhancer sequences to establish proper gene expression levels remains unexplored, and beyond a few anecdotal examples, the impact of their loss or mutation on developmental and disease processes is unknown. To examine the *in vivo* function of mammalian enhancers more systematically, we used CRISPR/Cas9 genome editing to engineer more than 25 mouse lines, each missing at least one developmental enhancer. This panel of deleted enhancers with tissue-specific activity in the developing heart, limb, forebrain, or face was selected from the Vista Enhancer Browser (enhancer.lbl.gov), a collection of enhancers whose *in vivo* activities were previously established in transgenic mouse assays. We perform RNA-seq on enhancer deletion mice to measure both *cis* and *trans* gene expression changes, and we assess potential morphological and disease phenotypes. As with gene knockouts, the effects of enhancer loss range from severe morphological or disease phenotypes to more subtle gene expression changes to cases of no discernable phenotype. For example, we identified enhancers active in the heart where loss of a single enhancer leads to large expression changes of neighboring cardiac myosin genes. Adult mice that are homozygous null for any one of these enhancer sequences are viable but display characteristics of heart failure. At the other end of the spectrum, individual loss of some enhancers leads to no overt gene expression or phenotypic changes, suggesting functional redundancy among enhancers regulating the same gene. To examine this possibility, some of these sites are now being deleted in combination with neighboring enhancers that have overlapping activity patterns. Our results underscore the potential for noncoding genomic sequences to harbor disease-causing mutations. The adoption of whole genome sequencing technologies for human disease studies results in a pressing need for continued functional characterization of noncoding sequences to distinguish deleterious regulatory mutations from neutral changes. Our data highlight the utility of large-scale *in vivo* CRISPR/Cas9-mediated genome editing in mice as a powerful tool for the exploration of such noncoding sequence changes.

HIGH RESOLUTION SIZE PROFILING OF PLASMA DNA: BIOLOGY AND CLINICAL APPLICATIONS

Yuk Ming Dennis Lo

The Chinese University of Hong Kong, Li Ka Shing Institute of Health Sciences, Hong Kong, China

There is much recent interest in the use of plasma DNA for noninvasive prenatal testing (NIPT), cancer liquid biopsy and transplantation monitoring. We have conducted a high resolution study of the size profile of plasma DNA in pregnant women using genomewide paired-end massively parallel sequencing. We have shown that the size profile shows a nucleosomal patterns, with circulating fetal DNA being shorter than the maternal background. Based on this information, we have developed a new method for noninvasive prenatal testing based on the size profiling of DNA molecules derived from different chromosomes. We have observed that pregnancies involving trisomy 21, trisomy 18 and trisomy 13 fetuses would have a shortening of the size profile of plasma DNA derived from the respective chromosomes. This size-based molecular diagnostic approach is synergistic with the count-based approach that is currently in clinical use and can be expected to further enhance the robustness of NIPT. We have also applied a similar approach in measuring the size profile of plasma DNA in patients with hepatocellular carcinoma. We compared the size profile of plasma DNA molecules derived from chromosomal regions that are amplified (relatively enriched in tumor-derived DNA) versus those that are deleted (relatively depleted with tumor-derived DNA). Such analysis shows that circulating tumor DNA is shorter than the background non-tumor-derived DNA. Such information could potentially be incorporated into new detection strategies for conducting liquid biopsies for cancer detection and monitoring.

MATERNAL AGE EFFECT AND SEVERE GERMLINE BOTTLENECK IN THE INHERITANCE OF HUMAN MITOCHONDRIAL DNA

Boris Rebolledo-Jaramillo*[1], Marcia Shu-Wei Su*[2], Nicholas Stoler[1], Jennifer A McElhoe[3], Benjamin Dickins[4], Daniel Blankenberg[1], Thorfinn Korneliussen[5], Francesca Chiaromonte[6], Rasmus Nielsen[7], Mitchell M Holland[3], Ian M Paul[8], Anton Nekrutenko**[1], Kateryna D Makova**[2]

[1]Penn State University, Biochemistry and Molecular Biology, University Park, PA, [2]Penn State University, Biology, University Park, PA, [3]Penn State University, Forensic Science Program, University Park, PA, [4]Nottingham Trent University, School of Science and Technology, Nottingham, United Kingdom, [5]University of Copenhagen, Centre for GeoGenetics, Copenhagen, Denmark, [6]Penn State University, Statistics, University Park, PA, [7]University of California, Integrative Biology, Berkeley, CA, [8]Penn State University, Pediatrics, College of Medicine, Hershey, PA

Most mitochondrial diseases show intraindividual mitochondrial DNA (mtDNA) variation (i.e., heteroplasmy). Unfortunately, transmission of deleterious alleles cannot be readily predicted due to the lack of data on the size of the mtDNA bottleneck during oogenesis, a process that may abruptly transform a benign (low frequency) variant in a woman into a disease-causing (high frequency) variant in her child.

Here we present a high-resolution study of heteroplasmy transmission conducted on blood and buccal mtDNA of 39 healthy mother-child pairs of European ancestry (156 samples, each sequenced at ~20,000x/site). With 99% power to detect heteroplasmies over sequencing error, we identified more than 150 heteroplasmic sites with minor allele frequency (MAF) ≥ 1%. On average, each individual carried one heteroplasmy, and one in eight individuals carried a disease-associated heteroplasmy with minor allele frequency ≥1%. We observed frequent drastic heteroplasmy frequency shifts between generations and estimated the effective size of the germline mtDNA bottleneck at only ~29-35. Strikingly, we found a positive association between the number of heteroplasmies in a child and maternal age at fertilization, likely attributable to oocyte aging. This study takes advantage of droplet digital PCR (ddPCR) to validate heteroplasmies and confirm a *de novo* mutation.

Our results can be utilized in predicting the transmission of disease-causing mtDNA variants. In particular, given the fact that the United Kingdom is moving forward towards allowing triparental assisted fertilization involving mtDNA, our approach can provide actionable information for the selection of mtDNA donors.

(*) Equal contribution
(**) Co-corresponding authors

RNA-SEQ ANALYSIS OF PLACENTAL TRANSCRIPTIONAL LANDSCAPE IN NORMAL AND COMPLICATED PREGNANCIES

Siim Sõber[1], Mario Reiman[1], Triin Kikas[1], Kristiina Rull[1], Rain Inno[1], Pille Vaas[2], Pille Teesalu[2], Jesus M Lopez Marti[3], Pirkko Mattila[3], Maris Laan[1]

[1]University of Tartu, Institute of Molecular and Cell Biology, Tartu, Estonia, [2]University of Tartu, Department of Obstetrics and Gynaecology, Tartu, Estonia, [3]Institute for Molecular Medicine Finland, Sequencing Unit, Helsinki, Finland

Approximately one in five pregnancies suffer from complications of varying severity, and prenatal conditions are increasingly recognized as major determinants in infant, adolescent and recently also adult health. We performed an extensive study of the placental transcriptome using RNA sequencing, the most powerful and contemporary method for gene expression analysis. Our samples (n=40) cover placentas representing normal gestation (n=8), preeclampsia (n=8), gestational diabetes (n=8), as well as aberrations in fetal growth defined as small- (n=8) and large-for-gestational-age (n=8) newborns. The generated dataset is the largest of its kind generated to date and covers the widest range of pregnancy outcomes. Sequencing yielded ~3.4 Billion 46 + 46 bp paired end reads (Mean: 84.4 M reads per sample; range: 48.4M – 145.2M). Of the total 164 billion sequenced bases 121 billion (73.5%) were successfully mapped to the human genome after quality control, filtering and alignment steps (49.2% mapped to known mRNAs).We observed a clear distinction in the severity of transcriptional disturbances in preeclampsia (215 differentially expressed genes) compared to other pregnancy complications which only exhibit differential expression in a few genes. A limited number of transcription factors including LRF, SP1 and AP2 are implicated as drivers of these changes. Our results also provide support to the hypothesis of shared placental responses to pregnancy complications as we observe substantial overlap of gene expression alterations in different pregnancy complications. Our data provide a rich resource for identification of potential biomarkers and therapeutic targets involved in gestational disturbances.

THE GENETIC REGULATORY LANDSCAPE OF THE HUMAN PANCREATIC ISLET TRANSCRIPTOME

Ana Viñuela[1,2,3], Martijn van de Bunt[4,5], Nikolay Oskolkov[6], Cédric Howald[1,2,3], João Fadista[6], Nikolaos Nikolaos[1,2,3], Petter Strom[6], Patrick E MacDonald[7], Anna L Gloyn[4,5], Leif Groop[6], Mark McCarthy[4,5,8], Emmanouil T Dermitzakis[1,2,3]

[1]University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva, Switzerland, [2]Swiss Institute of Bioinformatics, Geneva, Switzerland, [3]Institute of Genetics and Genomics in Geneva, Geneva, Switzerland, [4]Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, United Kingdom, [5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [6]Lund University Diabetes Centre, Department of Clinical Sciences, Malmö, Sweden, [7]Alberta Diabetes Institute, University of Alberta, Edmonton, Canada, [8]NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Understanding the molecular properties of GWAS signals underlying common disorders requires the study of relevant cell types and tissues. In type 2 diabetes (T2D), pancreatic islets of Langerhans are key in the pathogenesis of the disease and therefore investigating the genetics of gene regulation in this tissue is essential to understanding if and how the disease develops in certain individuals. However, obtaining enough samples to answer these questions has been a challenge. Here, we are overcoming this problem by combining RNAseq datasets of pancreatic islets generated by different laboratories in, to our knowledge, the largest dataset available from a mixed set of cases and controls for T2D. We are combining pancreatic islet expression with genotypic information from 286 individuals to uncover genetic regulatory variants (eQTL) relevant for the tissue. In addition, by combining phenotypic information, we are looking for disease specific genetic regulatory elements, including those manifesting as allelic imbalance in expression and splicing changes. Furthermore, using purified beta cells from 25 of the individuals we are characterizing the cell-specific transcription regulation in the context of the disease. In conclusion, our study aims to improve the current catalogue of genetic variants affecting gene expression in pancreatic islets; and moves us towards a mechanistic understanding of the pathology of T2D.

NOTES

NOTES

NOTES

NOTES

NOTES

TECHNICAL ABSTRACTS
FOR WORKSHOPS

ILLUMINA WORKSHOP
AT BIOLOGY OF GENOMES MEETING 2015

When: Wednesday, May 6, 2015

Time: 12:30 pm – 2:00 pm EST

Topics:

Illumina Portfolio Update: The Ever Expanding Genomics Tool Box

Carri-Lyn Mead, Ph.D.
Illumina, Inc.
Senior Manager, Technology Enablement

Abstract: The growing range of Illumina sequencing platforms and capabilities has been instrumental in transforming the use of sequence information in a wide range of genomic, genetic and biological studies. Here we present new developments in Illumina's sequencing portfolio, including an overview of new platforms, technical innovations, and recent data.

Building a Better Medical Genome

Michael A. Eberle, Ph.D.
Illumina, Inc.
Associate Director, Scientific Research

As sequencing moves into clinical applications, it is important that we assess and improve the accuracy of variant calls made by standard informatics pipelines and, where needed, develop targeted informatics pipelines to call medically relevant variants. To this end, Illumina recently developed a high quality truth data set of variant calls based on whole genome sequencing of the parents and eleven children from the CEPH/Utah pedigree 1463. We will describe how we improved the sensitivity of the Isaac SNP and indel calls by 1.5% and 35%, respectively with effectively no reduction in precision. We will also discuss a new algorithm that can detect pathological repeat expansions in one or more specified repeat regions in a given sample using paired 100mer reads from whole genome sequence data.

Register for the workshop at: **www.illumina.com/cshlbog2015**

Power your next big breakthrough.

Sequencing power for every scale.

NEW
HiSeq X™ Series

Population power
Maximum throughput and low cost population- and production-scale human whole-genome sequencing. Series includes the HiSeq X Ten and the new HiSeq X Five Systems.

NEW
HiSeq® Series

Production power
Maximum throughput and lowest cost for production-scale genomics. Series includes the new HiSeq 3000 and HiSeq 4000 Systems.

NEW
NextSeq® Series

Flexible power
Desktop speed and simplicity for everyday genomics. Series includes the NextSeq 500 and the new NextSeq 550 with cytogenomic array scanning.

MiSeq® Series

Focused power
Speed and simplicity for targeted and small-genome sequencing. Series includes the MiSeq and MiSeqDx™ Systems.*

Compare our new systems with our product selector at www.illumina.com/power.

FOR RESEARCH USE ONLY
*MiSeqDx™ is a 510(k) cleared, CE-marked instrument. See instructions for use.
©2015 Illumina, Inc. All rights reserved.

illumina®

Reduce Your NGS Sequencing Costs with Improved Library Complexity

Accel-NGS™ 2S DNA Library Kit for Illumina® Platforms

- Broad input range: 10 pg to 1 µg
- PCR-free libraries from 10 ng
- Superior library complexity
- Efficient adapter technology

Swift™ BIOSCIENCES

Soar Above. Discover More.

© 2015, Swift Bioscienes, Inc. The Swift logo and Accel-NGS are trademarks of Swift Biosciences. Illumina is a trademark of Illumina, Inc. 15-0188, 03/15

Accel-NGS™

www.swiftbiosci.com

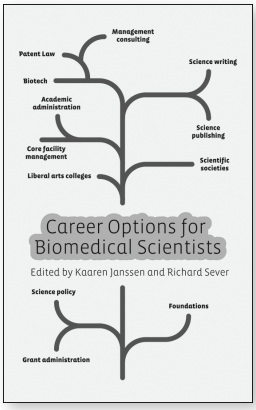
bioRxiv

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*

Participant List

Dr. illumina

Ms. Akweley Ablorh
Harvard T.H. Chan School of Public Health
aablorh@post.harvard.edu

Mr. Robert Aboukhalil
Cold Spring Harbor Laboratory
raboukha@cshl.edu

Dr. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Dr. Francois Aguet
Broad Institute
francois@broadinstitute.org

Dr. Derek Aguiar
Princeton University
daguiar@princeton.edu

Dr. Vitor Aguiar
University of São Paulo
vitor.aguiar@me.com

Dr. Keiko Akagi
Ohio State University
keiko.akagi@osumc.edu

Dr. Frank Albert
University of California, Los Angeles
falbert@mednet.ucla.edu

Dr. Jessica Alfoldi
Broad Institute of MIT and Harvard
jalfoldi@broadinstitute.org

Dr. Carl Anderson
Wellcome Trust Sanger Institute
ca3@sanger.ac.uk

Ms. Andrea Anderson
GenomeWeb
anderson@genomeweb.com

Dr. Aida Andres
Max Planck Institute for Evolutionary Anthropology
aida.andres@eva.mpg.de

Dr. Simeon Andrews
WCMC-Q
ssa2007@qatar-med.cornell.edu

Prof. Stylianos Antonarakis
University of Geneva
stylianos.antonarakis@unige.ch

Ms. Michaela Asp
SciLifeLab KTH
michaela.asp@scilifelab.se

Dr. Adam Auton
Albert Einstein College of Medicine
adam.auton@einstein.yu.edu

Ms. Bracha Avigdor (Erlanger)
The Johns Hopkins SOM
bracha@jhmi.edu

Dr. Maria Avila-Arcos
Stanford University
maricugh@gmail.com

Dr. Sharon Aviran
UC Davis
saviran@ucdavis.edu

David Aylor
North Carolina State University
dlaylor@ncsu.edu

Dr. Julien Ayroles
Princeton
ayroles@gmail.com

Dr. Paul Babb
Perelman School of Medicine, U. Penn
pbabb@mail.med.upenn.edu

Dr. Doris Bachtrog
UC Berkeley
dbachtrog@berkeley.edu

Dr. Soheil Baharian
McGill University
soheil.baharian@mcgill.ca

Dr. Orli Bahcall
Nature Reviews Genetics / Nature Genetics
o.bahcall@us.nature.com

Dr. Christopher Baker
The Jackson Laboratory
christopher.baker@jax.org

Dr. Michaela Banck
Mayo Clinic Rochester
michaela.banck@gmail.com

Dr. Antonio Barbadilla
University Autonoma Barcelona
antonio.barbadilla@uab.es

Dr. Luis Barreiro
University of Montreal
luis.barreiro@umontreal.ca

Dr. Jeffrey Barrett
Wellcome Trust Sanger Institute
barrett@sanger.ac.uk

Dr. Elizabeth Bartom
Northwestern University
ebartom@northwestern.edu

Mr. Dan Bar-Yaacov
Ben Gurion University of the Negev
danbary@post.bgu.ac.il

Ms. Mitra Barzine
EMBL
mitra@ebi.ac.uk

Dr. Timour Baslan
Memorial Sloan Kettering Cancer Center
baslant@mskcc.org

Dr. Chiara Batini
University of Leicester
cb334@le.ac.uk

Dr. Alexis Battle
Johns Hopkins University
ajbattle@cs.jhu.edu

Dr. Philippe Batut
CSHL
batut@cshl.edu

Dr. Serafim Batzoglou
Stanford
serafim@cs.stanford.edu

Mr. Kenneth Baughman
OIST
kbaughman@oist.jp

Dr. Tyler Beck
NHGRI, NIH
tyler.beck@nih.gov

Ms. Camille Berthelot
EMBL-EBI
cberthel@ebi.ac.uk

Dr. Johanna Bertl
Aarhus University
johanna.bertl@clin.au.dk

Dr. Andreas Beutler
Mayo Clinic
beutler.andreas@mayo.edu

Dr. Ewan Birney
EBI/EMBL
birney@ebi.ac.uk

Ms. Francis Blokzijl
Hubrecht Institute
f.blokzijl@hubrecht.eu

Dr. Alistair Boettiger
Harvard University
boettiger@fas.harvard.edu

Mr. Craig Bohrson
Johns Hopkins Medical School
craigbohrson@gmail.com

Dr. Kirsten Bomblies
Harvard University
kbomblies@oeb.harvard.edu

Dr. Mark Borowsky
NIBR
mark.borowsky@novartis.com

Mr. Lukasz Boryn
Research Institute of Molecular Pathology
lukasz.boryn@imp.ac.at

Dr. Laura Botigué
Stony Brook
laura.botigue@stonybrook.edu

Dr. Adam Boyko
Cornell University
boyko@cornell.edu

Dr. Alan Boyle
University of Michigan
apboyle@umich.edu

Ms. Debora Brandt
University of Sao Paulo
deboraycb@gmail.com

Dr. Nicolas Bray
University of California, Berkeley
nicolas.bray@gmail.com

Prof Michael Brent
Washington University
brent@wustl.edu

Ms. Alessandra Breschi
Centre for Genomic Regulation (CRG)
alessandra.breschi@crg.eu

Dr. Daniel Brewer
The Genome Analysis Centre / UEA
daniel.brewer@tgac.ac.uk

Dr. Lawrence Brody
National Human Genome Research Institute/NIH
lbrody@mail.nih.gov

Dr. Lisa Brooks
NHGRI
lisa.brooks@nih.gov

Dr. Christopher Brown
University of Pennsylvania
casey6r0wn@gmail.com

Ms. Sherry-Ann Brown
Mayo Clinic
brown.sherryann@mayo.edu

Dr. Andrew Brown
University of Geneva
andrew.brown@unige.ch

Dr. Katarzyna Bryc
23andMe
kbryc@23andme.com

Dr. Alfonso Buil
University of Geneva
alfonso.buil@unige.ch

Dr. Robert Bukowski
Cornell University
bukowski@cornell.edu

Dr. David Burt
Roslin Institute & R(D)SVS University of Edinburgh
dave.burt@roslin.ed.ac.uk

Dr. Wolfgang Busch
Gregor Mendel Institute
wolfgang.busch@gmi.oeaw.ac.at

Ms. Andrea Byrnes
Massachusetts General Hospital
abyrnes@broadinstitute.org

Dr. Mario Caceres
Universitat Autonoma de Barcelona
mcaceres@icrea.cat

Mr. Alex Cagan
Max Planck Institute for Evolutionary Anthropology
alexander_cagan@eva.mpg.de

Ms. Na Cai
University of Oxford
caina@well.ox.ac.uk

Dr. Eric Campbell
MGH/Harvard

Mr. Christopher Campbell
Duke University
c.ryan.campbell@duke.edu

Dr. Matthew Cannon
Cleveland Clinic Lerner Research Institute
cannonm3@ccf.org

Dr. Han Cao
BioNano Genomics
Han@bionanogenomics.com

Mr. Francesco Nicola Carelli
University Lausanne
francesconicola.carelli@unil.ch

Mr. Jedidiah Carlson
University of Michigan
jedidiah@umich.edu

Dr. Shai Carmi
Columbia University
scarmi@cs.columbia.edu

Dr. Piero Carninci
RIKEN
carninci@riken.jp

Mr. Francesco Paolo Casale
European Molecular Biology Laboratory
casale@ebi.ac.uk

Dr. Sònia Casillas
Universitat Autònoma de Barcelona
sonia.casillas@uab.cat

Ms. Sylvia Chang
Thermo Fisher Scientific
sylvia.chang@lifetech.com

Dr. Carole Charlier
University of Liège
carole.charlier@ulg.ac.be

Dr. Barbara Cheifet
Genome Biology
barbara.cheifet@genomebiology.com

Dr. Shann-Ching Chen
Thermo Fisher Scientific
Shann-Ching.Chen@thermofisher.com

Mr. Yuping Chen
Stony Brook University
yuping.chen@stonybrook.edu

Dr. Lei Chen
The Jackson Laboratory
lchen@jax.org

Dr. Michael Cherry
Stanford University
cherry@stanford.edu

Dr. Alessandra Chesi
CHOP
chesia@email.chop.edu

Mr. Colby Chiang
Washington University in St Louis
cc2qe@virginia.edu

Dr. Stirling Churchman
Harvard Medical School
churchman@genetics.med.harvard.edu

Dr. Michele Clamp
Harvard University
mclamp@g.harvard.edu

Dr. Andrew Clark
Cornell University
ac347@cornell.edu

Ms. Laura Clarke
EMBL-EBI
laura@ebi.ac.uk

Dr. Melina Claussnitzer
Harvard Medical School
melina@broadinstitute.org

Dr. Julie Collens
Illumina
jcollens@illumina.com

Dr. Francis Collins
National Institutes of Health
francis.collins@nih.gov

Dr. Don Conrad
Washington University School of Medicine
dconrad@genetics.wustl.edu

Dr. Montserrat Corominas
Universitat de Barcelona
mcorominas@ub.edu

Dr. Chris Cotsapas
Yale School of Medicine
cotsapas@broadinstitute.org

Dr. Ester Cuenca
Broad Institute-Call D'Hebron Research Institute
ecuenca@broadinstitute.org

Ms. Ciara Curtin
GenomeWeb
ccurtin@genomeweb.com

Dr. Darren Cusanovich
University of Washington
cusanovi@uw.edu

Ms. Ioana Cutcutache
Duke-NUS Graduate Medical School, Singapore
ioana.cutcutache@duke-nus.edu.sg

Dr. Agata Czyz
Illumina
aczyz@illumina.com

Dr. Matteo D'Antonio
University of California, San Diego
madantonio@ucsd.edu

Mr. Matthew Dapas
Northwestern University
matthew.dapas@northwestern.edu

Mr. Gregory Darnell
Princeton University
gdarnell@princeton.edu

Dr. Robert Darnell
New York Genome Center
jabusuttil@nygenome.org

Dr. Jyotishka Datta
Duke University
jd298@stat.duke.edu

Dr. George Davey Smith
University of Bristol
KZ.Davey-Smith@bristol.ac.uk

Mr. Robert Davies
University of Oxford
rwdavies@well.ox.ac.uk

Dr. Brian Davis
NHGRI / NIH
davisbw@mail.nih.gov

Mr. Joe Davis
Stanford University
joed3@stanford.edu

Dr. Aaron Day-Williams
Biogen Idec
aaron.day-williams@biogenidec.com

Dr. Marianne De Gorter
Stanford University
mdegorte@stanford.edu

Mr. Theodorus de Groot
University of Wisconsin - Madison
tedegroot@wisc.edu

Dr. Simone de Jong
King's College London
sdejongwork@gmail.com

Dr. Francisco De La Vega
Annai Systems, Inc.
delavefm@gmail.com

Dr. Ines de Santiago
University of Cambridge
ines.desantiago@cruk.cam.ac.uk

Mr. Brennan Decker
NIH-NHGRI
brennan.decker@nih.gov

Dr. Pieter deJong
CHORI
pdejong@chori.org

Dr. Olivier Delaneau
University of Geneva
olivier.delaneau@gmail.com

Dr. Laura DeMare
Genome Research/Molecular Case Studies
ldemare@cshl.edu

Dr. Scott Devine
Unversity of Maryland School of Medicine
sdevine@som.umaryland.edu

Ms. Valentina Di Francesco
NIH/NHGRI
vdifrancesco@mail.nih.gov

Dr. Federica Di Palma
Broad Institute/The Genome Analysis Center, UK
federica.di-palma@tgac.ac.uk

Dr. Anna Di Rienzo
University of Chicago
dirienzo@bsd.uchicago.edu

Ms. Tonya Di Sera
University of Utah
tony.disera@gmail.com

Dr. Diane Dickel
Lawrence Berkeley National Laboratory
dedickel@lbl.gov

Dr. John Didion
NIH
john.didion@nih.gov

Dr. Caroline Dive
Cancer Research UK Manchester Institute
caroline.dive@manchester.ac.uk

Dr. Alexander Dobin
CSHL
dobin@cshl.edu

Dr. Elisa Docampo
GIGA-ULg
elisa.docampo@ulg.ac.be

Dr. Drew Doering
University of Wisconsin-Madison
dtdoering@wisc.edu

Prof. Peter Donnelly
University of Oxford
directorpa@well.ox.ac.uk

Dr. Noah Dowell
HHMI- Univ. of Wisconsin at Madison
ndowell@wisc.edu

Mr. Jason Downing
Illumina
jdowning@illumina.com

Dr. Tom Druet
University of Liège
tom.druet@ulg.ac.be

Dr. Richard Durbin
The Wellcome Trust Sanger Inst.
rd@sanger.ac.uk

Dr. Keith Durkin
University of Liege
kdurkin@ulg.ac.be

Mr. Marwan El Khoury
University of Leicester
mek12@le.ac.uk

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Dr. Lel Eory
Roslin Institute, University of Edinburgh
Lel.Eory@roslin.ed.ac.uk

Dr. Michael Erdos
NHGRI
mikee@mail.nih.gov

Dr. Laurent Excoffier
University of Berne
laurent.excoffier@iee.unibe.ch

Ms. Maud Fagny
Institut Pasteur
maud.fagny@pasteur.fr

Dr. Fabiana Farias
University of Misouri
fabiana.farias@imbim.uu.se

Dr. Andrew Farrell
University of Utah
farrelac@bc.edu

Dr. Suzanne Fei
Oregon Health & Science University
feis@ohsu.edu

Dr. Elise Feingold
National Institutes of Health
feingold@nih.gov

Dr. Adam Felsenfeld
NIH/NHGRI
adam_felsenfeld@nih.gov

Dr. Juan Fernandez
Wellcome Trust Centre for Human Genetics
jfertaj@well.ox.ac.uk

Dr. Paul Flicek
EMBL-EBI
flicek@ebi.ac.uk

Dr. Jonathan Flint
University of Oxford - Wellcome Trust Centre for Human Genetics
jf@well.ox.ac.uk

Dr. Liliana Florea
Johns Hopkins University
florea@jhu.edu

Dr. Christine Fosker (nee Bird)
The Genome Analysis Centre
Christine.fosker@tgac.ac.uk

Mr. Gustavo Franca
University of Sao Paulo
gsfranca@gmail.com

Mr. Christopher Frank
Duke University
chris.frank@duke.edu

Dr. Andy Fraser
University of Toronto
andyfraser.utoronto@gmail.com

Dr. Kelly Frazer
UC San Diego
kafrazer@ucsd.edu

Dr. Rosa Fregel
Stanford University
rfregel@stanford.edu

Dr. Menachem Fromer
Icahn School of Medicine at Mount Sinai
menachem.fromer@mssm.edu

Dr. Audrey Fu
University of Chicago
audreyqyfu@gmail.com

Dr. Qiaomei Fu
Harvard Medical School
qiaomeifu@gmail.com

Dr. Daniel Gaffney
The Wellcome Trust Sanger Inst.
dg13@sanger.ac.uk

Dr. Julien Gagneur
Gene Center, LMU
gagneur@genzentrum.lmu.de

Dr. Irene Gallego Romero
University of Chicago
ireneg@uchicago.edu

Dr. Chuan Gao
Duke University
cg148@duke.edu

Dr. Manuel Garber
University of Massachusetts Medical School
manuel.garber@umassmed.edu

Dr. Eugene Gardner
University of Maryland School of Medicine
egardner@umaryland.edu

Mr. Kiran Garimella
University of Oxford
kiran@well.ox.ac.uk

Dr. Michel Georges
ULg
michel.georges@ulg.ac.be

Dr. Mark Gerstein
Yale University
piaa@gersteinlab.org

Dr. Stefania Giacomello
SciLifeLab
stefania.giacomello@scilifelab.se

Dr. Evgenia Giannopoulou
City University of New York
egiannopoulou@citytech.cuny.edu

Dr. Giuliana Giannuzzi
University of Lausanne
giuliana.giannuzzi@unil.ch

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Dr. David Gifford
MIT
dkg@mit.edu

Dr. Chris Gignoux
Stanford University
cgignoux@stanford.edu

Dr. Daniel Gilchrist
NHGRI
Daniel.Gilchrist@nih.gov

Ms. Carla Giner-Delgado
Universitat Autonoma de Barcelona
Carla.Giner@uab.cat

Dr. Thomas Gingeras
Cold Spring Harbor Laboratory
gingeras@cshl.edu

Prof. Donna Gitter
Baruch College, City University of New York
Donna.Gitter@baruch.cuny.edu

Mr. Craig Glastonbury
King's College London
craig.glastonbury@kcl.ac.uk

Dr. Gernot Gloeckner
University of Cologne
gernot.gloeckner@uni-koeln.de

Dr. Jonathan Goeke
Genome Institute of Singapore
gokej@gis.a-star.edu.sg

Dr. Omer Gokcumen
University at Buffalo
gokcumen@gmail.com

Dr. Angela Goncalves
Wellcome Trust Sanger Institute
ag14@sanger.ac.uk

Dr. Bettie Graham
National Institutes of Health
bettie_graham@nih.gov

Dr. Simon Gravel
McGill University
simon.gravel@mcGill.ca

Prof. Richard Green
University of California, Santa Cruz
ed@soe.ucsc.edu

Dr. Yongtao Guan
Baylor College of Medicine
yongtaog@bcm.edu

Dr. Rodrigo Gularte Mérida
GIGA -- Research
rodrigo.gularte@ulg.ac.be

Mr. James Gurtowski
Cold Spring Harbor Laboratory
gurtowsk@cshl.edu

Dr. Ira Hall
Washington University
ihall@genome.wustl.edu

Ms. Benika Hall
UNC Charlotte
bjohn157@uncc.edu

Dr. Pille Hallast
University of Leicester
ph116@le.ac.uk

Ms. Diana Han
Li Ka Shing Institute of Health Sciences
dianahansc@gmail.com

Mr. Bob Handsaker
Harvard Medical School
handsake@broadinstitute.org

Dr. Susan Harbison
NIH/NHLBI
susan.harbison@nih.gov

Dr. Manoj Hariharan
Salk Institute for Biological Studies
mhariharan@salk.edu

Dr. Tim Harkins
Swift Biosciences
harkins@swiftbiosci.com

Mr. Chad Harland
University of Liège
charland@ulg.ac.be

Dr. Jennifer Harrow
Wellcome Trust Sanger Institute
jla1@sanger.ac.uk

Dr. Christopher Hart
Isis Pharmaceuticals
chart@isisph.com

Dr. Shinichi Hashimoto
Kanazawa University
hashimoto@med.kanazawa-u.ac.jp

Dr. Alex Hastie
BioNano Genomics
ahastie@bionanogenomics.com

Mr. Jim Havrilla
University of Utah
semjaavria@gmail.com

Dr. Ben Hayes
Department of Environment
ben.hayes@ecodev.vic.gov.au

Mr. Yupeng He
Salk Institute for Biological Studies
yuhe@salk.edu

Dr. Brenna Henn
SUNY Stony Brook
brenna.henn@stonybrook.edu

Dr. Javier Herrero
UCL Cancer Institute
javier.herrero@ucl.ac.uk

Dr. Axel Himmelbach
IPK Gatersleben
himmelba@ipk-gatersleben.de

Ms. Angie Hinrichs
UC Santa Cruz
angie@soe.ucsc.edu

Dr. Yu-Jui Ho
Cold Spring Harbor Laboratory
yjho@cshl.edu

Dr. Margret Hoehe
Max Planck Institute for Molecular Genetics
hoehe@molgen.mpg.de

Dr. Eurie Hong
Stanford University
euriehong@stanford.edu

Dr. Daniel Howrigan
Massachusetts General Hospital
daniel.howrigan@gmail.com

Prof. Lusheng Huang
Jiangxi Agricultural University
lushenghuang@hotmail.com

Dr. Sanwen Huang
Chinese Academy of Agricultural Sciences
tianjuan@genomics.cn

Ms. Miriam Huntley
Baylor College of Medicine
mhuntley@alum.mit.edu

Dr. Daniel Hupalo
New York University
dh123@nyu.edu

Mr. Matthew Hymes
Swift Biosciences
hymes@swiftbiosci.com

Dr. Lilia Iakoucheva
University of California San Diego
lilyak@ucsd.edu

Dr. Hae Kyung Im
The University of Chicago
haky@uchicago.edu

Dr. Carlos Infante
University of Georgia
cinfante@uga.edu

Dr. John Irish

Dr. Koichi Itoh
The Institute for Theoretical Molecular Biology
koichiitoh@yahoo.co.jp

Dr. Vishy Iyer
University of Texas at Austin
vishy@utexas.edu

Dr. David Jaffe
The Broad Institute
jaffe@broadinstitute.org

Ms. Myrthe Jager
Hubrecht Institute
m.jager@hubrecht.eu

Mr. Jacob Jensen
Aarhus University
jmj@birc.au.dk

Ms. Shan (Mandy) Jiang
University of California, Irvine
jiangs2@uci.edu

Dr. Ying Jin
Cold Spring Harbor Laboratory
yjin@cshl.edu

Dr. Vijai Joseph
Memorial Sloan-Kettering Cancer Center
josephv@mskcc.org

Dr. Goo Jun
University of Texas Health Science Ctr Houston
Goo.Jun@uth.tmc.edu

Dr. Andre Kahles
Memorial Sloan Kettering Cancer Center
akahles@cbio.mskcc.org

Ms. Cynthia Kalita
Wayne State University
cakalita@gmail.com

Dr. Konrad Karczewski
Massachusetts General Hospital
konradk@broadinstitute.org

Dr. Elinor Karlsson
Univ Massachusetts Medical School
elinor@broadinstitute.org

Mr. Tomas Kazmar
Research Institute of Molecular Pathology
tomas.kazmar@imp.ac.at

Dr. Manolis Kellis
MIT / Broad Institute
kellis-admin@mit.edu

Mr. Thomas Kelly
BioNano Genomics
tkelly@bionanogenomics.com

Dr. Janet Kelso
Max Planck Institute for Evolutionary Anthropology
kelso@eva.mpg.de

Dr. Eimear Kenny
Icahn School of Medicine at Mount Sinai
eimear.kenny@mssm.edu

Dr. Ekta Khurana
Weill Cornell Medical College
ekk2003@med.cornell.edu

Dr. Helena Kilpinen
European Molecular Biology Laboratory
kilpinen@ebi.ac.uk

Dr. Seok-Won Kim
RIKEN Center for Integrative Medical Sciences
seokwon.kim@riken.jp

Dr. Daehwan Kim
Johns Hopkins University
infphilo@gmail.com

Dr. Dokyoon Kim
Pennsylvania State University
duk27@psu.edu

Dr. Martin Kircher
University of Washington
mkircher@uw.edu

Dr. Anthony Kirilusha
National Institutes of Health
akirilus@caltech.edu

Dr. Paul Kitts
NIH/NLM/NCBI
kitts@ncbi.nlm.nih.gov

Dr. David Knowles
Stanford University
dak33@stanford.edu

Dr. Kord Kober
University of California San Francisco
kord.kober@ucsf.edu

Dr. Sarah Kocher
Princeton
skocher@gmail.com

Mr. Nils Kölling
EMBL-EBI
nk@ebi.ac.uk

Dr. Jonas Korlach
Pacific Biosciences
jkorlach@pacificbiosciences.com

Mr. Joshua Korn
Novartis Institutes for Biomedical Research
joshkorn@gmail.com

Dr. Karl Kremling
Cornell University
kak268@cornell.edu

Mr. Sam Krerowicz
UW-Madison
skrerowicz@chem.wisc.edu

Ms. Anna Kropornicka
UW-Madison
kropornicka@wisc.edu

Ms. Kimberly Kukurba
Stanford University
kkukurba@stanford.edu

Mr. Runjun Kumar
Washington University in St. Louis
kumarr@wusm.wustl.edu

Dr. Deniz Kural
Seven Bridges Genomics
deniz.kural@sbgenomics.com

Mr. Jason Lajoie
UW-Madison
jason.lajoie87@gmail.com

Mr. Ernest Lam
BioNano Genomics
elam@bionanogenomics.com

Dr. Eric Lander
The Broad Institute of MIT & Harvard
lander@broadinstitute.org

Dr. Tuuli Lappalainen
New York Genome Center & Columbia University
tlappalainen@nygenome.org

Dr. David Lawrie
University of Southern California
dlawrie@usc.edu

Mr. Ryan Layer
University of Virginia
rl6sf@virginia.edu

Amanda Lea
Duke University
amanda.lea@duke.edu

Mr. Hyung Joo Lee
Washington University School of Medicine
hyungjoo.lee@wustl.edu

Mr. Dillon Lee
University of Utah
dlee123@gmail.com

Dr. Wan-Ping Lee
Seven Bridges Genomics
wanping.lee@sbgenomics.com

Dr. Andreas Leha
Wellcome Trust Sanger Institute
al22@sanger.ac.uk

Dr. Susanna Lemmelä
Finnish Institute of Occupational Health
susanna.lemmela@ttl.fi

Mr. Liron Levin
Ben Gurion University of the Negev
levinl@post.bgu.ac.il

Dr. Ruowang Li
Penn State
rvl5032@psu.edu

Dr. Yang Li
Stanford University
yangili@stanford.edu

Dr. Jiani Li
Baylor College of Medicine
jianil@bcm.edu

Dr. Jun Li
University of Michigan
junzli@med.umich.edu

Dr. Yanzhu Lin
NIH/NHLBI
yanzhu.lin@nih.gov

Dr. Stephen Lincoln
Invitae
steve.lincoln@me.com

Dr. Z. Lewis Liu
USDA-ARS-NCAUR
Zlewis.liu@ars.usda.gov

Dr. Dajiang Liu
Penn State College of Medicine
dajiang.liu@psu.edu

Dr. Dennis Lo
The Chinese University of Hong Kong
loym@cuhk.edu.hk

Ms. Rebecca Lowdon
Washington University in St. Louis
rebecca.lowdon@wustl.edu

Dr. David Lowry
Michigan State University
davidbryantlowry@gmail.com

Dr. Fei Lu
Cornell Unviersity
fl262@cornell.edu

Dr. Francesca Luca
Wayne State University
fluca@wayne.edu

Dr. Jie Luo
Zhejiang Academy of Agricultural Sciences
luojie@mail.zaas.ac.cn

Mr. Yi-Jyun Luo
Okinawa Institute of Science and Technology
locke.yj@gmail.com

Dr. Gholson Lyon
Cold Spring Harbor Laboratory
gholsonjlyon@gmail.com

Dr. Daniel MacArthur
Massachusetts General Hospital
macarthur@atgu.mgh.harvard.edu

Dr. Iain MacCallum
Broad Institute
iainm@broadinstitute.org

Mr. Matthew MacGilvray
UW-Madison
macgilvray@wisc.edu

Ms. Sorina Maciuca
University of Oxford
sorina.maciuca@well.ox.ac.uk

Mr. Sho Maekawa
The University of Tokyo
smaekawa@hgc.jp

Dr. Thomas Mailund
Aarhus University
mailund@birc.au.dk

Dr. Robert Majovski
Broad Institute
rmajovsk@broadinstitute.org

Dr. Vladimir Makarov
Swift Biosciences
makarov@swiftbiosci.com

Dr. Kateryna Makova
Penn State University
kdm16@psu.edu

Dr. Joel Andrew Malek
Weill Cornell Medical College in Qatar
jom2042@qatar-med.cornell.edu

Dr. Allison Mandich
National Institutes of Health
mandicha@mail.nih.gov

Dr. Allison Mandich
National Institutes of Health
mandicha@mail.nih.gov

Dr. Elaine Mardis
Washington University School of Medicine
emardis@wustl.edu

Dr. Elliott Margulies
Illumina
emargulies@illumina.com

Dr. Rob Mariman
Ulg
rob.mariman@ulg.ac.be

Dr. John Marioni
EMBL
marioni@ebi.ac.uk

Dr. Gabor Marth
University of Utah School of Medicine
gmarth@genetics.utah.edu

Ms. Hilary Martin
University of Oxford
hilary.martin@well.ox.ac.uk

Mr. Martin Mascher
IPK Gatersleben
mascher@ipk-gatersleben.de

Dr. Iain Mathieson
Harvard Medical School
iain_mathieson@hms.harvard.edu

Dr. Richard McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Dr. Shannon McCurdy
UC Berkeley
smccurdy@berkeley.edu

Mr. Tarang Mehta
The Genome Analysis Centre (TGAC)
Tarang.Mehta@tgac.ac.uk

Dr. Marta Melé
CRG / Harvard University
marta.mele.messeguer@gmail.com

Dr. Pall Melsted
University of Iceland
pmelsted@hi.is

Ms. Hannah Meyer
EMBL-EBI
hannah@ebi.ac.uk

Prof. Diogo Meyer
University of Sao Paulo
diogo@ib.usp.br

Dr. Michael Michalkiewicz
Aurora Health Care
michael.michalkiewicz@aurora.org

Dr. Franziska Michor
Harvard School of Public Health/Dana Farber Cancer Institute
michor@jimmy.harvard.edu

Mr. Chase Miller
Boston College
chmille4@gmail.com

Dr. Ryan Mills
University of Michigan
remills@umich.edu

Mr. Benjamin Minkoff
University of Wisconsin-Madison
bminkoff@wisc.edu

Dr. Dan Mishmar
Ben-Gurion University of the Negev
dmishmar@bgu.ac.il

Dr. Tejaswini Mishra
Stanford University School of Medicine
tejaswini.mishra@stanford.edu

Ms. Andrea Moffitt
Duke University
andrea.moffitt@duke.edu

Dr. Stephen Montgomery
Stanford University
smontgom@stanford.edu

Dr. Shinichi Morishita
University of Tokyo
moris@cb.k.u-tokyo.ac.jp

Dr. Ali Mortazavi
University of California Irvine
ali.mortazavi@uci.edu

Dr. Sara Mostafavi
University of British Columbia
saram@stat.ubc.ca

Mr. Yuichi Motai
The University of Tokyo
motights@yahoo.co.jp

Dr. Simon Moxon
The Genome Analysis Centre
business.support@tgac.ac.uk

Mr. Gregory Moyerbrailean
Wayne State University
gmoyerbr@wayne.edu

Dr. Jonathan Mudge
Wellcome Trust Sanger Institute
jm12@sanger.ac.uk

Mr. Felix Muerdter
Research Institute of Molecular Pathology
felix.muerdter@imp.ac.at

Dr. Jim Mullikin
NHGRI/NIH
mullikin@mail.nih.gov

Dr. Kasper Munch
Aarhus University
kaspermunch@birc.au.dk

Ms. Shaila Musharoff
Stanford University
shailam@stanford.edu

Dr. Ramaiah Nagaraja
National Institute on Aging
nagarajar@mail.nih.gov

Dr. Ryo Nakaki
The University of Tokyo
nakaki@genome.rcast.u-tokyo.ac.jp

Dr. Maria Nattestad
Cold Spring Harbor Laboratory
mnattest@cshl.edu

Dr. Arcadi Navarro
University Pompeu Fabra
arcadi.navarro@upf.edu

Dr. Jeff Nelson
Morgridge Institute for Research
nelson@morgridge.org

Mr. Dominic Nelson
McGill University
dominic.nelson@mail.mcgill.ca

Dr. Mark Nolte
University of Wisconsin-Madison
mjnolte@wisc.edu

Dr. Kelly Nunes
University of Sao Paulo
knunes_bio@yahoo.com.br

Dr. Jared O'Connell
Illumina
joconnell@illumina.com

Dr. Daniel O'Connor
Wellcome Trust

Ms. Anna Okula
Pennsylvania State University
azo121@psu.edu

Dr. Hanna Ollila
Stanford University School of Medicine
hannao@stanford.edu

Dr. Hiroaki Onda
The Jackson Laboratory
hiroaki.onda@jax.org

Dr. Dayna Oschwald
New York Genome Center
doschwald@gmail.com

Dr. Stephan Ossowski
Center for Genomic Regulation
stephan.ossowski@crg.eu

Mr. Omead Ostadan
Illumina, Inc.
oostadan@illumina.com

Dr. Elaine Ostrander
National Human Genome Research Institute/NIH
eostrand@mail.nih.gov

Ms. Natasha Pacheco
University of Alabama at Birmingham
npacheco@uab.edu

Dr. Athma Pai
Massachusetts Institute of Technology
athma@mit.edu

Dr. Aarno Palotie
PNGU/CHGR
aarno.palotie@helsinki.fi

Dr. Andy Wing Chun Pang
BioNano Genomics
apang@bionanogenomics.com

Dr. Ji Yeon Park
Seoul National University
parkji7@snu.ac.kr

Dr. Heidi Parker
NHGRI/National Institutes of Health
hgparker@mail.nih.gov

Dr. Stephen Parker
University of Michigan
scjp@umich.edu

Dr. Leopold Parts
EMBL
sabine.blum@embl.de

Dr. Justin Paschall
EMBL-EBI
paschall@ebi.ac.uk

Dr. Ashwini Patil
University of Tokyo
ashwini@hgc.jp

Dr. Michael Pazin
NHGRI, NIH
pazinm@mail.nih.gov

Dr. Nathan Pearson
New York Genome Center
npearson@nygenome.org

Dr. Sarah Pendergrass
Geisinger Health Research
sap29@psu.edu

Dr. Sarah Perry
Nature Biotechnology
sarah.perry@nature.com

Dr. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Dr. Hedi Peterson
University of Tartu
hedi.peterson@ut.ee

Dr. Galina Petukhova
Uniformed Services University of the Health Sciences
galina.petukhova@usuhs.edu

Dr. Hemali Phatnani
New York Genome Center
hphatnani@nygenome.org

Dr. Joseph Pickrell
New York Genome Center
jkpickrell@nygenome.org

Mr. Harold Pimentel
UC Berkeley
hpimentel@berkeley.edu

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Dr. Maria Polychronidou
Molecular Systems Biology
maria.polychronidou@embo.org

Prof. Christopher Ponting
University of Oxford
chris.ponting@dpag.ox.ac.uk

Dr. Timothy Powell
King's College London
timothy.powell@kcl.ac.uk

Dr. Joseph Powell
University of Queensland
joseph.powell@uq.edu.au

Dr. Hadass Pri Chen

Dr. Alkes Price
Harvard School of Public Health
aprice@hsph.harvard.edu

Dr. Jonathan Pritchard
Stanford University
ttrim@stanford.edu

Dr. Molly Przeworski
Columbia University
molly.przew@gmail.com

Dr. Marta Puig
Universitat Autonoma de Barcelona
marta.puig@uab.cat

Dr. Yi Qiao
University of Utah
yi.qiao@genetics.utah.edu

Dr. Pengfei Qin
Max Planck Institute for Evolutionary Anthropology
pengf.qin@gmail.com

Dr. Wei Qu
University of Tokyo
quwei00@gmail.com

Prof. Francis Quetier
University of Evry/GIP Genopole
francis.quetier@gmail.com

Dr. Raquel Rabionet
Center for Genomic Regulation
rut.carbonell@crg.eu

Dr. Towfique Raj
Brigham and Women's Hospital
traj@rics.bwh.harvard.edu

Dr. Anil Raj
Stanford University
rajanil@stanford.edu

Ms. Deepthi Rajagopalan
Duke University
deepthi.rajagopalan@duke.edu

Dr. Sohini Ramachandran
Brown University
sramachandran@brown.edu

Dr. Vijay Ramani
University of Washington
vramani@uw.edu

Mr. Gokul Ramaswami
Stanford University
gokulr@stanford.edu

Dr. Joshua Randall
Wellcome Trust Sanger Institute
joshua.randall@sanger.ac.uk

Dr. Gunnar Ratsch
Memorial Sloan-Kettering Cancer Center
raetsch@cbio.mskcc.org

Mr. Boris Rebolledo-Jaramillo
Pennsylvania State University
berebolledo@gmail.com

Dr. Anupama Reddy
Duke University
anupama.reddy@duke.edu

Dr. Caroline Relton
University of Bristol
caroline.relton@bristol.ac.uk

Prof. Jun Ren
Jiangxi Agricultural University
renjunjxau@hotmail.com

Ms. Leonor Rib
University of Lausanne
leonor.rib@unil.ch

Dr. Stephen Richards
Baylor College of Medicine

Dr. Samuli Ripatti
FIMM Institute for Molecular Medicine Finland
samuli.ripatti@helsinki.fi

Dr. Marylyn Ritchie
Pennsylvania State University
marylyn.ritchie@psu.edu

Dr. Roland Roberts
PLOS Biology
rroberts@plos.org

Dr. Nicolas Robine
New York Genome Center
nrobine@nygenome.org

Ms. Nicole Rockweiler
Washington University in St. Louis
nrockweiler@wustl.edu

Dr. Dan Roden
Vanderbilt University School of Medicine
dan.roden@vanderbilt.edu

Dr. Eli Rodgers-Melnick
Cornell University
er432@cornell.edu

Dr. Laura Rodriguez
National Human Genome Research Institute
rodrigla@mail.nih.gov

Dr. Jeffrey Rogers
Baylor College of Medicine
jr13@bcm.edu

Dr. Mostafa Ronaghi
Illumina, Inc.
mronaghi@illumina.com

Dr. Frederick Roth
University of Toronto
fritz.roth@utoronto.ca

Dr. Maxime Rotival
Institut Pasteur
mrotival@pasteur.fr

Mr. Andrei Rozanski
Hospital Sirio-Libanes
rozanski.andrei@gmail.com

Mr. Konrad Rudolph
EMBL-EBI
konrad.rudolph@ebi.ac.uk

Dr. Anna Rychkova
Stanford University
rychkova@stanford.edu

Dr. Tina Saey
Science News
tina.hesman@gmail.com

Dr. William Salerno
Baylor College of Medicine
William.Salerno@bcm.edu

Mr. Fredrik Salmén
SciLifeLab/KTH
fredrik.salmen@scilifelab.se

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Mr. Didac Santesmases
Centre for Genomic Regulation (CRG)
didac.santesmasses@crg.eu

Dr. Stephen Schaffner
Broad Institute
sfs@broadinstitute.org

Dr. Michael Schatz
Cold Spring Harbor Laboratory
mschatz@cshl.edu

Dr. Steven Scherer
Baylor College of Medicine
sscherer@bcm.edu

Dr. David Schlessinger
National Institute on Aging
schlessingerd@mail.nih.gov

Dr. Daniel Schlingman
Illumina
dschlingman@illumina.com

Ms. Ellen Schmidt
University of Michigan
schellen@umich.edu

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Dr. Korbinian Schneeberger
Max Planck Institute for Plant Breeding Research
schneeberger@mpipz.mpg.de

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@ncbi.nlm.nih.gov

Ms. Molly Schumer
Princeton University
schumer@princeton.edu

Prof. David Schwartz
University of Wisconsin - Madison
dcschwartz@wisc.edu

Ms. Renee Sears
Washington University in St. Louis
r.sears@wustl.edu

Dr. Jonathan Sebat
UC San Diego
jsebat@ucsd.edu

Dr. Fritz Sedlazeck
Simons Center for Quantitative Biology, Cold Sprin
fsedlaze@chsl.edu

Dr. Laura Shannon
Cornell
lms395@cornell.edu

Dr. Andrew Sharp
Icahn School of Medicine at Mount Sinai
andrew.sharp@mssm.edu

Dr. Yufeng Shen
Columbia University
ys2411@columbia.edu

Dr. Xinghua Shi
University of North Carolina at Charlotte
x.shi@uncc.edu

Dr. Atsushi Shimizu
Iwate Medical University
ashimizu@iwate-med.ac.jp

Dr. Suyash Shringarpure
Stanford University
suyashs@stanford.edu

Dr. Arend Sidow
Stanford University
arend@stanford.edu

Dr. Petr Simecek
The Jackson Laboratory
petrs@jax.org

Dr. Sonal Singhal
Columbia University
sonal.singhal1@gmail.com

Ms. Alyza Skaist
Johns Hopkins University
alyza@jhu.edu

Dr. Magdalena Skipper
Nature
m.skipper@nature.com

Mr. Laurits Skov
Aarhus University
lauritsskov2@gmail.com

Dr. Kerrin Small
Kings College London
kerrin.small@kcl.ac.uk

Dr. Michael Snyder
Stanford University School of Medicine
mpsnyder@stanford.edu

Dr. Siim Sõber
University of Tartu
siims@ut.ee

Dr. Heidi Sofia
NIH/NHGRI/DGM
heidi.sofia@nih.gov

Mr. Li Song
Johns Hopkins University
lsong10@jhu.edu

Dr. Nicole Soranzo
Wellcome Trust Sanger Institute
ns6@sanger.ac.uk

Mr. Dylan Spalding
EMBL-EBI
spalding@ebi.ac.uk

Dr. Noah Spies
Stanford University/NIST
nspies@stanford.edu

Dr. Frank Steemers
Illumina, Inc
fsteemers@illumina.com

Dr. Oliver Stegle
EMBL European Bioinformatics Institute
oliver.stegle@ebi.ac.uk

Dr. Nils Stein
IPK Gatersleben
stein@ipk-gatersleben.de

Dr. Derek Stemple
Wellcome Trust Sanger Institute
ds4@sanger.ac.uk

Mr. Nicholas Stoler
Penn State University
nbs128@psu.edu

Dr. Barbara Stranger
University of Chicago
bstranger@uchicago.edu

Mr. Georg Stricker
Gene Center Munich
stricker@genzentrum.lmu.de

Ms. Lakshmi Subramanian
Indian Institute of Technology Madras
lakshmi1386@gmail.com

Dr. Luo Sun
New England Biolabs
tonello@neb.com

Dr. Qi Sun
Cornell University
qisun@cornell.edu

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Mr. Yuta Suzuki
The University of Tokyo
ysuzuki@cb.k.u-tokyo.ac.jp

Dr. Yutaka Suzuki
University of Tokyo
ysuzuki@hgc.jp

Mr. Yohihiko Suzuki
The University Tokyo
ys.neoteny@gmail.com

Dr. David Swarbreck
The Genome Analysis Centre
business.support@tgac.ac.uk

Dr. David Symer
Ohio State University Comp Cancer Ctr
david.symer@osumc.edu

Dr. Haruko Takeda
University of Liège
htakeda@ulg.ac.be

Dr. Michael Talkowski
Massachusetts General Hospital
talkowski@chgr.mgh.harvard.edu

Dr. Oliver Tam
Cold Spring Harbor Laboratory
tam@cshl.edu

Mr. Leland Taylor
National Institutes of Health
leland.taylor@nih.gov

Dr. Todd Taylor
RIKEN
taylor@riken.jp

Mr. João Teixeira
Max Planck Institute for Evolutionary Anthropology
joao_teixeira@eva.mpg.de

Ms. Natalie Telis
Stanford University
ntelis@stanford.edu

Dr. Jim Thomas
NIH
jthomas@nhgri.nih.gov

Dr. James Thomas
NIH
thomasjw4@mail.nih.gov

Dr. Barbara Thomas
National Institutes of Health
bthomas@csr.nih.gov

Dr. Hagen Tilgner
Stanford University
htilgner@stanford.edu

Dr. Marta Tomaszkiewicz
Penn State University
marta@bx.psu.edu

Mr. Shingo Tomioka
The University of Tokyo
s.tomioka.103@gmail.com

Dr. Richard Trembath
Queen Mary University of London
vp-health@qmul.ac.uk

Dr. Gosia Trynka
Wellcome Trust Sanger Institute
gosia@sanger.ac.uk

Dr. Aristotelis Tsirigos
NYU School of Medicine
aristotelis.tsirigos@nyumc.org

Dr. Gene Tsvid
U Wisconsin-Madison
ytsvid@wisc.edu

Dr. Taru Tukiainen
Massachusetts General Hospital
taru@atgu.mgh.harvard.edu

Dr. Tom Tullius
Boston University
tullius@bu.edu

Dr. Jenny Tung
Duke University
jt5@duke.edu

Dr. Igor Ulitsky
Weizmann Institute
igor.ulitsky@weizmann.ac.il

Dr. Catalina Vallejos Meneses
EMBL-EBI
catalina@ebi.ac.uk

Dr. Anton Valouev
University of Southern California
valouev@usc.edu

Dr. Anne Van den Broeke
Jules Bordet Institute - ULB
anne.vandenbroeke@bordet.be

Dr. Francesco Vezzi
SciLifeLab -- National Genomics Infrastructure
francesco.vezzi@scilifelab.se

Dr. Alain Vignal
INRA
alain.vignal@toulouse.inra.fr

Dr. Ana Viñuela
University of Geneva
ana.vinuela@unige.ch

Dr. Benjamin Voight
The University of Pennsylvania
bvoight@upenn.edu

Dr. Manja Wachsmuth
Max-Planck-Institute for Evolutionary Anthropology
manja_wachsmuth@eva.mpg.de

Prof. Claes Wadelius
Uppsala University
Claes.Wadelius@igp.uu.se

Mr. Florian Wagner
Duke University
florian.wagner@duke.edu

Dr. Per Wahlberg
Uppsala University
per.wahlberg@medsci.uu.se

Dr. Klaudia Walter
Wellcome Trust Sanger Institute
kw8@sanger.ac.uk

Dr. Xiufeng(Henry) Wan
MIssissippi State University
wan@cvm.msstate.edu

Dr. Xiaoyue Wang
CAMS, Peking Union Medical College
pumcwangxy@163.com

Dr. Ting Wang
Washington University
twang@genetics.wustl.edu

Dr. Minghui Wang
Cornell University Bioinformatics
mw729@cornell.edu

Dr. Jinhua Wang
NYU School of Medicine
jinhua.wang@nyumc.org

Dr. Bo Wang
Cold Spring Harbor Lab
bwang@cshl.edu

Dr. Alistair Ward
University of Utah
AlistairNWard@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory/USDA ARS
ware@cshl.edu

Dr. Stephen Watt
Wellcome Trust Sanger Institute
sbw@sanger.ac.uk

Dr. Matthew Webster
Uppsala University
matthew.webster@imbim.uu.se

Dr. Sharon Wei
CSHL
weix@cshl.edu

Dr. Wu Wei
Stanford University
wuwei5@stanford.edu

Dr. Zhiping Weng
University of Massachusetts Medical School
zhiping.weng@umassmed.edu

Prof. Lorenz Wernish
Medical Research Council
lorenz.wernisch@mrc-bsu.cam.ac.uk

Dr. Sarah Wheelan
The Johns Hopkins University School of Medicine
swheelan@jhmi.edu

Dr. John Willis
Duke University
jwillis@duke.edu

Dr. Richard Wilson
The Genome Institute
rwilson@wustl.edu

Dr. David Winter
Arizona State University
david.winter@gmail.com

Dr. Genevieve Wojcik
Stanford University
gwojcik@stanford.edu

Ms. Brooke Wolford
National Human Genome Research Institute
brooke.wolford@nih.gov

Dr. Emily Wong
University of Southern California
emilyhwo@usc.edu

Dr. Kim Worley
Baylor College of Medicine
kworley@bcm.edu

Dr. Steven Wu
Arizona State University
stevenwu@asu.edu

Mr. Hualin Xi
Pfizer
hualin.xi@pfizer.com

Dr. Chunlin Xiao
NIH/NLM/NCBI
xiao2@ncbi.nlm.nih.gov

Mr. Tao Yang
Pennsylvania State University
txy146@psu.edu

Ms. Lu Yang
Princeton University
luy@princeton.edu

Dr. Huanming Yang
BGI
yanghm@genomics.org.cn

Dr. Haiwang Yang
NIH
haiwang.yang@nih.gov

Mr. Alexander Young
University of Oxford
ay@well.ox.ac.uk

Dr. Bing Yu
University of Texas Health Science at Houston
Bing.Yu@uth.tmc.edu

Dr. Feng Yue
Pennsylvania State University
fyue@hmc.psu.edu

Dr. Nataliya Yutin
NCBI NLM NIH
yutinn@ncbi.nlm.nih.gov

Dr. Laura Zahn
AAAS
lzahn@aaas.org

Dr. Judith Zaugg
EMBL
judith.zaugg@embl.de

Mr. Daniel Zerbino
EMBL-EBI
kwalsh@ebi.ac.uk

Dr. Xiuqing Zhang
BGI-Shenzhen
zhangxq@genomics.cn

Dr. Bo Zhang
Washington University in St. Louis
bzhang29@wustl.edu

Dr. Yu Zhang
The Pennsylvania State University
yzz2@psu.edu

Ms. Jenny Zhang
Duke University
jenny.zhang@DUKE.EDU

Dr. Dongyan Zhao
Michigan State University
zhaodon4@msu.edu

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# 62275** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail only
Lower level: Word processing 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
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes
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 62275 (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**)
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**)
Executive Limo	631-696-8000 (**1047**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
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