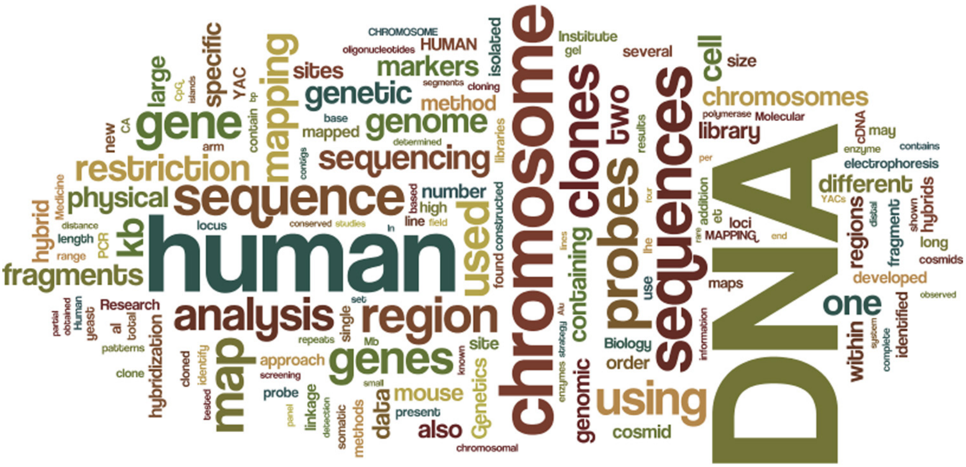
Abstracts of papers presented
at the 2017 meeting on

THE BIOLOGY OF GENOMES

May 9–May 13, 2017

1988



2017

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2017 meeting on

THE BIOLOGY OF GENOMES

May 9–May 13, 2017

Arranged by

Michel Georges, *University of Liege, Belgium*
Matthew Hurles, *Wellcome Trust Sanger Institute*
Dana Pe'er, *Sloan Kettering Institute*
Jonathan Pritchard, *Stanford University*

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

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

Corporate Benefactor

Regeneron

Corporate Sponsors

Agilent Technologies
Bristol-Myers Squibb Company
Calico Labs
Celgene
Genentech, Inc.
Thermo Fisher Scientific
Merck
Monsanto Company
New England BioLabs
Pfizer

Corporate Partners

Alexandria Real Estate
Gilead Sciences
Novartis Institutes for Biomedical Research
Sanofi

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: Three decades of genome science at Cold Spring Harbor. Word clouds from the full text of abstracts presented at the 1988 *Genome Mapping and Sequencing* and 2017 *Biology of Genomes* meetings.

Word clouds generated by David Stewart using Wordle™

THE BIOLOGY OF GENOMES

Tuesday, May 9 – Saturday, May 13, 2017

Tuesday	7:30 pm	**1** Cancer and Medical Genomics
Tuesday	*following eve. session*	*Happy Hour sponsored by Illumina*
Wednesday	9:00 am	**2** Complex Traits and Microbiome
Wednesday	2:00 pm	**3** Poster Session I
Wednesday	4:30 pm	*Wine and Cheese Party**
Wednesday	7:30 pm	**4** Evolutionary and Non-human Genomics
Thursday	9:00 am	**5** Functional Genomics
Thursday	2:00 pm	**6** Poster Session II
Thursday	4:30 pm	**7** ELSI Panel and Discussion
Thursday	7:30 pm	**8** Translational Genetics and Genomics
Thursday	*following eve. session*	*Happy Hour sponsored by Swift Biosciences*
Friday	9:00 am	**9** Population 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** Computational Genomics

Workshops (immediately following morning sessions)
Illumina, Wednesday
Oxford Nanopore, Thursday

* *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

Printed on 100% recycled paper.

PROGRAM

TUESDAY, May 9—7:30 PM

SESSION 1 CANCER AND MEDICAL GENOMICS

Chairperson: **Elaine Mardis,** Nationwide Children's Hospital, Columbus, Ohio
Christina Curtis, Stanford University, California

Genomic characterization of breast cancer progression
Elaine R. Mardis, Christopher A. Miller, Marni Siegel, Katherine Hoadley, Jeremy Hoog, Sherri Davies, Lisa Carey, Joel Parker, Charles Perou, Matthew J. Ellis.
Presenter affiliation: Nationwide Children's Hospital, Columbus, Ohio. 1

A pan cancer analysis of promoter activity highlights the regulatory role of alternative transcription start sites and their association with noncoding mutations
Deniz Demircioglu, Tannistha Nandi, Engin Cukuroglu, Claudia Calabrese, Nuno Fonseca, Andre Kahles, Kjong Lehmann, Steve Rozen, Bin Tean Teh, Oliver Stegle, Alvis Brazma, Angela Brooks, Gunnar Raetsch, Patrick Tan, Jonathan Goeke.
Presenter affiliation: Genome Institute of Singapore, Singapore. 2

Expanding discovery from cancer genomes by integrating network analyses with massively parallel *in vivo* tumorigenesis assays
Heiko Horn, Michael Lawrence, Candace Chouinard, Yashaswi Shresta, Jessica Hu, Elizabeth Worstell, Emily Shea, Nina Ilic, Ejung Kim, Atanas Kamburov, Alireza Kashani, William Hahn, Joshua Campbell, Jesse Boehm, Gad Getz, Kasper Lage.
Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 3

The genomic autopsy—Using whole exome and whole genome sequencing to solve complex fetal and neonatal presentations
Alicia B. Byrne, Jinghua Feng, Andreas W. Schreiber, Peter J. Brautigan, Nathalie J. Nataren, Sui Yu, Yee Khong, Lynette Moore, Marcel E. Dinger, Christopher N. Hahn, Karin S. Kassahn, Christopher P. Barnett, Hamish S. Scott.
Presenter affiliation: Centre for Cancer Biology, Adelaide, Australia. 4

Happy Hour
Sponsored by **Illumina**

WEDNESDAY, May 10—9:00 AM

SESSION 2 COMPLEX TRAITS AND MICROBIOME

Chairperson: **Benjamin Neale,** Massachusetts General Hospital, Boston
Moran Yassour, Broad Institute of MIT and Harvard, Cambridge, Massachusetts

WEDNESDAY, May 10—2:00 PM

SESSION 3 POSTER SESSION I

WEDNESDAY, MAY 10—4:30 PM

Wine and Cheese Party

WEDNESDAY, May 10—7:30 PM

SESSION 4 EVOLUTIONARY AND NON-HUMAN GENOMICS

Chairperson: **Karine van Doninck,** Université de Namur, Belgium
Nitin Phadnis, University of Utah, Salt Lake City

THURSDAY, May 11—9:00 AM

SESSION 5 FUNCTIONAL GENOMICS

Chairperson: **Howard Chang,** Stanford University School of Medicine, California
Barbara Treutlein, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

THURSDAY, May 11—2:00 PM

SESSION 6 POSTER SESSION II

THURSDAY, May 11—4:30 PM

SESSION 7 ELSI PANEL DISCUSSION

What's in a Name? Diversity and the Future of Genomic Research

Moderator: **Dave Kaufman, Ph.D.,** NIH/National Human Genome Research Institute

Panelists

Eimar Kenny, Icahn School of Medicine at Mt. Sinai
Sandra Lee, Stanford University
Alondra Nelson, Columbia University
Aliya Saperstein, Stanford University

The lack of representation of individuals from understudied and ancestral minority populations in genomic research, and in most gene/variant databases, GWAS catalogs and variant databases is now widely recognized. When members of racial and ethnic minority groups are not adequately included in genomic research, important information about disease biology can be missed. This may limit the ability of researchers to confirm the broader relevance of population-specific findings and to investigate heterogeneous disease risks across populations that might be attributable to gene-gene or gene-environment interactions.

More research focused on people from under-studied populations must be conducted. As genomics addresses that challenge we should try to attempt define with sufficient precision and rigor the populations whose data are needed, we should understand the impact of the definitions we choose on our research findings, and we should consider how the interpretation of those findings may impact the populations being studied. An array of approaches to classify or assess the diversity of those represented in our studies and mutation databases are currently used, including self-identified race and ethnic group (which are often based on imprecise U.S. Census categories), grandparental geographical origins, geocoding based on participants' home and workplace, and ancestral informative genetic markers.

How membership in understudied populations is defined will influence the utility of future genomics databases. It will have implications for our power to do subgroup analyses and assess a broad range of gene-environment interactions. The choice may influence the sensitivity of our databases to interpret and classify variants as genetic causes of disease across populations. It will have implications for our power to do subgroup analyses, to assess a broad range of gene-environment interactions, and to disentangle social and genetic determinants of health.

This panel includes individuals with expertise in population genetics, demographics, ethics and the sociology of genomics. We will discuss the scientific and social importance of addressing disparities in the genomic databases; the implications of using different criteria to define and recruit populations; and rigorous methods to address the problem of inadequate inclusion of minority populations.

THURSDAY, May 11—7:30 PM

SESSION 8 TRANSLATIONAL GENETICS AND GENOMICS

Chairperson: **Mathew Garnett,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
Jay Shendure, University of Washington, Seattle

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

Genome-wide CRISPR/Cas9 knockout screening and poly-genomic interrogation of primary leukemia cells identify novel mechanisms of glucocorticoid resistance in pediatric B-lineage ALL
Robert J. Autry, Steven W. Paugh, Joseph R. McCorkle, Calvin E. Lau, Erik J. Bonten, Jordan A. Beard, Kristine R. Crews, Wenjian Yang, Cheng Cheng, Deqing Pei, Seth E. Karol, Kathryn G. Roberts, Stanley Pounds, Charles G. Mullighan, Sima Jeha, Ching-hon Pui, Mary V. Relling, William E. Evans.
Presenter affiliation: St. Jude Children's Research Hospital (SJCRH), Memphis, Tennessee; The University of Tennessee Health Science Center, Memphis, Tennessee. 232

Dissecting the regional heterogeneity and microenvironment of human glioblastoma using massively parallel single-cell RNA-seq
Jinzhou Yuan, Hanna M. Levitin, Veronique Frattini, Peter Canoll, Jeffrey N. Bruce, Antonio Iavarone, Anna Lasorella, Peter A. Sims.
Presenter affiliation: Columbia University Medical Center, New York, New York. 233

Adaptive resistance to chemotherapy in triple-negative breast cancer revealed by single cell DNA and RNA sequencing
Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Nicola Crosetto, Theodoros Foukakis, Nicholas Navin.
Presenter affiliation: The University of Texas MD Anderson Cancer Center, Houston, Texas. 234

Multiplex strategies for measuring variant effects
Jay Shendure.
Presenter affiliation: University of Washington, Seattle, Washington.

Molecular dissection of complex germline structural variation using a multi-omics approach
Sjors Middelkamp, Judith Vlaar, Wigard Kloosterman, Ewart Kuijk, Edwin Cuppen.
Presenter affiliation: UMC Utrecht, Utrecht, Netherlands. 235

Spatial maps of prostate cancer transcriptomes reveals an unexplored landscape of heterogeneity
Emelie Berglund, Jonas Maaskola, Niklas Schultz, Maja Marklund, Joseph Bergenstråhle, Stefanie Friedrich, Firas Tarish, Anna Tanoglidi, Christoph Ogris, Erik Sonnhammer, Thomas Helleday, Patrik Ståhl, Joakim Lundeberg.
Presenter affiliation: Science for Life Laboratory, Stockholm, Sweden. 236

Ultrafast metagenomics at the bench and in the clinic
Aurélie Kapusta, Steven Flygare, Chase Miller, Yi Qiao, Gabor Marth, Edgar J. Hernandez, Qing Li, Helena Safavi-Hemami, Samuel D. Robinson, Aiping Lu, Baldomero Olivera, Guochun Liao, Martin G. Reese, Robert Schlaberg, Mark Yandell.
Presenter affiliation: University of Utah, Salt Lake City, Utah; USTAR, Salt Lake City, Utah. 237

Happy Hour
Sponsored by **Swift Biosciences**

FRIDAY, May 12—9:00 AM

SESSION 9 POPULATION GENOMICS

Chairperson: **Molly Schumer,** Harvard University / Columbia University
Shamil Sunyaev, Brigham & Women's Hospital,
Harvard Medical School, Boston, Massachusetts

Natural selection and local recombination rates shape the genome evolution of swordtail hybrids
Molly Schumer.
Presenter affiliation: Harvard University / Columbia University, New York, New York. 238

A wolf in sheep's clothing—A selfish element disguised as a linked pair of developmental genes underlies a genetic incompatibility in *C. elegans*
Eyal Ben-David, Alejandro Burga, Leonid Kruglyak.
Presenter affiliation: UCLA / Howard Hughes Medical Institute, Los Angeles, California. 239

DNA sequencing of single sperm using a novel approach for whole genome amplification provides critical insights into meiosis and recombination
Anjali Hinch, Gang Zhang, Philipp Becker, Ben Davies, Daniela Moralli, Cath Green, Rory Bowden, Peter Donnelly.
Presenter affiliation: University of Oxford, Oxford, United Kingdom. 240

Background selection is the dominant mode of linked selection in humans
David Murphy, Guy Sella.
Presenter affiliation: Columbia University, New York, New York. 241

Shamil Sunyaev.
Presenter affiliation: Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts.

FRIDAY, May 12—2:00 PM

SESSION 10 POSTER SESSION III

Protein-altering and regulatory genetic variants near *GATA4* implicated in bicuspid aortic valve
Wei Zhou, Bo Yang, Jiao Jiao, Jonas B. Nielsen, Michael Mathis, Simon C. Body, Gonçalo Abecasis, Kim Eagle, Alan P. Boyle, Bicuspid Aortic Valve Consortium, Cristen J. Willer.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 337

Integrative personal omics profiles during periods of weight gain and loss
Brian Piening, Wenyu Zhou, Kevin Contrepois, Hannes Röst, Gucci Gu, Tejaswini Mishra, Blake Hansen, Eddy Bautista, Shana Leopold, Christine Yeh, Daniel Spakowicz, Kimberly Kukurba, Dalia Perelman, et al., Erica Sodergren, Tracey McLaughlin, George Weinstock, Michael Snyder.
Presenter affiliation: Stanford University School of Medicine, Stanford, California. 338

FRIDAY, May 12—4:30 PM

GUEST SPEAKERS

Andrew Clark
Cornell University

“Junk Evolution”

Aviv Regev
Broad Institute of MIT and Harvard

“Reconstructing intra- and inter-cellular circuits with single cell genomics”

FRIDAY, May 12—6:00 PM

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SATURDAY, May 13—9:00 AM

SESSION 11 COMPUTATIONAL GENOMICS

Chairperson: **Anshul Kundaje,** Stanford University, California
Alexis Battle, Johns Hopkins University, Baltimore, Maryland

AUTHOR INDEX

GENOMIC CHARACTERIZATION OF BREAST CANCER PROGRESSION

Elaine R Mardis[1], Christopher A Miller[2,4], Marni Siegel[3], Katherine Hoadley[3], Jeremy Hoog[4], Sherri Davies[4], Lisa Carey[3], Joel Parker[3], Charles Perou[3], Matthew J Ellis[5]

[1]Nationwide Children's Hospital, Institute for Genomic Medicine, Columbus, OH, [2]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [3]University of North Carolina at Chapel Hill, Lineberger Comprehensive Cancer Center, Chapel Hill, NC, [4]Washington University School of Medicine, Department of Medicine, Division of Oncology, St. Louis, MO, [5]Baylor College of Medicine, Lester and Sue Smith Breast Center, Houston, TX

Large-scale cancer genomics efforts such as TCGA and ICGC have provided a comprehensive view of the primary, treatment-naïve landscape of human breast cancers. However, our understanding of how this landscape changes in the course of progression to metastatic disease is poorly understood, primarily because medical practice does not routinely evaluate metastatic biopsies. By consenting patients prior to their deaths from metastatic breast cancers, rapid autopsy provides a means of banking material from multiple metastatic sites that can be studied using modern genomics, often in comparison to the banked primary tumor. In this study, we performed NGS on tumor and matched metastatic sites using exome or whole genome sequencing, and performed RNAseq from these tumors. We applied integrated analytical approaches to tease out aspects of heterogeneity changes in progression and to understand the relationships between primary and metastatic disease across different breast cancer subtypes. Remarkably, our results support the notion that multi-clonal, rather than mono-clonal seeding from primary to metastasis can occur. This mode of seeding metastatic sites appears to be more prevalent in the triple negative subtype but can occur in other subtypes as well. Also surprising was the extent of similarity between different metastatic lesions when compared to the primary disease, indicating that very few metastasis-unique lesions emerge during progression. This finding reinforces the idea that if treatment regimens can address the genomic drivers in primary breast cancer, our ability to eradicate the disease prior to progression would be potentially curative.

A PAN CANCER ANALYSIS OF PROMOTER ACTIVITY HIGHLIGHTS THE REGULATORY ROLE OF ALTERNATIVE TRANSCRIPTION START SITES AND THEIR ASSOCIATION WITH NONCODING MUTATIONS

Deniz Demircioglu[1], Tannistha Nandi[1], Engin Cukuroglu[1], Claudia Calabrese[2], Nuno Fonseca[2], Andre Kahles[3], Kjong Lehmann[3], Steve Rozen[4], Bin Tean Teh[4], Oliver Stegle[2], Alvis Brazma[2], Angela Brooks[5], Gunnar Raetsch[3], Patrick Tan[1,2], Jonathan Goeke[1]

[1]Genome Institute of Singapore, Singapore, [2]EMBL-EBI, Hinxton, United Kingdom, [3]ETH, Zuerich, Switzerland, [4]Duke-NUS, Singapore, [5]University of California Santa Cruz, Santa Cruz, CA, [6]PCAWG Network, PCAWG Network, MA

The promoter holds a central position in regulation of gene expression, making it vulnerable to mutational processes or alterations in epigenetic modifications. This is particularly relevant to cancer, where alterations that affect gene products and their regulation often have severe consequences. Interestingly, most genes have more than 2 different alternative promoters, suggesting that the choice of promoter is as important as its level of transcriptional activity. However, data such as ChIP-Seq or CAGE-Tag that is aimed at identifying promoters is not available for most cancer studies, and the role of alternative promoters in cancer remains largely unknown.

Here we demonstrate that active promoters can be identified from RNA-Seq data alone, enabling the simultaneous analysis of promoter activity with somatic mutations across more than 1,000 samples in 30 cancer types. We find that alternative promoters are a major contributor to tissue-specific regulation of isoform expression, and that alternative promoters are frequently deregulated in cancer. These cancer-associated alternative promoters regulate expression of isoforms of known cancer-genes and novel candidates, some of which are deregulated across almost all cancer types. By screening for active promoters that show tissue-specific accumulation of mutations, we identify several candidate mutational hotspots that might be involved in driving cancer.

In summary, our study demonstrates the pervasive role of alternative promoters in context-specific isoform expression, regulation of isoform diversity, and highlights aberrant promoter activation in cancer. We provide a comprehensive catalog of active promoters and their expression pattern across 29 cancer types that will be a highly useful resource to understanding the roles of gene regulation and noncoding mutations in cancer. Our results suggest that promoter choice can have a profound effect on gene and isoform expression, opening many opportunities to further explore the interplay of regulatory mechanism and mutational processes with transcriptional aberrations in cancer.

EXPANDING DISCOVERY FROM CANCER GENOMES BY INTEGRATING NETWORK ANALYSES WITH MASSIVELY PARALLEL *IN VIVO* TUMORIGENESIS ASSAYS

Heiko Horn [1,2], Michael Lawrence[2,3], Candace Chouinard[2], Yashaswi Shresta[2], Jessica Hu[1,2], Elizabeth Worstell[1,2], Emily Shea[2], Nina Ilic[2,4], Ejung Kim[2,4], Atanas Kamburov[2,3], Alireza Kashani[1,2], William Hahn[2,4], Joshua Campbell[2,5], Jesse Boehm[2], Gad Getz[2,3], Kasper Lage[1,2]

[1]Massachusetts General Hospital, Department of Surgery, Boston, MA, [2]Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, [3]Massachusetts General Hospital, Department of Pathology, Boston, MA, [4]Dana Farber Cancer Institute, Department of Medical Oncology, Boston, MA, [5]Boston University, Department of Medicine, Boston, MA

Gene-based statistical tests to find cancer genes look for increased rates of somatic mutations or genomic copy number changes in cancer genomes. However, considerable sample sizes are required to find driver genes with intermediate or low mutation frequencies, and additional cancer genes remain to be discovered. Previous analyses have shown that cancer mutations in some cases converge on specific functional genomics sub networks. This suggests that mutations in a genes' functional network can be predictive of whether it is a cancer gene itself. However, this hypothesis has never been systematically explored across hundreds of known cancer genes, tens of tumor types, and thousands of cancer genomes. More importantly, analyses of cancer gene networks have not previously been coupled to systematic experimental validation assays and their predictive power to provide new insight into tumor biology remains unclear. We develop a statistic (NetSig) that combines molecular protein network information and existing cancer sequencing data to identify genes with a significantly mutated gene network (excluding data on the gene itself). We apply NetSig to data from 4,742 tumors spanning 21 tumor types and identify known and recently proposed driver genes in most (~ 60%) tumor types. NetSig also identifies 62 other genes with a significantly mutated gene network many suggesting new cancer biology. We test 25 known driver genes (positive controls), 33 NetSig candidates, and 79 random genes (random controls) in a massively parallel *in vivo* tumorigenesis cell assay. We demonstrate that the NetSig candidates induce tumors at rates that are comparable to the known driver genes and eightfold higher than random genes when injected into mouse models. Guided by the NetSig results and functional validation experiments, we looked for mutations and copy number changes in these genes that could explain 242 (out of a total of 660) lung adenocarcinomas without any known driver event; the analysis identified significant amplifications of several NetSig candidates in this patient subgroup. Overall, we present an integrated workflow that complements gene-based statistical tests by combining molecular network information, cancer sequencing data, and in vivo tumorigenesis assays to find and validate new driver genes in existing cancer genome data. The framework we describe is scalable to the rapid production of data and should become increasingly powerful as more tumors are sequenced in the future.

THE GENOMIC AUTOPSY: USING WHOLE EXOME AND WHOLE GENOME SEQUENCING TO SOLVE COMPLEX FETAL AND NEONATAL PRESENTATIONS

Alicia B Byrne[1,7], Jinghua Feng[2,8], Andreas W Schreiber[2,8], Peter J Brautigan[1], Nathalie J Nataren[1], Sui Yu[3], Yee Khong[4], Lynette Moore[4,9], Marcel E Dinger[5], Christopher N Hahn[1,9], Karin S Kassahn[3,8], Christopher P Barnett[6,9], Hamish S Scott[1,3,7,9]

[1]Centre for Cancer Biology, Molecular Pathology Research Laboratory, Adelaide, Australia, [2]Centre for Cancer Biology, ACRF Cancer Genomics Facility, Adelaide, Australia, [3]SA Pathology, Genetics and Molecular Pathology, Adelaide, Australia, [4]SA Pathology, Surgical Pathology, Adelaide, Australia, [5]Garvan Institute of Medical Research, Kinghorn Centre for Clinical Genomics, Sydney, Australia, [6]Women's and Children's Hospital, SA Clinical Genetics Service, Adelaide, Australia, [7]University of South Australia, School of Pharmacy and Medical Sciences, Adelaide, Australia, [8]University of Adelaide, School of Biological Sciences, Adelaide, Australia, [9]University of Adelaide, School of Medicine, Adelaide, Australia

Congenital abnormalities are the most frequent reason for termination of pregnancy, stillbirth, and neonatal death. Despite complex and thorough investigation, current autopsy examinations are unable to identify the cause in the majority of cases. A further ~10% of all perinatal deaths remain completely unexplained. In this study, we examine the utility of a ‘genomic autopsy’ for elucidating the genetic causes underlying congenital abnormalities and unexplained perinatal death. Whole exome sequencing (WES) and whole genome sequencing (WGS) are being performed using either an Illumina HiSeq 2500 (WES) or X Ten System (WGS). High priority cases are those where i) parents are consanguineous, ii) congenital abnormalities are multiple, or iii) abnormalities have recurred in families. Major phenotypes investigated include neurologic abnormalities, ciliopathy-like disorders, urinary tract malformations, non-immune hydrops, and musculoskeletal abnormalities. All cases are pre-screened for indicated Mendelian causes. An analysis pipeline incorporating bioinformatic and experimental laboratory techniques has been developed to identify and causally link novel variants and genes to disease. Of 13 families investigated to date (9 WES and 4 WGS), 6 have resulted in the identification of causative genetic variants; 2 in known disease genes, 2 expanding the phenotypic spectrum of known disease genes, and 2 in novel disease genes. 4/6 families are using this information for pre-implantation genetic diagnosis, already leading to a successful reproductive outcome for 1 family. A further 2 cases likely representing novel discoveries are currently under investigation. All findings are supported by in-depth functional evidence. This ongoing work demonstrates the power of a genomic autopsy in not only providing affected families with a diagnosis and facilitating family planning, but also in identifying novel genes critical in embryonic developmental pathways and furthering our understanding of early human development.

GENETIC DIVERSITY IN MULTI-REGION SEQUENCING DATA REFLECTS THE MODE AND TEMPO OF TUMOR EVOLUTION

Christina Curtis

Stanford University, Medicine & Genetics, Stanford, CA

Given the implications of tumor dynamics for precision medicine, there is need for a systematic classification of different modes of evolution in diverse solid tumors. In particular, although selection is fundamental to tumorigenesis, methods to infer the relative strength of selection within established human tumors are lacking. By simulating spatial tumor growth under different modes of evolution and examining patterns of genetic variation in multi-region sequencing (MRS), we demonstrate that it is feasible to distinguish strong positive selection from neutral tumor evolution, whereas weak selection and neutral evolution were indistinguishable. We developed a classifier based on features of the site frequency spectrum and fit experimental MRS data into model space, revealing different modes of evolution both within and between solid tumor types. These findings have implications for defining the drivers of tumor growth and inform practical guidelines for characterizing human tumor evolution.

FREQUENCY AND PROPERTIES OF MOSAIC SOMATIC MUTATIONS IN A NORMAL DEVELOPING BRAIN

Taejeong Bae[1], Jessica Mariani[2], Livia Tomasini[2], Bo Zhou[3], Alexander E Urban[3], Alexej Abyzov[1], Flora M Vaccarino[2]

[1]Mayo Clinic, Health Sciences Research, Rochester, MN, [2]Yale University, Child Study Center, New Haven, CT, [3]Stanford University, Psychiatry and Genetics, Palo Alto, CA

As mounting evidence indicates, each cell in the human body has its own genome, a phenomenon called somatic mosaicism. Few studies have been conducted to understand post-zygotic accumulation of mutations in cells of the healthy human body. Starting from single cells, directly obtained from three fetal brains, we established 31 separate colonies of neuronal progenitor cells, and carried out whole-genome sequencing on DNA from each colony. The clonal nature of these colonies allows a high-resolution analysis of the genomes of the founder progenitor cells without being confounded by the artifacts of in vitro single cell whole genome amplification. Across the three brains we detected between 100 and 300 non-germline SNVs per clone. Validation experiments (with PCR, digital droplet PCR, and capture deep sequencing) revealed high specificity (>95%) and sensitivity (>80%) of the SNVs as well as confirmed the presence of over 50% of SNVs in the original brain tissues, thereby proving that the detected SNVs represent genuine mosaic variants present in neuronal progenitors.

The per-cell number of mosaic SNVs increased linearly with brain age allowing us to estimate the mutation rate at 0.5-4.5 per cell division (95% CI). Dozens of SNVs were genotyped in multiple different regions of a brain and even in blood, suggesting that they have likely occurred prior to gastrulation. On a coarse-grained scale mosaic SNVs were distributed uniformly across the genome and were enriched in mutational signatures observed in medulloblastoma, neuroblastoma, as well as in a signature observed in all cancers and in de novo variants and which, as we previously hypothesized, is a hallmark of normal cell proliferation. Correlations with histone marks further strengthened the similarity of mosaic mutations in normal fetal brain with somatic mutations reported for medulloblastoma. On a smaller scale SNVs were mostly benign, tended to avoid DNAse hypersensitive sites and were enriched in genes related to synapse development and functions. These findings reveal a large degree of somatic mosaicism in the developing human brain, link de novo and cancer mutations to normal mosaicism and set a baseline for mosaic genome variation related to human brain development and function.

A POPULATION PHYLOGENY APPROACH TO UNDERSTANDING MITOCHONDRIAL HETEROPLASMY

Peter R Wilton[1], Thorfinn S Korneliussen[1], Marcia Su[2], Arslan Zaidi[2], Kateryna Makova[2], Rasmus Nielsen[1]

[1]UC Berkeley, Department of Integrative Biology, Berkeley, CA, [2]Penn State University, Department of Biology, University Park, PA

Mitochondrial genomic variation within an individual, termed heteroplasmy, is the major cause of mitochondrial disease and has roles in aging and tumorigenesis. Recently, heteroplasmy has also been shown to be a part of healthy human biology. Heteroplasmy is created and maintained by processes that can be viewed from the perspective of population genetics: mitochondria proliferate and diverge from one another during development, replace one another throughout adult life, mutate, and potentially experience natural selection within tissues. Here, we present a novel population genetic framework for studying heteroplasmy in which we model the mitochondria of several tissues sampled from a family as populations related by an ontogenetic phylogeny reflecting human development and reproduction. In this framework, we calculate a likelihood for observed heteroplasmy frequencies, accounting for genetic drift, mutation, and natural selection, and perform inference using Bayesian MCMC. We apply this inference procedure to an in-house dataset of mitochondrial genomes sampled from multiple tissues in >100 human families and sequenced to high depth (~ 5000x), as well as to publicly available datasets. By inferring how different population-genetic forces act along each branch of the ontogenetic phylogeny, we learn about the processes shaping genetic variation in mitochondria within an individual and between related individuals. We find that genetic drift, mutation, and natural selection each have a role in explaining mitochondrial genetic variation amongst tissues and that these forces continue to act on mitochondria throughout adult life. This work both refines previous estimates of the germline mitochondrial bottleneck, an important determinant of mitochondrial disease inheritance, and provides a new, quantitative understanding of the proliferation of mitochondria in the body and the differentiation of cell types during ontogenesis.

H3K27M AND THE BALANCE BETWEEN H3K27ME3 AND H3K27AC IN DIPG CELLS

Andrea Piunti[1], Rintaro Hashizume[1,2], Marc A Morgan[1], Elizabeth Thomas Bartom[1], Craig M Horbinski[2], Stacy A Marshall[1], Emily J Rendleman[1], Quanhong Ma[1,2], Yoh-hei Takahashi[1], Ashley R Woodfin[1], Alexander V Misharin[3], Nebiyu A Abshiru[4], Rishi R Lulla[5], Amanda M Saratsis[2,6,7], Neil L Kelleher[4], C David James[1,2,7], Ali Shilatifard[1,7]

[1]Northwestern University Feinberg School of Medicine, Biochemistry and Molecular Genetics, Chicago, IL, [2]Northwestern University Feinberg School of Medicine, Neurological Surgery, Chicago, IL, [3]Northwestern University Feinberg School of Medicine, Medicine, Chicago, IL, [4]Northwestern University, Chemistry, Evanston, IL, [5]Lurie Children's Hospital of Chicago, Hematology, Oncology, Neuro-Oncology & Stem Cell Transplantation, Chicago, IL, [6]Lurie Children's Hospital of Chicago, Surgery, Chicago, IL, [7]Northwestern University Feinberg School of Medicine, Robert H Lurie Comprehensive Cancer Center, Chicago, IL

Heterozygous mutation of lysine 27 of histone H3 to methionine (H3K27M) is found in the majority of cases of the pediatric brain cancer DIPG (Diffuse Intrinsic Pontine Glioma). While rare, this disease is invariably fatal and devastating to affected families. It has been previously reported that presence of the H3K27M histone is associated with a global decrease in trimethylation (H3K27me3) and global increase in acetylation (H3K27ac) at this residue. A model has been developed whereby H3K27M interferes with the function of Polycomb Repressive Complex 2 (PRC2), binding to it and sequestering it. We performed a ChIP-seq analysis of the genomic distribution of histone H3K27M, H3K27me3, H3K27ac, and components of the PRC2 complex in three different patient derived DIPG cell lines. We find peaks of PRC2 activity, but they are always associated with peaks of H3K27me3, and almost never with peaks of H3K27M. Instead we find that H3K27M peaks are strikingly co-located with H3K27ac and frequently with the bromodomain containing proteins Brd2 and Brd4. A pulldown with an antibody against H3K27M histones also pulls down H3K27ac histones, indicating the presence of heterotypic H3K27M/K27ac nucleosomes. Furthermore, we find that treatment of a xenograft model of the disease with bromodomain inhibitors suppresses tumor growth. We propose that H3K27M affects the balance of H3K27me3 and H3K27ac in the genome primarily by selectively increasing H3K27ac at specific genomic loci through the formation of the H3K27M/ac heterotypic nucleosomes that exclude PRC2 binding and recruit Brd2/4 to drive a DIPG oncogenic transcriptional network.

INHERITED GENETIC VARIATION AT MANY LOCI SHAPES SOMATIC GENOME EVOLUTION AND CLONAL EXPANSIONS IN APPARENTLY HEALTHY PEOPLE

Po-Ru Loh[1,2], Giulio Genovese[1,3,4], Samuel Bakhoum[5,6], Hilary K Finucane[1,7], Yakir A Reshef[8], Pier Francesco Palamara[1,2], Robert E Handsaker[1,3], Steven A McCarroll[1,3,4], Alkes L Price[1,2,9]

[1]Broad Institute, Medical and Population Genetics, Cambridge, MA, [2]Harvard, Epidemiology, Boston, MA, [3]Harvard, Biostatistics, Boston, MA, [4]Broad Institute, Stanley Center, Cambridge, MA, [5]Harvard, Genetics, Boston, MA, [6]MSKCC, Radiation Oncology, New York, NY, [7]Weill Cornell Medicine, Meyer Cancer Center, New York, NY, [8]MIT, Mathematics, Cambridge, MA, [9]Harvard, Computer Science, Cambridge, MA, [10]Harvard, Biostatistics, Boston, MA

Selective pressures that shape clonal evolution in healthy individuals are largely unknown. Earlier work showed that a surprising fraction (>10%) of individuals over 65 harbors clonally expanded blood cells.

Clonally expanded cells harboring structural alterations distort representations of alleles at heterozygous sites. Using population data from ~150,000 individuals from the UK Biobank to phase such alleles at long genomic distances, we greatly increased power to detect such events at allelic fractions as low as 1%. We identified 8,342 large mosaic structural alterations (27% deletions, 16% duplications, and 57% copy-number neutral loss of heterozygosity events (CNN-LOHs)), one order of magnitude more than in any previous study.

We identified five separate loci at which inherited variation acts in cis to shape either the likelihood of nearby somatic events or the selective pressure upon such events: one known near JAK2 (p<10-12), and four novel at MPL (p<10-16), 10q25.3 (p<10-50), 15q26.3 (p<10-25), and Xp11.21 (p<10-20), each with a different genomic or selective model to explain the association.

Data revealed many relationships: a general correlation of clonal expansions with inherited variation regulating telomere length; >30 regions with an excess of focal deletions and/or CNN-LOHs; non-random overlaps between focal deletions and CNN-LOHs, suggesting convergent patterns of selection; and we found that chromosomes with a lower-than-average burden of focal deletions are more likely to be duplicated.

In analyses of health outcome data from the UK cancer and death registries (median follow-up ~5 years), we observed that several mosaic events commonly seen in patients with chronic lymphocytic leukemia (CLL) and myeloproliferative neoplasms (MPN) were strong risk factors for these malignancies (OR>100).

This extensive atlas of structural alterations in clonal expansions from blood sheds light on the multitude of patterns that will be visible in early stages of clonal expansions from blood and it presages the future development of similar atlas for other tissues.

MOLECULAR AND FUNCTIONAL VARIATION IN IPSC-DERIVED SENSORY NEURONS.

Jeremy Schwartzentruber[1], Stefanie Foskolou[2], Helena Kilpinen[3], Julia Rodrigues[1], Kaur Alasoo[1], Andrew J Knights[1], Minal Patel[1], Angela Goncalves[1], Rita Ferreira[2], Caroline L Benn[2], Anna Wilbrey[2], Magda Bictash[2], Emma Impey[2], Lishuang Cao[2], Sergio Lainez[2], Paul J Whiting[2], Alex Gutteridge[2], Daniel Gaffney[1]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, [2]Pfizer, Neuroscience and pain research unit, Cambridge, United Kingdom, [3]European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Cellular disease models are critical for understanding the molecular mechanisms of disease and for the development of novel therapeutics. Advancements in induced pluripotent stem cell (iPSC) technology are enabling the development of these models in many human cell types. Initial uses of iPSCs for disease modeling have focused mostly on highly penetrant, rare coding variants with large phenotypic effects. However, there is growing interest in using iPSCs to model the effects of common genetic variants that drive complex disease. Regulatory genetic variants have been identified for gene expression, chromatin accessibility and transcription factor binding in primary tissues and immortalized cell lines, however, for many tissues, pure cultures of individual cell types are either unavailable or are too scarce to enable a variety of molecular assays to be performed. iPSC-derived cells are a renewable source of cells which can be genetically manipulated to investigate causal genetic effects.

Here, we present the first large-scale study of common genetic effects in a cell type differentiated from human stem cells, iPSC-derived sensory neurons (IPSDSNs). To investigate genetic contributions to human sensory function, we performed 123 differentiations of iPSCs from 103 unique donors to a sensory neuronal fate, and measured gene expression, chromatin accessibility, and neuronal excitability. Compared with primary dorsal root ganglion, where sensory nerves collect near the spinal cord, gene expression was more variable across iPSC-derived neuronal cultures, particularly in genes related to differentiation and nervous system development. Single cell RNA-sequencing revealed that although the majority of cells are neuronal and express the expected marker genes, a substantial fraction have a fibroblast-like expression profile.

By applying an allele-specific method we identify 3,778 quantitative trait loci influencing gene expression, 6,318 for chromatin accessibility, and 2,097 for RNA splicing at FDR 10%. A number of these overlap with common disease associations, and suggest candidate causal variants and target genes. These include known causal variants at SNCA for Parkinson’s disease and TNFRSF1A for multiple sclerosis, as well as new candidates for migraine, Parkinson’s disease, and schizophrenia.

IDENTIFYING GENETIC VARIANTS THAT AFFECT VIABILITY IN LARGE COHORTS

Hakhamanesh Mostafavi[1], Tomaz Berisa[2], Felix R Day[3], John R Perry[3], Molly Przeworski*[1], Joseph K Pickrell*[1,2]

[1]Columbia University, Biological Sciences, New York, NY, [2]New York Genome Center, New York, NY, [3]University of Cambridge, Institute of Metabolic Science, MRC Epidemiology Unit, Cambridge, United Kingdom

A number of open questions in human evolutionary genetics would become tractable if we were able to directly measure evolutionary fitness. As a step towards this goal, we developed a method to test whether individual genetic variants, or sets of genetic variants, currently influence viability. The approach consists in testing whether the frequency of an allele varies across ages in large cohorts, accounting for variation in ancestry. It is similar in spirit to a genome-wide association study for survival, but also allows us to identify possible non-monotonic effects at different ages and to learn about sex differences. We applied it to two recent datasets: to 57,696 individuals of European ancestry from California from the Resource for Genetic Epidemiology Research on Aging (GERA) Cohort and, by proxy, to the parents of 95,513 individuals of European ancestry surveyed as part of the UK Biobank. In the GERA cohort, the top signal is the APOE ε4 allele ($P < 1e-15$), a known major risk factor for Alzheimer's disease. In the UK Biobank, the strongest signals are detected in fathers only, and are for variants near CHRNA3 ($P \sim 4e-8$), previously shown to be associated with smoking quantity. That there exist so few common variants with effects on survival only late in life, despite considerable power, suggests that even variants with late onset effects are kept at low frequency by purifying selection. We further tested for the joint effect of sets of genetic variants on viability, using genome-wide significant associations previously identified in genome-wide association studies for 42 polygenic traits. In the UK Biobank data, variants that delay puberty timing are associated with a higher chance of survival ($P \sim 2e-7$), consistent with epidemiological studies. Similarly, variants associated with later age at first birth are enriched among longer-lived mothers ($P \sim 6e-4$); interestingly, they are also associated with lower number of children, pointing to an apparent trade-off between effects on fertility and longevity. Other signals are also observed for variants influencing cholesterol level, heart disease risk, and body mass index, but only among the fathers and only in the UK Biobank. Differences in the findings between datasets suggest gene-environment interactions that differ between California in the early 21th century (GERA) and the UK in the mid to late 20th century. More generally, our analysis serves as a proof of principle for how upcoming biomedical datasets can be used to learn about selection effects in contemporary humans.

STRAIN-LEVEL IDENTIFICATION OF MOTHER-TO-CHILD BACTERIAL TRANSMISSION DURING THE FIRST FEW MONTHS OF LIFE

Moran Yassour[1,2], Larson Hogstrom[1], Eeva Jason[3], Heli Siljander[4], Jenni Selvenius[3], Sami Oikarinen[5], Heikki Hyöty[5], Jorma Ilonen[6], Suvi Virtanen[7], Hera Vlamakis[1,2], Eric S Lander[1], Mikael Knip[4], Ramnik J Xavier[1,2]

[1]The Broad Institute of MIT and Harvard, Cambridge, MA, [2]Massachusetts General Hospital, Center for Computational and Integrative Biology, Boston, MA, [3]Tampere University Hospital, Dept. of Pediatrics, Tampere, Finland, [4]Children's Hospital, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland, [5]University of Tampere, Dept. of Virology, School of Medicine, Tampere, Finland, [6]University of Turku and Turku University Hospital, Immunogenetics Lab, Turku, Finland, [7]National Institute for Health and Welfare, Dept. of Health, Helsinki, Finland

The human gut microbiota is established during the first few years of life, yet we know remarkably little about the natural history of how the microbiome forms in children. Early events in microbial colonization have a profound effect on physiology and immune education in the gut, thereby impacting disease susceptibility (e.g., obesity, asthma, and other inflammatory disorders later in life).

Using a newly established prospective birth cohort, we studied the gut microbiome of 24 infants and their mothers, sampled longitudinally in the first few months of the child's life. We compared the microbial communities across and within families to identify bacterial transmission events.

We found that newborn samples collected within two days of delivery exhibit low-complexity microbial communities, where often 1-4 species account for >85% of the community. Using species-level abundance profiles, we can identify communities dominated by either *E. coli, Bifidobacterium*, or *Staphylococcus* species. We next used the deep metagenomic sequencing to identify strains by comparing SNPs across the genomes of these highly abundant species. Despite the species-level similarity across unrelated children, strain-level analysis reveals unique shared strains that are transferred from mothers to their children.

US IMMIGRATION WESTERNIZES THE HUMAN GUT MICROBIOME

Pajau Vangay[1], Chaisiri Angkurawaranon[2], Rose McGready[3], Shannon Pergament[4], Kathleen Culhane-Pera[4], Dan Knights[5]

[1]University of Minnesota, Bionformatics and Computational Biology, Minneapolis, MN, [2]Chiang Mai University, Family Medicine, Chiang Mai, Thailand, [3]Oxford University, Tropical Medicine, Mae Sot, Thailand, [4]West Side Community Health Services, Quality and Research, St. Paul, MN, [5]University of Minnesota, Computer Science and Engineering, Minneapolis, MN

Immigrants in the US, such as the Hmong and Karen in Minnesota, are developing chronic diseases, such as obesity, at alarming rates. Drastic and permanent changes in dietary and environmental exposures, characteristic of immigration, can lead to disruption of gut homeostasis. This project tests the hypotheses that immigration from developing countries to the USA induces loss of native microbial species, predisposing immigrants to obesity, and that increasing dietary fiber intake supports maintenance of the native microbiome. We collected cross-sectional dietary data, anthropometric data, and stool samples from 550 first- and second-generation Hmong and Karen female immigrants in the USA and Thailand. We also describe microbiome adaptation upon arrival in the USA in a 6-month longitudinal cohort of 20 individuals. Using 16s rRNA gene sequencing as well as deep shotgun metagenomics, we characterize how the gut microbiome changes with increasing US residency, and we identify associations with obesity. We found that lower gut microbiome diversity is associated with higher obesity risk, and that gut microbiomes of foreign-born individuals become indistinguishable from their US-born counterparts within approximately 5 years. In addition, we found that relative abundances of Western-associated *Bacteroides* increase over time in the US, and non-Western-associated *Prevotella* decreases over time in the US. This study is a critical first step towards understanding how the gut microbiome mediates or prevents chronic disease development among the rapidly growing US foreign-born population.

GUT MICROBIOTA INDUCES SPECIES-SPECIFIC GENE REGULATION ACROSS PRIMATES

Allison Richards[1], Adnan Alazizi[1], Michael Burns[2], Andres Gomes[2], Jonathan Clayton[3], Klara Petrzelkova[4], Amanda Muehlbauer[2], Roger Pique-Regi[1], Francesca Luca[1], Ran Blekhman[2]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Minnesota, Genetics, Cell Biology, and Development, Minneapolis, MN, [3]University of Minnesota, Veterinary and Biomedical Sciences, St. Paul, MN, [4]Czech Academy of Sciences, Institute of Vertebrate Biology, Brno, Czech Republic

The gut microbiota can regulate and train host immune response, perform important metabolic functions, produce nutrients, and protect against pathogen infection. Exciting new research is unraveling an extreme variation in the composition of gut microbial communities across primate species. However, we know very little about how this cross-species variation affects host health. An important mechanism by which the microbiome can affect host physiology is by altering gene expression in proximal colonic epithelial cells. Thus, it has been hypothesized that variation in the microbiome can control species-specific gene expression and potentially affect human-specific diseases.

Here, we used a novel experimental system based on colonic epithelial cells co-cultured with live microbiomes extracted from four primate hosts (human, chimpanzee, gorilla, and orangutan), with 4-8 individuals from each species. This allowed us to dynamically profile host gene expression changes (via RNA-seq) that are directly modulated by the microbiome. We found a conserved signature whereby 958 genes consistently respond to microbiomes from all four species. These genes are involved in cell adhesion, metabolic functions (such as cholesterol biosynthesis), and immune pathways, including interleukin signaling. We also identified genes whose response is driven by the abundance of specific microbial taxa across all species. In addition, we identified species-specific host gene response, with ~2500 genes that respond to microbiomes from one primate species and not the others. Host genes that respond specifically to human microbiomes are significantly enriched with genes that have been previously identified in GWAS studies as associated with microbiome-related health conditions, such as HDL cholesterol, obesity-related traits, celiac disease, and inflammatory bowel disease. Our study provides the first evidence that symbiosis with gut microbial communities affects species-specific gene regulation in primates, and demonstrates how human-specific microbiomes may influence host health.

TEMPORAL DYNAMICS OF METATRANSCRIPTION IN INFLAMMATORY BOWEL DISEASE

Melanie Schirmer[1,2], Eric A Franzosa[1,2], Jason Lloyd-Price[1,2], Alexandra Sirota-Madi[1,3], Lauren McIver[2], Randall Schwager[2], Hera Vlamakis[1], Ramnik J Xavier[1,3], Curtis Huttenhower[1,2]

[1]The Broad Institute of MIT and Harvard, Infectious Disease and Microbiome, Cambridge, MA, [2]Harvard T.H. Chan School of Public Health, Department of Biostatistics, Boston, MA, [3]Massachusetts General Hospital and Harvard Medical School, Boston, MA

Inflammatory bowel disease (IBD) is a group of chronic diseases of the digestive tract with no effective long-term treatment options and increasing incidence rates worldwide. One of the major breakthroughs in recent years was the elucidation of the human microbiome's role in the onset and exacerbation of IBD. In particular, the mechanisms associating gut microbial dysbioses and aberrant immune responses with the disease remain largely unknown. The integrative Human Microbiome Project (iHMP or HMP2) seeks to close these gaps by examining the dynamics of microbiome functionality in disease. As part of this effort, the IBD Multi'omics Database (IBDMDB) profiled the gut microbiomes of 100 individuals with new-onset and established disease using multiple high-throughput, functional genomic screens over the course of one year each. Here, we present the results based on the pilot data including 300 stool metagenomes and 78 metatranscriptomes, spanning 117 individuals with up to 17 time points per individual. While some microbial organisms exhibited similar expression patterns on the DNA and RNA level, we also detected species-specific biases in metabolic activity. For example, while *Faecalibacterium prausnitzii* was in many cases not the most prevalent organisms in a sample, it was for many pathways identified as the major transcriber. Further, certain disease characteristics were particularly detectable at the transcript level, such as pathways metagenomically uniform between IBD patients compared to non-IBD controls but metatranscriptomically contributed by distinct organisms (e.g. *Alistipes putredinis*). These IBDMDB pilot studies thus pave the way for analysis of the iHMP data incorporating metabolomics, host transcriptomics, epigenetics, and genetics as well as provide new opportunities for the discovery of novel diagnostic and therapeutic approaches in IBD.

Funding: This work is supported by NIDDK (NIH U54DE023798).

GENOMIC CLASSIFICATION OF GASTRIC CANCER FOR TREATMENT SELECTION

Akihiro Suzuki[1,2], Miwako Kakiuchi[1], Shumpei Ishikawa[3], Hiroyuki Aburatani[1]

[1]The University of Tokyo, Research Center for Advanced Science and Technology, Tokyo, Japan, [2]Yokohama City University Graduate School of Medicine, Department of Gastroenterology and Hepatology, Yokohama, Japan, [3]Tokyo Medical and Dental University, Medical Research Institute, Tokyo, Japan

Genomic sequencing analysis have been applied to classify gastric cancer and led to discovery of a potential therapeutic target, e.g. RHOA mutation in 25% of diffuse-type gastric cancer cases (Kakiuchi 2014). To further explore driver mutations and biomarkers for gastric cancer in Japanese population, we performed comprehensive genomic profiling of gastric carcinoma, namely 300 tumor/normal samples, archived at Yokohama City University and Tokyo University hospital, were analyzed by exome, RNA-seq, and DNA methylation profiling. Based on preliminary results from a subset of 103 cases, *TP53, CDH1, RHOA, ARID1A, PIK3CA*, and *KRAS* are recurrently mutated, as previously reported, while *CDH1* and *RHOA* were mutated mostly in diffuse-type gastric cancer. Hypermutator cases that are mostly caused by *MLH1* promoter methylation were observed at 16%, while high amplification of *ERBB2, EGFR, KRAS* and *VEGFA* accompanied with their overexpression, were seen in 13.6%, 3.9%, 2.9% and 4.9%, respectively. Overexpression of *PD-L1* was observed in 4 %, who would presumably respond well to immune checkpoint inhibitor treatment. Currently we are integrating all the data with clinicopathological data to identify a distinct subgroup and a new therapeutic target or signaling pathway in gastric cancer.

PREDICTING GENE EXPRESSION FROM TEMPORAL CHANGES OF THE REGULATORY LANDSCAPE IN HUMAN DENDRITIC CELLS

Shaked Afik[1], Pranitha Vangala[2], Elisa Donnard[2], David Fisher[3,4], Barbara Tabak[2], Patrick McDonel[2], Jeremy Luban[5], Manuel Garber[2,5], Nir Yosef[3]

[1]Computational Biology Graduate Group, University of California, Berkeley, Berkeley, CA, [2]Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, [3]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, [4]Department of Computer Science, ETH Zurich, Zurich, Switzerland, [5]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA

Activation and maturation of cells following environmental stimuli are controlled by transcriptional changes of thousands of genes. Those changes are mediated by a complex regulatory network that is comprised of non-coding DNA sequences, chromatin structure and transcription factors (TFs). However, the code linking these variables in a way that temporal changes in gene expression can be predicted has yet to be deciphered. We aim to model such code in human dendritic cells (hDCs), antigen presenting cells that help initiate the immune response, as they mature in response to lipopolysaccharide (LPS), a component of gram-negative bacteria.

In order to characterize the temporal changes of the regulatory regions in hDCs, we profiled the chromatin landscape at six time points following LPS stimulation with both ATAC-seq and ChIP-seq of few selected histone marks. Thus, we were able to identify thousands of non-coding genomic regions that exhibit significant changes in their accessibility and activity. We classified each region based on its temporal behavior across all assays, revealing various temporal patterns of the regulatory landscape.

We next combine the local features of the regulatory regions such as DNA composition and chromatin accessibility with the expression patterns of the genes that they regulate. By taking a supervised learning approach we will be able to discover the regulatory features that are important for each temporal pattern of gene expression. Our results, in combination with TF binding information, will allow us to uncover the grammar of transcriptional regulation.

The large expression changes as well as the various temporal transcriptional responses makes hDCs activation an ideal system to understand general mechanisms of gene regulation and gain a better grasp of the human immune system. We aim to build a combined computational and experimental platform that could be applied to study many other systems.

PIGMENTOR: ACCURATE PREDICTION OF MULTIPLE PIGMENTATION PHENOTYPES

Babak Alipanahi[1], Pierre Fontanillas[1], 23andMe Research Team[1], Steve Pitts[2], Robert Gentleman[2]

[1]23andMe, Inc., Research, Mountain View, CA, [2]23andMe, Inc., Therapeutics, South San Fransisco, CA

Eye, hair, and skin pigmentation are amongst the most heritable traits in humans, and although they are often presented as an example of simple Mendelian inheritance, they are actually highly polygenic. More than one hundred genetic loci have been significantly associated with human pigmentation traits. Perhaps not surprisingly, the genetic architectures of these three pigmentation traits are highly overlapping. However, this feature has not been used so far for predicting pigmentation. Here, we applied a custom-designed deep neural network on self-reported pigmentation phenotypes from a large cohort of 23andMe research participants of European ancestry for joint prediction of eye, hair, and skin colors. To exhaustively identify SNPs that could discriminate between self-reported pigmentation levels, for each pigmentation trait, we performed pairwise and one-vs-others GWASes, followed by conditional analyses on each of the significant loci. Then, we found all significant SNP:SNP and SNP:covariate interactions. This process helped us discover several novel SNPs in highly significant loci with complex haplotype structures, such as the *MC1R* and *OCA2-HERC2* loci. Our model, called "Pigmentor", trains very quickly on tens of thousands of samples and shows excellent predicting performance on held-out test data (not used in feature selection and model training): eye color, hair color, and skin color are predicted with weighted average AUCs over 75%. Pigmentor's inner neural network mechanics also provide insights into the genetic architecture of each pigmentation trait, by supplying an "importance" measure for each of the predictors. Pigmentor is a useful tool to predict the pigmentation phenotypes, to understand the architecture of these traits, and to analyze their involvement in skin cancer risk or other pigmentation-associated disorders.

THE IMPACT OF *PRDM9* EXPRESSION ON THE GENOMIC AND TRANSCRIPTOMIC LANDSCAPE OF CANCER

Armande Ang Houle[1,2], Mawusse Agbessi[1], Vanessa Bruat[1], PCAWG Consortium[1], Lincoln Stein[1,2], Philip Awadalla[1,2]

[1]Ontario Institute for Cancer Research, Informatics and Bio-Computing, Toronto, Canada, [2]University of Toronto, Molecular Genetics, Toronto, Canada

Homologous recombination is a process allowing for the exchange of genetic information between homologous chromosomes, and when impaired, has been linked to multiple genomic failings ranging from indels to larger scale aneuploidies. During meiosis, PRDM9 binding sites determine positions of double strand breaks by recruiting SPO11, a topoisomerase-like protein that catalyzes double strand breaks leading to the initiation of meiotic recombination. Previously, allelic variation at the *PRDM9* loci was associated with pediatric acute lymphoblastic leukaemia, suggesting a role for PRDM9 in some cancers. Although PRDM9 has a meiosis-specific function, and therefore is normally expressed solely in testis and in foetal ovaries, we found *PRDM9* expression in over 260 tumors across several cancer types from the International Cancer Genome Consortium (n=1256), even after stringent homology correction. *PRDM9* expression levels are significantly different from those found in healthy tissues, implicating cancer-specific expression of *PRDM9* in somatic cells. In particular, cancers originating from the ovary and from the liver show high proportions of tumors expressing *PRDM9. PRDM9* expression appears to also impact the transcriptional landscape, and potentially influence carcinogenesis through the down regulation of known tumor suppressors. Genomic rearrangements frequently neighbour PRDM9 binding sites within the genomic range of PRDM9's H3K4me3 epigenetic mark, hinting at an association between the location of PRDM9 and the initiation of double strand breaks, reminiscent of PRDM9's function in meiotic cells. We observe a similar colocalization between PRDM9 binding sites and gene fusions occurrences in ovarian tumours. This study is the first to highlight the role of the meiosis-specific gene *PRDM9* on the transcriptomic and genomic landscape of tumors, and suggests a mechanism to explain aberrant homologous recombination in cancers.

DETECTING CONFOUNDING DUE TO RESIDUAL POPULATION STRUCTURE IN GENOME-WIDE ASSOCIATION SUMMARY STATISTICS

Georgios Athanasiadis[1], Miguel M Álvarez-Álvarez[2], Kalle Leppälä[1], Mikkel H Schierup[1], Thomas Mailund[1], Bjarni J Vilhjálmsson[1]

[1]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark, [2]University of Barcelona, Department of Evolutionary Biology, Ecology and Environmental Sciences, Barcelona, Spain

Linkage disequilibrium (LD) score regression has been used widely for a less conservative adjustment of genome-wide association study (GWAS) summary statistics that are potentially inflated by population structure. However, LD between markers is not only attributable to physical proximity on the chromosome (i.e. *real* LD), but also to population structure (i.e. *long-range* LD). Here, through the use of a new principal component analysis (PCA)-based statistic, we show that LD score is potentially biased by population structure. Thus, a multiple regression using both LD- and PCA-based scores can be suboptimal unless the two variables are uncorrelated. To address this issue, we also introduce an adjusted version of LD score, which is largely free from population structure and relatedness, capturing only real LD. We regress the new LD score together with the PCA-based score against GWAS summary statistics and use the parameters of the model to adjust summary statistics from various GWASs.

IDENTIFYING GENE-BY-ENVIRONMENT INTERACTIONS IMPROVES SKIN CANCER RISK PREDICTION.

Pierre Fontanillas, 23andMe Research Team, Adam Auton

23andMe, Research, Mountain View, CA

Disease risks are determined by a complex interplay of genetics and environmental factors. In the last decade, genome-wide association studies (GWAS) were successful at identifying thousand of genetic risk loci. On the other hand, collecting and identifying environmental factors has proven to be challenging and data are lacking. Here, we are presenting exposure risk predictions and gene by environment (GxE) interaction analyses for skin cancers in a deep genotype/phenotype cohort. We surveyed a total of 112 thousand research participants via an online survey over a course of 4 months, of which 27 thousand reported having had skin cancer. Our survey contained >50 questions covering a variety factors relating to susceptibility, exposure, and family history. Using this data, we built a predictive risk model containing 29 variables that explained ~25% of the skin cancer variance. Using predicted risk factors as covariates, we included GxE interactions within a GWAS analysis. In doing so, we were able to assess the contribution of environment in modulating genetic risk.

microRNAs EXECUTE A BIMODULAR DEVELOPMENTAL PROGRAM IN THREE ANIMAL PHYLA

Gal Avital[1,2], Gustavo S França[2], Itai Yanai[2]

[1]Technion – Israel Institute of Technology, Biology, Haifa, Israel, [2]NYU School of Medicine, Institute for Computational Medicine, New York, NY

While RNA was historically relegated by the “central dogma” to an intermediate role, compelling evidence now abounds for myriad RNA classes playing central roles in most cellular processes. While developmental gene networks are typically framed in terms of protein-coding genes, the role of small RNA remains generally unexplored in this context. Here, we report a highly-resolved time-course of both miRNA and mRNA expression throughout embryogenesis of the nematode *C. elegans*. Examining miRNA developmental temporal profiles, we found that a majority of miRNA’s exhibit dynamic gene expression during this period. Strikingly, patterns are organized into a dichotomy of early and late gene expression modules, as opposed to gradual temporal expression throughout embryogenesis. To test for the universality of this observation, we further generated miRNA time-courses for *Drosophila* and zebrafish. Again, we found a general bimodality in terms of the main set of miRNA expression profiles. Although the miRNA repertoire is highly divergent across species, the orthologs generally show conserved early/late modality. Together, these results suggest that while miRNAs and their targets evolve quickly, the underlying regulatory program of development is deeply conserved and may correspond to a signature of animal development.

LEARNING CAUSAL GENE REGULATORY NETWORKS WITH MENDELIAN RANDOMIZATION

Md B Badsha[1], Audrey Q Fu[1]

[1]University of Idaho, Statistical Science, Institute for Bioinformatics and Evolutionary Studies, Center for Modeling Complex Interactions, Moscow, ID

It has been challenging to identify the regulatory (or causal) relationships among genes, both experimentally and computationally. Is it possible to go beyond correlation and learn causal relationships directly from genomic data collected in vivo? Is it possible to establish these relationships for many genes simultaneously? Mendelian Randomization (MR) makes tackling these questions possible. MR views genetic variants (SNPs, indels, and copy number variation) in a natural population as perturbations randomly performed by Nature, and provides a reasonable and potentially powerful assumption for studying causal relationships among genes. Here we present MPC, a novel, MR-based algorithm that use both genotype and gene expression data and efficiently learn a causal gene regulatory network (or a causal graph of genes). Our algorithm is a variant of the classic PC algorithm for learning causal graphs. Similar to PC, our algorithm also conducts a series of statistical tests for marginal and conditional independence. However, unlike existing PC algorithms, our MPC algorithm incorporates MR, and further controls the False Discovery Rate (FDR) and reduces the impact of outliers. We use MPC to study the regulatory relationships among frequently mutated cancer genes, which are identified from the TCGA data on breast cancer patients.

REMAP : AN INTEGRATIVE CHIP-SEQ ANALYSIS OF REGULATORY ELEMENTS

Jeanne Cheneby[1], Marius Gheorghe[2], Anthony Mathelier[2], Benoit Ballester[1]

[1]INSERM AMU, TAGC U1090, Marseille, France, [2]Centre for Molecular Medicine Norway, UiO Faculty of Medicine, Oslo, Norway

The large collections of ChIP-seq data rapidly accumulating in public data warehouses provide genome-wide binding region maps for hundreds of transcription factors (TFs). However, the extent of the regulatory occupancy space in the human genome has not yet been fully apprehended by integrating public ChIP-seq data sets and combining it with ENCODE TFs map.

In 2015 to enable genome-wide identification of regulatory elements we have collected, analysed and retained 395 available ChIP-seq data sets merged with ENCODE peaks covering a total of 237 TFs. This enhanced repertoire complements and refines current genome-wide occupancy maps by increasing the human genome regulatory search space by 14% compared to ENCODE alone, and also increases the complexity of the regulatory dictionary.

In 2017, we are updating this catalogue with the latest TF ChIP-seq data sets from GEO and ArrayExpress and all current ENCODE TF ChIP-seq data, both mapped to the hg38 genome assembly. This unprecedented catalogue of binding regions will include a set of 564 TFs.

To facilitate the exploration of the regulatory elements by the scientific community, we created an online resource called ReMap (http://tagc.univ-mrs.fr/remap/) to display information about TFs, peaks and data sets. Genomic tracks containing the 2015 hg19 data are available as tracks in the UCSC Genome Browser for browsing and visual exploration. The 2017 hg38 catalogue will be updated soon on the ReMap website.

Finally, we also developed a tool to allow the annotation of genomic regions provided by users. Those regions are compared against the ReMap catalogue returning statistical enrichments of TFs present within input regions compared to random expectations. It thus becomes possible to study bindings of specific TFs overrepresented in those regions.

INTEGRATING MOLECULAR MECHANISMS INTO GWAS SUMMARY RESULTS

Alvaro N Barbeira[1], Scott P Dickinson[1], Jiamao Zheng[1], Hae K Im[1], Rodrigo Bonazzola[1], Jason M Torres[2], Nancy J Cox[4], Eric S Torstenson[4], Heather E Wheeler[3], Kaanan P Shah[1], Todd Edwards[4], Dan L Nicolae[1]

[1]The University of Chicago, Section of Genetic Medecine, Chicago, IL, [2]The University of Chicago, Committee on Molecular Metabolism and Nutrition, Chicago, IL, [3]Loyola University Chicago, Departments of Biology and Computer Science, Chicago, IL, [4]Vanderbilt University Medical Center, Vanderbilt Genetic Institute, Nashville, TN

Genome-Wide Association Studies (*GWAS*) and Genome-Wide Association Meta Analysis Studies (*GWAMAS*) have been successful in identifying genetic loci that robustly associate with human complex traits. However, the mechanistic understanding of these discoveries is still limited, hampering the translation of this knowledge into actionable targets.

In order to leverage the wealth of information available in public *GWAS* and *GWAMAS*, we have proposed **MetaXcan**, a method that can integrate variant associations with information from molecular traits. By building models of molecular mechanisms such as gene expression, **MetaXcan** allows to test the association of a genetic feature to any given trait via said molecular mechanism.

In this work we present the application of **MetaXcan** to a set of over 60 traits, using models of Genetic Expression built from **Genotype-Tissue Expression Project** (*GTEx*), a large multiple tissue transcriptome data set with sample size up to **n=338**, and the Depressive Genes and Networks study on Whole Blood with sample size **n=922**. We discuss novel genetic findings concerning these traits, and discuss preliminary findings on other molecular mechanisms.

We also compare **MetaXcan** with recently published methods that integrate functional mechanisms with *GWAS* and *GWAMAS* information.

We make the results of this large in silico gene-trait association study publicly available at **http://gene2pheno.org/**

EXPLORING TUMOR EVOLUTION IN ZEBRAFISH USING SINGLE-CELL RNA-SEQ

Maayan Baron[1], Isabella S Kim[2], Richard M White[2,3], Itai Yanai[1]

[1]NYU School of Medicine, Institute for Computational Medicine, New York, NY, [2]Memorial Sloan Kettering Cancer Center, Department of Cancer Biology & Genetics, New York, NY, [3]Memorial Sloan Kettering Cancer Center, Department of Medicine, New York, NY

Cancer progresses as an evolutionary process due the force of selective pressures acting on the body's heterogeneous cell population. To study the precise molecular mechanisms underlying this process, we use the well-studied zebrafish melanoma model to characterize the gene expression profiles at single-cell resolution. Our transgenic zebrafish develop spontaneous tumors due to dominant-negative *p53* and *mitfa*-promoter induction of hBRAF^{V600E}. Using a single zebrafish host, we can repeatedly biopsy one tumor at several time-points followed by single cell RNA-seq analysis. This approach allows us to characterize the heterogeneity of the tumor cell population as well as observe the dynamic shifts in cellular state over time. This method holds the potential to understand and characterize the divergent patterns of transcriptional states within an individual tumor, as well as the potential to converge across animals.

THE ADVENT OF AGRICULTURE SHAPED INNATE IMMUNE RESPONSES TO PATHOGENS IN HUMANS

Genelle F Harrison[1,2], Joaquin Sanz-Remon[2], Christina M Bergey[4], Anne Dumaine[2], Jean-Christophe Grenier[2], Vania Yotova[2], Lluis Quintana-Murci[5], George H Perry[4], Luis B Barreiro[2,3]

[1]McGill University, Department of Human Genetics, Montreal, Canada, [2]Sainte-Justine Hospital Research Centre, Department of Genetics, Montreal, Canada, [3]University of Montreal, Department of Pediatrics, Montreal, Canada, [4]Pennsylvania State University, Departments of Anthropology and Biology, State College, PA, [5]Institut Pasteur, Département Génomes et Génétique, Paris, France

The shift of societies from a hunter-gatherer to agricultural method of resource accumulation in Africa is considered to have facilitated the emergence of many problematic pathogens. The extent to which hunter-gatherers and agricultural population diverge in their immune response has not been evaluated, nor have we gauged the role of selection in contributing to these differences. In this study, we collected peripheral blood mononuclear cells (PBMCs) from both hunter-gatherer (Batwa) and agricultural (Bakiga) populations in Uganda. We stimulated the PBMCs using viral (Gardiquimod–GARD) and bacterial (lippopolysacharide– LPS) ligands to mimicking infection and looked for a divergence in transcriptional regulation of innate immune response. We evaluated transcriptional differences between the Batwa and Bakiga populations (PopDE), and found 1,664 PopDE genes that differed in their overall expression for LPS and 2,242 PopDE for GARD. Among these PopDE genes we found an increase in the anti-viral activity in the Batwa population with the increased expression of genes in interferon (IFN) pathways. We next mapped expression quantitative trait loci (eQTL) for 1,097 genes. We show that genes with cis-eQTL are enriched among PopDE genes, suggesting that a significant fraction of transcriptional differences are genetically controlled. The higher expression of genes in the IFN-pathways found in the Batwa cannot be explained by cis genetic regulatory variants suggesting that either a trans eQTL or non-genetic factors are responsible for the increased activity of anti-viral responses. Finally, we show that several of the eQTL driving population differences in immune regulation have been targeted by recent positive selection.

CRISPR-MEDIATED ISOLATION OF SPECIFIC MEGABASE SEGMENTS OF GENOMIC DNA

Pamela E Bennett-Baker, Jacob L Mueller

University of Michigan Medical School, Human Genetics, Ann Arbor, MI

Throughout eukaryotic genomes there are megabase-sized regions of complex genomic structures harboring genes with important biological functions. In many cases, the reference genome sequences of these regions are incompletely assembled owing to their highly repetitive nature and huge size. Moreover, the inter-individual and inter-species variation of these regions is highly polymorphic on many size-scales, indicating a plethora of undiscovered structural and functional variation. In the few cases where the DNA sequence of such regions has been accurately determined, important insights into processes such as immunity (e.g. immunoglobulin heavy-chain locus) and reproduction (e.g. human Y chromosome) are revealed. Nevertheless, the methods involved are labor intensive and expensive, requiring whole genome clone libraries and haplotype-specific iterative mapping and sequencing. Here we describe a strategy to overcome these challenges, CISMR (CRISPR-mediated isolation of specific megabase-sized regions of the genome), enabling us to perform targeted explorations of megabase-sized segments of the genome. By designing custom pairs of specific, single guide RNAs to flanking sequences, we excise megabase-size regions from the whole genome in vitro, using the specificity of CRISPR enzymology. The DNA segments are isolated by pulsed-field gel electrophoresis (PFGE), purified, and directly sequenced using standard techniques. Resulting sequencing libraries show >100-fold enrichment of the targeted megabase-sized regions. CISMR combines the specificity of in vitro CRISPR with the sensitivity of next-generation sequencing, for a more direct, targeted, and affordable strategy to isolate and sequence complex and repetitive, megabase-sized regions of the genome to assess their biological significance.

DETECTING POLYGENIC ADAPTATION USING GWAS DATA

Claude Bherer[1], Fernando Racimo[1], Joseph K Pickrell[1,2]

[1]New York Genome Center, New York, NY, [2]Columbia University, Department of Biological Sciences, New York, NY

Natural selection on a polygenic phenotype may drive local adaptation via small shifts in allele frequencies at many loci. To test for this mode of adaptation in humans, we introduce a hierarchical model for detecting selection on a phenotype from changes in allele frequency at loci identified by genome-wide association studies (GWAS). Our method uses the normal approximation to genetic drift to jointly model neutrality and selection in a three-population tree, and a MCMC to estimate the strength of selection. Simulations show that our three-population test has increased power to detect selection compared to related methods. We test for polygenic adaptation in contemporary human populations using GWAS data for over 40 traits, and publicly available genomic data from 3,657 individuals from 40 populations. We confirm selection for increased height in northern European populations (and decreased height in southern European populations), and find selection signals in populations from other continents. In addition, we detect selection on other anthropometric traits, including waist-to-hip ratio and unibrow. Our study shows how to leverage results from GWAS studies to gain insights into the evolutionary history of alleles that contribute to current phenotypic variation, notably disease risk.

HORIZONTALLY TRANSFERRED GENES ARE OFTEN SHARED BETWEEN CLOSELY RELATED BACTERIAL SPECIES

Evgeni Bolotin, Ruth Hershberg

The Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Department of Genetics and Developmental Biology, Haifa, Israel

Horizontal gene transfer (HGT) is a major contributor to the evolution of bacterial gene content. In this study we used a pangenome-based approach to identify genes that were horizontally transferred into four closely related species, belonging to the Enterobacteriaceae family. This analysis enabled us to show that a surprisingly high percent of the horizontally transferred genes are shared between these species. Furthermore, we demonstrate significant differences between horizontally transferred genes that are shared between more than one species and those that are not. Specifically, acquired genes shared between the four analyzed species tend to be conserved in a larger number of additional species, and are better optimized for expression in their host genomes. Combined, these results demonstrate the existence of a large pool of horizontally transferred genes that are shared between species and that these genes display unique characteristics that differentiate them from other horizontally transferred genes.

RAPID EVOLUTION OF THE DROSOPHILA FEMALE MEIOSIS-SPECIFIC GENE *MATRIMONY*

Amanda M Bonner[1], R. Scott Hawley[1,2]

[1]Stowers Institute for Medical Research, Kansas City, MO, [2]University of Kansas Medical Center, Department of Physiology, Kansas City, KS

Sex-biased genes, especially those expressed in a single reproductive tissue, have been shown to evolve rapidly in numerous organisms. In *Drosophila melanogaster*, many such genes, particularly those with roles in reproduction, diverge rapidly due to positive selection. Here we investigate the rapid evolution of *matrimony* (*mtrm*), a gene that is essential in *D. melanogaster* females but dispensable in males. Mtrm protein is required for proper segregation of achiasmate chromosomes during the first meiotic division, where it physically interacts with and inhibits Polo kinase (Polo). *mtrm* homologs have not been identified outside the Drosophila genus, and even within the genus sequence conservation is poor. To explore the rapid divergence of *mtrm*, we expressed Mtrm proteins from 12 different Drosophila species in *D. melanogaster* females. Distantly-related Mtrm homologs are still able to both rescue the meiotic defects seen in *mtrm* mutants and to physically interact with *D. melanogaster* Polo. However, because these distant homologs are not properly degraded after meiosis, their maternal expression has a dominant-negative effect in the early embryo. These data, along with phylogeny and sequence-based analyses, suggest that different regions of the *mtrm* gene are evolving at different rates, with *mtrm*'s least-conserved regions showing some evidence of positive selection. Conversely, a small, highly-conserved region near Mtrm's N-terminus, which contains multiple phosphorylated resides required for the Mtrm::Polo interaction, can be used to identify homologs in other Dipterans.

COGG (CORRELATION OPTIMIZATION OF GENETICS AND GEODEMOGRAPHICS) REVEALS SOCIOLINGUISTIC STRATIFICATION IN INDIAN SUBCONTINENT

Aritra Bose[1,2], Daniel E Platt[2], Laxmi Parida[2], Peristera Paschou[3], Petros Drineas[1]

[1]Purdue University, Computer Science, West Lafayette, IN, [2]IBM T.J. Watson Research Center, Computational Genomics, Yorktown Heights, NY, [3]Purdue University, Biological Sciences, West Lafayette, IN

Human population genomics have revolutionized the research into many demographic aspects, such as, tracing the origin of a language group, studying migration routes across continents, cultural diffusion among migrants, investigating the relationships of endogamy and genetics, etc. To this end, there has been a lack of novel computational model which takes into account the external demographic factors (such as society, languages, etc.) that has shaped the genetic structure of a population and quantify their influences.

We propose COGG (Correlation Optimization of Genetics and Geodemographics), a novel optimization method to maximize the correlation of genetic relationships provided by the principal components, with demographic factors such as society, languages, occupation and geographical coordinates. More precisely, using prior information of external factors influencing the genetics of a population belonging to a particular region or country, then, we analytically solve an optimization problem to obtain a closed form solution for the coefficients which quantify the influence of the external factors on the genetic stratification of the population under study. Thereafter, we conduct feature selection using a greedy algorithm on the external factors to understand which features play the most important role in shaping the genetic structure. We show that COGG can also be extended to a Canonical Correlation Analysis (CCA) setup and gives similar statistically significant, high correlations.

To test our model we apply it on carefully selected samples from all parts of the Indian subcontinent and study genetic variation in an assembled data set of 1,163 individuals across 49,357 common SNPs from 77 well-defined ethno-linguistic groups in India. Our results demonstrate that COGG returns high correlations, with p-values lower than 10^{-8}. Identification of significant components such as Austro-Asiatic and Tibeto-Burmese speaking tribal nomads and Forward Caste groups in India, simplifies the complex structure of the Indian subcontinent. We identify few populations in India being closer to the Eurasians, finding a route for the dispersal of Indo-European and Dravidian languages, from Siberian populations and the Steppes into the Indian subcontinent. We highlight the strong shared ancestry along the hierarchical social structure and the demographic history of tribal groups belonging to different language families across the country.

EVIDENCE FOR THE PARTICIPATION OF MOST TRANSCRIBED GENES IN COMPLEX TRAIT GENE NETWORKS

Evan A Boyle[1], Yang I Li[1], Jonathan K Pritchard[1,2,3]

[1]Stanford University, Genetics, Stanford, CA, [2]Stanford University, Biology, Stanford, CA, [3]HHMI, Chevy Chase, MD

A central goal of genetics is to understand the links between genetic variation and disease. Intuitively, one might expect disease risk to be explained by a small number of disease-causing variants that cluster in or near core genes and pathways.

However recent GWASs have revealed that most complex traits, including height and schizophrenia risk, are highly polygenic. While the strongest of these associations sometimes map to genes directly linked to the trait, the majority of association signals are found across much of the genome, including near genes with housekeeping-like functions. For example, we found that over half of the genomic SNPs are in high linkage disequilibrium with a SNP that has an estimated effect of increasing height by an average of 0.145mm.

To better understand how these widespread signals contribute to complex traits, we focused on three diseases for which causal cell-types are relatively well defined: schizophrenia, Crohn's disease, and rheumatoid arthritis. Again, we found that association signals were widely dispersed. We further found that the causal signal was present sometimes exclusively in regions marked by active chromatin in the relevant cell types (45, ~100 and ~100%), but vastly depleted or absent from regions that are generally inactive across cell-types.

While variation in cell type-specific gene networks contributes to complex disease risk, we show evidence that genes with housekeeping-like functions cumulatively account for a greater fraction of total SNP heritability. As expected, we found that relevant gene sets exhibited the greatest enrichment in trait heritability. However, we also observed a strong linear relationship between the size of the gene sets and the proportion of heritability they explained, further supporting the hypothesis that most if not all transcribed genes in the relevant cell-type(s) contribute to disease risk.

Together, these findings imply a need for rethinking models of complex traits. We propose that gene regulatory networks are sufficiently interconnected for all genes expressed in disease-relevant cells to be liable to affect the functions of core disease-related genes. Consequently, the bulk of the genetic effects on disease are mediated through genes without any direct relationship to disease function, and variation in non-disease genes previously thought to be innocuous may in fact drive complex disease risk in human populations.

CHARACTERIZING CAUSAL *CIS*-REGULATORY VARIANTS USING COMPUTATIONAL APPROACHES AND CRISPR/CAS9 GENOME EDITING

Margot K Brandt[1,2], Ana Vasileva[1,2], Tuuli Lappalainen[1,2]

[1]New York Genome Center, New York, NY, [2]Columbia University, Department of Systems Biology, New York, NY

Expression quantitative trait loci (eQTL) and splicing quantitative trail loci (sQTL) studies have identified thousands of common human genetic variants associated with gene expression levels and splicing patterns, respectively. Because of linkage disequilibrium, the true causal variants and therefore the mechanisms underlying the effect on gene expression or splicing have yet to be determined at many loci. We aim to both identify characteristics of causal variants genome-wide and discriminate causal variants from those in close linkage on a local scale.

We analyzed eQTLs and sQTLs from the GTEx project with a Bayesian hierarchical model fgwas for fine-mapping of causal variants accounting for functional genomic annotation. We find that eQTL variants are significantly enriched for annotations affecting both transcriptional regulation, such as promoter-associated variants, and post-transcriptional regulation, such as RNA-binding protein binding sites and miRNA target sites. Comparison of the same data analyzed with CAVIAR, another statistical method for identifying causal variants, reveals high concordance between the two methods. We also find that sQTLs are enriched in the ends of exons, canonical splice sites, and synonymous variants. Using these enriched annotations genome-wide, we are able to identify likely causal variants in linkage disequilibrium to further validate experimentally. So far, we have validated two eQTL variants for different genes by isolating monoclonal cell lines after CRISPR/Cas9 editing. We are developing a novel approach to experimentally test hundreds of common and rare putative causal regulatory variants in coding regions by pooled editing with CRISPR/Cas9 followed by multiplexed targeted DNA and RNA sequencing.

THE MOLECULAR BASIS OF THE CELLULAR TAXONOMY OF THE HUMAN BODY

Alessandra Breschi[1], Carrie A Davis[2], Sarah Djebali[3], Manuel Muñoz[1], Dmitri D Pervouchine[4], Alex Dobin[2], Jesse Gillis[2], Thomas R Gingeras[2], Roderic Guigó[1]
[1]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Bioinformatics and Genomics, Barcelona, Spain, [2]Cold Spring Harbor Laboratory, Functional Genomics, Cold Spring Harbor, NY, [3]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosane, France, [4]Skolkovo Institute of Science and Technology, Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Odintsovkiy District, Russia

Organs and tissues are complex structures composed of millions of cells of diverse morphology and function. Tissue transcriptomes, thus, represent the average behaviour of genes across a highly heterogeneous collection of primary cells. How the transcriptional profiles of these cells relate to the profiles of tissues and organs is still poorly understood. Here we have monitored by RNA-Seq the transcriptome of multiple primary cells from multiple tissues. We found that their transcriptional profiles cluster into a few broad cell types: endothelial, epithelial and mesenchymal. The clustering is recapitulated using independent transcriptomic and epigenomic data, which also shows that blood cells, as well as other specialized cells, cluster separately. The tissue of origin contributes very little (<4%) to the transcriptomes of primary cells. Regulation of transcription plays the main role in defining these broad cell types, compared to post-transcriptional regulation (splicing), which plays a comparatively more important role in refining the characteristic transcriptomes of primary cells within a given type. We identified about 3000 genes specific to the these three cell types. These include a core set of transcription factors (TFs), showing strong co-expression patterns, and thus likely candidates to drive cell type specificity. Cell type specific genes are enriched for motifs of TFs specific to the same cell type. Cell type specific genes are mostly a vertebrate innovation, appearing early in the evolution of this lineage, with epithelial specific genes being the most evolutionary dynamic. We employed a method to estimate the proportion of cells from each type in a given tissue from gene expression values in that tissue in the GTEx collection of tissue transcriptomes. We found that, although we characterized a very limited number of primary cells from a limited number of tissues, the three basic cell types capture a large proportion (>70%) of cellular composition of many human tissues. Through the analysis of the GTEx catalogue of histological images and the associated annotations, we show that our inferred cellular composition precisely defines tissue type and captures morphological heterogeneity in the tissue samples. We identified changes in the cell type composition occurring with age and sex in a few tissues. We found that departures from the normal cellular composition of tissues correlate with histological phenotypes associated to diseases. Alterations of the cellular composition are particularly relevant in cancer, where they can be even associated to different stages of disease progression. The collection of primary cells transcriptomes produced here is a unique resource to understand tissue biology, serving as interface between tissue and single cell transcriptomics.

RARE VARIANTS AND PARENT-OF-ORIGIN EFFECTS ON WHOLE BLOOD GENE EXPRESSION ASSESSED IN LARGE FAMILY PEDIGREES

Andrew A Brown[1], Ana Viñuela[1], Angel Martinez-Perez[2], Nikolaos Panousis[1], Olivier Delaneau[1], Andrey Ziyatdinov[2], Maria Sabater-Lleal[3], Anders Hamsten[3], Juan C Souto[2], Alfonso Buil[1], Jose M Soria[2], Emmanouil T Dermitzakis[1]

[1]University of Geneva Medical School, Genetic Medicine & Development, Geneva, Switzerland, [2]Sant Pau Hospital, Barcelona, Spain, [3]Karolinska Institute, Medicine, Stockholm, Sweden

Studying genetic effects on gene expression in related individuals provides insights inaccessible when using unrelated individuals, such as heritability estimates, rare (in population) regulatory variants commonly observed in the pedigree, imprinting, and parent-of-origin effects. We recruited 935 individuals from 35 pedigrees, with an average of 27 individuals per pedigree and a total of 8654 related pairs as part of the GAIT2 study. In addition to gene expression, quantified using whole blood, blood cell count and extensive phenotypic information was available.

To better understand the genetic regulation of expression, we first estimated median heritability of expression as 0.22; this is larger than the proportion of variance in expression explained by cell count composition (median 0.17). Furthermore, we identified 11,297 eQTLs (FDR<0.01, $\pi 1 = 0.98$) using a cis association mapping, with variance components to consider the familial structure in the data. We tested whether the eQTLs discovered could be rare in the general population, examining the MAF of the lead variants in similar populations of 1000 Genomes. Compared to eQTLs from the Depression Genes and Networks study with whole blood expression from unrelated individuals, we see an excess of variants with MAF < 0.01 (9.6% eQTL compared to 0%, median MAF is 0.11 in GAIT2, 0.27 in DGN).

Finally, we used the trios within the pedigrees to look for parent-of-origin effects on regulatory variants. We performed a cis scan to find variants where the effect of the reference allele in heterozygotes depended on whether it was maternally or paternally inherited (called here parent-of-origin in expression QTL or poeQTL). We found 12 significant poeQTLs (FDR<0.05). Six of these affect known imprinted genes, implying a cis-eQTL whose effect is masked on one haplotype. However, for two of the remaining six genes (IFITM3 and C3AR1) measures of allelic ratios were also available, and showed expression of both parental alleles. This suggests these poeQTL act by a different mechanism which we intend to interrogate further, integrating our data and reference chromatin resources.

Both rare variants and parent-of-origin genetic effects have been shown to be relevant for human disease, studies such as this allow a deeper understanding of their action and implications at the cellular level.

EPIGENETIC FINE-MAPPING OF CARDIOVASCULAR DISEASE LOCI IN THE LIVER

Minal Caliskan[1], YoSon Park[1], Marco Trizzino[1], Julian Segert[1], Evanthia Pashos[1], Kiran Musunuru[2], Daniel Rader[1,2], Barbara Engelhardt[3], Christopher Brown[1]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]University of Pennsylvania, Medicine, Philadelphia, PA, [3]Princeton University, Computer Science, Princeton, NJ

Genome-wide association studies (GWAS) have identified more than 200 loci that contribute to cardiovascular disease risk (CVD). As with other complex phenotypes, the majority of the heritability of CVD risk lies within the noncoding regions of the genome. This has led to the hypothesis that the causal variants at GWAS associated loci lead to changes in local gene expression. As a result of linkage disequilibrium and the fact that cis-regulatory elements (CREs) may target genes over large distances, it is often unclear which variant or gene affects disease risk. However, their identification will improve understanding of disease etiology and identify targets for novel therapeutic development. Recent work has demonstrated that histone modification state data can be used to identify CREs harboring disease-causal variants. Existing studies have focused on easily ascertained cell types, while the liver, which plays a critical role in regulating cholesterol and lipid metabolism, and where many CVD associated variants likely affect gene expression, has remained understudied. To identify the specific variants and genes that affect CVD risk, we have deeply phenotyped liver biopsies and iPSC derived hepatocytes from more than 400 donors, collecting RNA-seq along with histone modification and transcription factor ChIP-seq data. We have used these data to identify thousands of genetic variants associated with allele-specific transcription factor binding, histone modification, gene expression, and splicing. Comparison to data from the GTEx and Roadmap Epigenomics projects demonstrate that many of these associations are specific to the liver. We demonstrate that multi-phenotype molecular trait mapping improves statistical power to detect associations and results in improved resolution at identified loci. We have integrated these data with CVD GWAS data using a novel multi-phenotype causal inference framework based on Mendelian randomization to predict the precise variants, CREs, and genes that underlie CVD risk. Using a combination of massively parallel reporter assays, genome-edited stem cells, CRISPR interference, and in vivo mouse models, we establish rs2277862-CPNE1, rs10889356-ANGPTL3, rs10889356-DOCK7, and rs10872142-FRK as causal SNP-gene sets for CVD. These results demonstrate that a molecular trait mapping framework can rapidly identify causal genes and variants contributing to complex human traits and demonstrates that, at many GWAS loci, candidate genes have been falsely implicated based on proximity to the lead SNP.

DYNAMIC HYPER EDITING UNDERLINES TEMPERATURE ADAPTATION IN DROSOPHILA

Ilana Buchumenski[1], Osnat Bartok[2], Varun Pandey[2], Reut Ashwall-Fluss [2], Hagit Porath[1], Erez Y Levanon[1], Sebastian Kadener[2]

[1]Bar Ilan University, The Mina and Everard Goodman Faculty of Life Sciences, Ramat Gan, Israel, [2]The Hebrew University of Jerusalem, Biological Chemistry Department, Silberman Institute of Life Sciences, Jerusalem, Israel

In *Drosophila*, A-to-I editing is highly prevalent in the brain and mutations in the editing enzyme dADAR correlate with specific behavioral defects. As editing sites are usually engaged in secondary structures, temperature is predicted to impact the level and nature of editing.

Here we demonstrate a role for ADAR in temperature adaptation in *Drosophila*. Briefly, we found that despite the higher level of editing at lower temperatures, there are more editing sites at 29°C. This is due to a less specific activity of ADAR at this temperature, which edits sites which are less evolutionary conserved, more disperse, less committed in secondary structures and more likely to be located in exons. These results strongly support the notion that at 29°C, RNA editing is less deterministic and might even have deleterious effects. Interestingly, hypomorph mutants for ADAR display a weaker transcriptional response to temperature changes. In addition, and in agreement with the differences on the head transcriptome, ADAR Hypomorph flies display a highly abnormal behavioral response while adapting to temperature changes. In sum, our data shows that ADAR is essential for proper temperature adaptation, a key behavior trait, which is essential for the survival of flies in the wild.

THE MICROBIOME OF THE FEMALE REPRODUCTIVE TRACT AND PREGNANCY

Gregory A Buck, Vaginal Microbiome Consortium at VCU

Virginia Commonwealth University, Microbiology and Immunology, Richmond, VA

The microbiome of the female reproductive tract has a major impact on women's reproductive health and well-being, including but not limited to health during pregnancy and its adverse outcomes including preterm birth and still birth. Bacterial vaginosis, with an obvious yet undefined microbial etiology, has a prevalence of up to 30% and carries a risk for adverse outcomes of pregnancy. Over 10% of pregancies terminate prematurely and some demographic groups; e.g., African Americans, experience a significantly higher incidence. The estimated annual costs associated with preterm birth in the US alone exceed $25 billion. The Vaginal Microbiome Consortium at VCU and its collaborators, including the Global Alliance to Prevent Prematurity and Stillbirth based at Seattle Children's Hospital, has collected cross-sectional samples from over 6,000 women, over 1,000 of whom were pregnant in the Vaginal Human Microbiome Project (VaHMP), and longitudinal samples from over 1,500 pregnant women in the Multi-Omic Microbiome Study-Pregnancy Initiative (MOMS PI). Samples include cervical, vaginal, buccal, anal and skin swabs, blood and plasma, and urine from pregnant and non-pregnant women, placenta, cord and cord blood taken at birth, and buccal, meconium, skin and first stool samples from neonates. These samples (>200,000) have been stabilized and archived in the Research Alliance for Microbiome Science (RAMS) Registry at VCU for multi-omic analyses.

We have analyzed >10,000 of these samples for microbiome profiles, several thousand for cytokine and lipidome profiles, and >500 by metagenomic/ metatranscriptomic sequencing. Additional analyses, including metabolomic and proteomic analysis are ongoing.

Our results confirm a uniquely complex microbiome in the female reproductive tract that shows racial biases, is altered during pregnancy, and is impacted by environmental and clinical factors. Multi omic analysis of our preliminary data shows correlations between multi omic profiles; i.e., taxonomic, cytokine, and lipidome profiles, and clinical observations. These results are altering the traditional view of women's vaginal and reproductive health, leading to identification mechanisms of pathogenesis leading to adverse health and pregnancy outcomes, permitting earlier prediction of adverse reproductive events, and permit early intervention in disease.

This work was supported by grants from the NIH Common Fund Human Microbiome Project program (1UH2/UH3AI08326 and 8U54HD080784 and to G. Buck, K. Jefferson and J. Strauss). We thank members of the Vaginal Microbiome Consortium at VCU (vmc.vcu.edu), and the Global Alliance for the Prevention of Prematurity and Stillbirth for their participation in these projects.

ELUCIDATING CENTROMERIC PROTEIN-DNA INTERACTIONS VIA HYBRIDIZATION CAPTURE AND MASS SPECTROMETRY

Katherine E Buxton[1], Julia Kennedy-Darling[1], Michael R Shortreed[1], Nur Zafirah Zaidan[2], Michael Olivier[3], Mark Scalf[1], Rupa Sridharan[2,4], Lloyd M Smith[1,5]

[1]University of Wisconsin, Department of Chemistry, Madison, WI, [2]University of Wisconsin, Wisconsin Institute for Discovery, Epigenetics Theme, Madison, WI, [3]Texas Biomedical Research Institute, Department of Genetics, San Antonio, TX, [4]University of Wisconsin, Department of Cell and Regenerative Biology, Madison, WI, [5]University of Wisconsin, Genome Center of Wisconsin, Madison, WI

The centromere is the chromosomal locus where the kinetochore forms, and is critical for ensuring proper segregation of sister chromatids during cell division. A substantial amount of effort has been devoted to understanding the characteristic features and roles of the centromere, yet some fundamental aspects of the centromere, such as the complete list of elements that define it, remain obscure. To move towards a better understanding of the structure and function of the centromere, it is essential to determine its constituent molecular components, such as proteins and nucleic acids. It has long been known that human centromeres include a highly repetitive class of DNA known as alpha satellite, or alphoid, DNA. We present here the first DNA-centric examination of human protein-alpha satellite interactions, employing an approach known as HyCCAPP (hybridization capture of chromatin-associated proteins for proteomics) to identify centromeric proteins in a human cell line. Using HyCCAPP, cross-linked alpha satellite chromatin was isolated from cell lysate, and captured proteins were analyzed via mass spectrometry. After comparing to proteins identified in whole cell lysate and control pulldown experiments, a total of 81 proteins were identified as specifically enriched at alphoid DNA. This list included many known centromere-binding proteins in addition to a number of novel alpha satellite-binding proteins, such as LRIF1, a heterochromatin-associated protein. The ability of HyCCAPP to reveal both known as well as novel centromeric protein-DNA interactions highlights the validity and utility of this approach.

JBROWSE AND AWS S3: CHEAP AND EASY GENOME BROWSING

Scott Cain

Ontario Institute for Cancer Research, Stein Lab, Toronto, Canada

JBrowse is a widely used GMOD (Generic Model Organism Database, http://gmod.org/wiki/JBrowse) project for displaying a variety of genomic feature data types. Here we present a method for implementing JBrowse in a very low maintenance fashion by loading the data and software for JBrowse into Amazon Web Service's S3 data storage service and serving the web pages and data directly out of S3, without the need for any other web server. We outline the methods for implementation as well as the issues that implementers may want to consider, including cost and security.

DISSECTING HUMAN AND CHIMPANZEE CEREBRAL ORGANOIDS USING SINGLE-CELL RNA-SEQ

J. Gray Camp[1], Sabina Kanton[1], Farhath Badsha[2], Marta Florio[2], Ben Vernot[1], Wieland Huttner[2], Svante Pääbo[1], Barbara Treutlein[1,2]

[1]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany, [2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

Cerebral organoids have emerged as powerful models of human brain development, and offer the potential to study uniquely human brain evolution. However, the extent to which cerebral organoid systems recapitulate fetal gene expression networks remains unclear. Here we use single-cell RNA sequencing (scRNA-seq) to dissect and compare cell composition and progenitor-to-neuron lineage relationships in human and chimpanzee cerebral organoids and fetal human neocortex. We find that human and chimpanzee organoid cortical cells use gene expression programs remarkably similar to those of the fetal tissue in order to organize into cerebral cortex-like regions. We identify genes that are differentially expressed in human progenitors and neurons relative to chimpanzee, and highlight modern human genetic changes that can be studied in organoid cultures. More broadly, this strategy can be extended to other organoid systems modeling human and chimpanzee development and disease.

DIRECT ESTIMATION OF GERMLINE MUTATION RATE IN MOUSE LEMURS, GENUS *MICROCEBUS*

C. Ryan Campbell[1], Kelsie E Hunnicutt[1], Rachel C Williams[1], Peter A Larsen[1], Mario dos Reis[2], Anne D Yoder[1]

[1]Duke University, Biology, Durham, NC, [2]Queen Mary University of London, School of Biological and Chemical Sciences, London, United Kingdom

Germline mutation rate is a vital evolutionary statistic and a well-studied metric among model organisms. Mutations in the germline are the raw material on which evolution acts, and knowing the rate at which they occur is vital to understanding how evolution acts at both long and short timescales. While these rates are well documented among model organisms, such as humans and mice, there is far less known outside of these key species. As the use of genomic technologies spread from model to non-model organisms it is crucial to assess underlying metrics such as mutation rate in novel species, both to learn about how they vary across the tree of life and to correctly place genomic findings in an evolutionary context. The mouse lemurs, genus *Microcebus*, are 25 species of small nocturnal primates that constitute a 10 million year old evolutionary radiation on the island of Madagascar. The accurate estimation of their mutation rate from deep sequencing of family pedigrees at the Duke Lemur Center shows how the rate in lemurs compares with the known rate of model organisms with respect to phylogenetic position, body size, life span, and effective population size. Furthermore, the mutation rate affects results when studying the evolutionary relationships between these species in the wild. Accurately dating these evolutionary relationships is crucial to understanding the species interactions and the evolutionary and ecological history of Madagascar.

FGFR2 RISK SNPs CONFER BREAST CANCER RISK BY AUGMENTING ESTROGEN RESPONSIVENESS

Thomas Campbell[1], Mauro Castro[2], Bruce Ponder[1], Kerstin Meyer[1]

[1]University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom, [2]Federal University of Paraná (UFPR), Bioinformatics and Systems Biology Lab, Curitiba, Brazil

Extensive epidemiological studies have consistently demonstrated that there are associations between variants in the second intron of the FGFR2 gene and breast cancer risk, with risk variants conferring increased risk for estrogen receptor-positive (ER$^+$) disease. Due to the relatively common nature of the risk alleles at this locus, it is believed that the locus contributes to up to 16% of all breast cancers, suggesting a significant disease burden due to FGFR2. However, the mode of action of the risk variants has remained controversial. Here, we employ a systems biology approach, in which we use gene set enrichment analysis (GSEA) on gene expression data, to demonstrate that signalling via FGFR2 counteracts the estrogen response in ER$^+$ breast cancer cells. In the presence of estrogen, the estrogen receptor (ESR1) regulon is in an active state. However, signalling via FGFR2 is able to reverse the activity of the ESR1 regulon. This effect is seen in multiple distinct FGFR2 signalling model systems, across multiple cell lines, and is dependent on the presence of FGFR2.

Increased exposure to estrogen has long been associated with an increased risk of developing breast cancer. We therefore hypothesised that the FGFR2 risk variants, which have been shown to interact with the FGFR2 promoter, should reduce FGFR2 expression and subsequent signalling. Indeed, transient transfection experiments assaying the three independent variants of the FGFR2 risk locus (rs2981578, rs35054928 and rs45631563) in their normal chromosomal context show that the single nucleotide polymorphisms (SNPs) map to transcriptional silencer elements and that the risk alleles augment silencer activity to reduce FGFR2 gene expression. This results in an increased cellular response to estrogen, and thus, breast cancer risk.

Having demonstrated that reduced FGFR2 signalling results in an increased estrogen response in ER$^+$ breast cancer cells, we hypothesised that inhibition of FGFR2 should sensitise breast cancer cells to anti-estrogen therapy. Indeed, treatment of ZR751 cells with two different FGFR2 inhibitors, AZD4547 and PD173074, enhances the anti-proliferative effect of tamoxifen and fulvestrant on the cells.

We therefore propose a molecular mechanism by which FGFR2 can confer increased breast cancer risk that is consistent with estrogen exposure as a major driver of breast cancer development. Our findings demonstrate a mechanistic follow-up study from genome wide association study (GWAS) data, and may have implications for the clinical use of FGFR2 inhibitors.

NEXT-GENERATION MAPPING: APPLICATION TO CLINICALLY RELEVANT STRUCTURAL VARIATION ANALYSIS

Alex Hastie, A Pang, J Lee, E Lam, T Anantharaman, W Andrews, M Saghbini, Han Cao

Bionano Genomics, Inc, R & D, San Diego, CA

Next-generation mapping (NGM) from Bionano Genomics allows researchers to interrogate genomic structural variations (SVs) in the range of one kilobase pairs and above. It uses extremely long range information to span interspersed and long tandem repeats making it suitable for elucidating the structure and copy number of complex regions of the human genome, such as loci with complex pseudogene and paralogous gene families. Because NGM is a de novo process and because molecules analyzed are longer than almost all genomic repeats, NGM is able to detect a wide range of SVs including insertions of novel sequence, tandem duplications, interspersed duplications, deletions, inversions and translocations, a range of SV types detectable by NGM alone. Because of the high speed and comprehensiveness of the SV types detected, NGM is increasingly being applied to the analysis of clinical genomes for the detection of SVs potentially involved in disease pathogenesis. We present several in silico and biological validation experiments to demonstrate the sensitivity and specificity of NGM for insertion, deletion and translocation SVs and compare it to benchmark studies using short read and long read sequencing. We also show the application of NGM to studying somatic variation in a breast cancer cell line, finding hundreds of somatic structural variations. Finally, we applied NGM to several leukemia patient samples to find more than 50 cancer related SVs in each patient. NGM is a fast and cost effective method for detection of a broad range of traditionally refractory SVs across the genome.

EVOLUTIONARY TURNOVER OF REGULATORY ELEMENTS ACTIVITY IN MAMMALS

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

[1]University of Cambridge, Gurdon Institute, Cambridge, United Kingdom, [2]University of Heidelberg, ZMBH - Center for Molecular Biology, Heidelberg, Germany

Promoters and enhancers control gene transcription, and their specific activities therefore underlie organismal development, physiology and behaviour. Although these regulatory elements have long been distinguished from each other on the basis of their function, recent work highlighted similarities in their chromatin composition and functionality. The common architecture of regulatory elements suggests that subtle inheritable alterations in their chromatin makeup and sequence context might underlie evolutionary changes in their activity. Here, through the analysis of chromatin and transcriptional profiles across multiple mammalian organs and tissues, we provide support for this hypothesis by detecting hundreds of regulatory elements showing signatures of functional turnover (repurposing) within the rodent and primate lineages. We determined the directionality of several turnover events and found that the evolution of enhancers into promoters represents the most common kind of repurposing event. Finally, we found that sequence and motif composition of repurposed regulatory elements distinguish them from other regulatory regions, and we observed that evolutionary changes in the regulatory activity correlated with changes in GC content and in the U1-PAS frequencies at these loci. Overall, our work suggests functional repurposing of mammalian regulatory elements as a potential mechanism underlying evolutionary changes in regulatory networks.

FANTOM6 REVEALS BROAD FUNCTIONAL PROPERTIES OF LONG NON-CODING RNAs

Michiel de Hoon, Jay Shin, Jordan Ramilowski, Saumya Agrawal, Chi-Wai Yip, Chung-Chao Hon, Masayoshi Itoh, Ken Yagi, Yasushi Okazaki, Piero Carninci

RIKEN Center fir Life Science Technologies, Division of Genomic Technologies, Yokohama, Japan

FANTOM (Functional ANnoTation Of the Mammalian genome) is an international research consortium aiming at a comprehensive identification and annotation of mammalian transcripts. The latest FANTOM catalog estimates 124,245 expressed loci based on a compilation of five transcript assemblies together with CAGE to establish the precise 5' end of transcripts (Resource to FANTOM CAGE-Associated Transcripts (CAT): http://fantom.gsc.riken.jp; and Hon C.C. et al.: "An atlas of human long non-coding RNAs with accurate 5' ends". Nature, in press (2017)). The vast majority of these loci encode long non-coding RNAs (lncRNAs) and the genome-wide association studies as well as expression QTL analysis suggested biological significance of these lncRNAs. However, a large-scale functional genomics screening effort is required to functionally annotate the exact role of each lncRNA.

In FANTOM6, we use high-throughput knockdown strategies to probe the function of lncRNAs, followed by transcriptome profiling using CAGE to assess the molecular phenotype. An unbiased approach was used to select lncRNAs for knockdown in human dermal fibroblast, including both known and novel lncRNAs, to generate a broad catalog of lncRNA function. Importantly, our knockdown strategy specifically targets the transcript rather than the genomic locus, allowing us to directly dissect the function of the lncRNA itself. Additionally, by profiling the molecular phenotype of transcriptome changes upon knockdown, we are able to investigate the broader functions of lncRNAs without restricting ourselves to a particular cellular phenotype.

Overall, about 70% of the knockdowns resulted in a distinct transcriptome change, suggesting that the majority of lncRNAs may be functional. Integration of the knockdown data with genomic, conservation, and expression features of lncRNAs, as annotated in the FANTOM CAT, revealed that in particular polyadenylated and dermal fibroblast-specific lncRNAs are were more likely to be functional. Importantly, the impact of lncRNA knockdown on the transcriptome was not significantly correlated with their expression level, indicating that lowly expressed lncRNAs are as functional.

EPISTASIS WITHIN GENES SHAPES HUMAN GENETIC VARIATION AND DISEASE RISK

Stephane E Castel[1,2], Alejandra Cervera[1], Tuuli Lappalainen[1,2]

[1]New York Genome Center, New York, NY, [2]Columbia University, Department of Systems Biology, New York, NY

Non-additive interaction between genetic variants, known as epistasis, has been hypothesized to contribute to phenotypic diversity and variant penetrance, but examples have been largely restricted to model organisms. Searching for effects of epistasis between genes is statistically challenging due to the large search space and many confounding factors, however searching for epistasis within genes, including cis-regulatory regions, is a more tractable approach that has a direct mechanistic interpretation. In this work, we use human population scale genetic and functional genomic data to demonstrate that epistasis between variants affecting a single gene, which we call haplotype epistasis, has shaped the human genome and plays a role in disease risk.

We primarily studied how haplotype epistasis can modify the penetrance of coding variants. We began by modeling interactions between regulatory and coding variants and found that higher frequencies of lower expressed haplotypes minimize negative epistatic interactions. Investigating allele frequencies of GTEx v6 eQTLs showed a significant reduction in the frequency of gain versus loss of expression regulatory variants ($p < 6e-8$), suggesting that purifying selection has acted to minimize negative haplotype epistasis. Supporting this, we found using GTEx allele specific expression (ASE) data that putatively deleterious coding variants are more often found on the lesser expressed haplotype ($p < 3e-3$). Furthermore, using GTEx exon inclusion measurements we found that putatively deleterious coding variants are enriched in exons that are excluded by splicing ($p < 1e-7$). Finally, we used our method phASER on GTEx individuals to phase rare coding variants with expression and splice QTL variants using WGS and RNA-seq reads and found that putatively deleterious coding variants are enriched on haplotypes with lower predicted expression and splice inclusion.

Having demonstrated that at the population level purifying selection acts to reduce negative epistatic interactions we next sought to quantify its contribution to disease. Using the Simons Simplex Collection, we found evidence that in individuals with autism, rare inherited disrupting coding variants in autism associated genes are more often found in regulatory haplotype configurations that would result in negative epistatic interactions as compared to synonymous variants ($p < 1e-100$). We next used The Cancer Genome Atlas data to read back phase somatic coding variants with germline regulatory variants and found a significant enrichment of negative haplotype epistasis in tumor suppressor genes versus control genes ($p < 1e-93$). These results suggest that regulatory haplotype configuration of disease-causing rare coding variants affects their penetrance via haplotype epistasis.

Altogether, our results show that epistasis between variants on the same haplotype is relatively common, and has shaped the patterns of both coding and regulatory variation in humans. Furthermore, our results indicate that haplotype epistasis may play a significant role in disease. More broadly, our results highlight the importance of analyzing genetic variants in the context of the entire haplotype.

HETEROCHROMATIN TURNOVER AMONG GREAT APES: SPECIES AND GENDER DIFFERENCES

Monika Cechova[1], Robert S Harris[1], Francesca Chiaromonte[2], Kateryna D Makova[1]

[1]Penn State, Department of Biology, University Park, PA, [2]Penn State, Department of Statistics, University Park, PA

Heterochromatin comprises intriguing repeat-rich and gene-deprived genomic regions that play an essential role in the cell, both mechanistically and functionally. Proper pairing of the chromosomes is ensured by the satellite repeats in centromeric arrays, while telomeric repeats protect chromosome ends during cell division. Such important processes as X-chromosome inactivation and tissue-specific gene repression are driven by heterochromatin. Moreover, even small segments where the heterochromatin is unbalanced can lead to mitotic failure and non-viable offspring in hybrids. Thus, rapid heterochromatin turnover can drive speciation. Indeed, heterochromatic repeats show remarkable diversity in composition and variation in repeat counts among species, populations, and individuals. Due to its repetitive nature, heterochromatin has been difficult to study and its substantial portions are still missing from existing reference assemblies (e.g., estimated 5-10% of the human genome). Here, using short- and long-read resequencing data, we conduct the first detailed genome-wide investigation of heterochromatin turnover among great apes. We show that each species harbors only a handful of repeats, e.g., the (GGAAT)n pentamer in human and homologous 32-mers in chimpanzee, bonobo, and gorilla. Principal Component Analysis allowed us to successfully differentiate great ape species (but not sexes) based on abundances of the 10 most common repeats. We observed small interindividual variation in repeat counts in humans, whereas large interindividual variation characterizes gorillas and chimpanzee subspecies. We identified eight male-biased repeats that comprise up to 3.8% of the male genomes. To study repeat length distributions, we developed a novel algorithm, NoisyRepeatFinder that can analyze repeats in long and noisy PacBio reads. Lastly, we found that, for most linkage functions, unsupervised hierarchical clustering does not reproduce the expected species phylogeny, illustrating a remarkably high tempo of heterochromatin turnover in great apes that might have contributed to their speciation.

THE BURDEN OF SELECTION AT HLA GENES OVER THEIR GENOMIC NEIGHBORHOOD.

Jonatas Cesar[1], Fabio Mendes[2], Diogo Meyer[1]

[1]University of São Paulo, Departament of Genetics and Evolutionary Biology, São Paulo, Brazil, [2]Indiana University, Departament of Biology, Bloomington, IN

The extreme polymorphism of HLA genes results from balancing selection and reflects the importance of HLA proteins in responding to a plethora of pathogenic attacks. As shown by Lenz et al 2016 (https://doi.org/10.1093/molbev/msw127), the strong selection on HLA genes also affects linked neighboring loci, increasing the frequency of deleterious variants in the MHC region. Nevertheless, some questions regarding the effect that selection on HLA genes has over its neighborhood remain open: a) which regions within the MHC are more susceptible to the increased load? b) how much more frequent are deleterious mutations in the MHC with respect to the rest of genome? c) is there an increase in the number of segregating deleterious sites around HLA genes? d) how does the influence of the HLA genes over its neighborhood vary among populations? To answer these questions we defined the peri-HLA as the region in the neighborhood of HLA genes, within which diversity is above the genomic average, consistent with the effects of linked selection. Using the 1000 genomes data and a gene-based bootstrap methodology for statistical testing, which accounts for differences in the site frequency spectrum of HLA region and the remainder of the genome, we showed that in the peri-HLA the deleterious variants are on average 2.9 times more frequent in comparison with genome, for all continental populations (AFR, EUR, EAS, SAS). In other regions of the MHC, which do not flank selected loci, there is no significant increase in the frequency of deleterious variants. We also show that the total number of deleterious variants is significantly greater in the peri-HLA in comparison with genomic control, and that the African population is the one with the largest number of deleterious variants.

UNEXPECTED DISTRIBUTION OF GLUCOSE-6-PHOSPHATE DEHYDROGENASE DEFICIENCY HAPLOTYPES IN AN ADMIXED MALAGASY POPULATION.

Ernest R Chan[1], Rosalind E Howes[2,3], Tovonahary Rakotomanga[4], Seth T Schulte[3], Arsene C Ratsimbasoa[4], Peter A Zimmerman[3]

[1]Case Western Reserve University, Institute for Computational Biology, Cleveland, OH, [2]University of Oxford, Oxford Big Data Institute, Oxford, United Kingdom, [3]Case Western Reserve University, Center for Global Health and Diseases, Cleveland, OH, [4]Ministry of Health, National Malaria Control Programme, Antananarivo, Madagascar

The human X-chromosome-linked *G6PD* gene is characterized by hundreds of single nucleotide sequence polymorphisms (SNPs), over 200 of these influence G6PD enzyme deficiency. The G6PD deficiency polymorphisms have most likely arisen under the selective pressure of malaria. Geographically dispersed populations have experienced malaria independently over hundreds of generations and different polymorphisms underlying this important enzyme deficiency exhibit unique origins in different malaria-endemic regions, Africa and Southeast Asia, in particular. The African-Austronesian origins of the Malagasy population suggest that a complex mixture of genetic variation will be present in the *G6PD* gene and across the human genome. This would have implications for the widespread use of *Plasmodium vivax* therapy using the drug, primaquine, to which G6PD deficient people experience severe hemolytic anemia. Two study regions in the *P. vivax*-endemic western foothills region of Madagascar were selected for G6PD screening. Both the qualitative fluorescent spot test phenotyping and *G6PD* genotyping were used to screen all participants. A total of 365 unrelated male volunteers from the Tsiroanomandidy, Mandoto, and Miandrivazo districts of Madagascar were screened and 12.9% were found to be phenotypically G6PD deficient. Full gene Illumina sequencing of 95 samples identified 16 SNPs, which were integrated into a genotyping assay. Genotyping (n=291) found one individual diagnosed with the severe *G6PD Mediterranean*^{C563T} mutation, while the remaining G6PD deficient samples had mutations of African origin, *G6PD A-* and *G6PD A*. Interestingly, despite evidence in these 291 individuals of Austronesian genetic markers in genes encoding proteins localized to the red blood cell (e.g. Duffy blood group) and in cytochrome P450 2D6, we noted a complete absence of *G6PD* deficiency alleles shown to be common in Austronesian populations. Our results suggest an incompatibility of the African and Austronesian G6PD deficiency alleles.

INTEGRATED METADATA-DRIVEN ACCESS OF ENCODE, MODENCODE, REMC, GGR AND MODERN DATA THROUGH A COMMON PORTAL

Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Marcus Ho, Aditi K Narayanan, Kathrina C Onate, J Seth Strattan, Forrest Tanaka, Ulugbek Baymuradov, Christopher Thomas, Cricket A Sloan, Benjamin C Hitz, J Michael Cherry

Stanford University, Department of Genetics, Stanford, CA

The efforts of large-scale NIH-funded projects such as the Encyclopedia of DNA Elements (ENCODE), its model organism corollary (modENCODE), and the Roadmap Epigenomics Mapping Consortium (REMC) have resulted in the accumulation of genome-wide maps of candidate functional sequence elements and epigenetic marks in a large variety of cells in human and other model organisms. Consideration of the provenance of experimental reagents and transparency of computational analyses is crucial to the interpretation and comparison of these data. Tracking this information consistently across different biochemical assays employed in thousands of experiments across hundreds of cell and tissue types is particularly challenging at the scale of these projects.

The ENCODE Data Coordination Center (DCC) has developed a flexible and rich data model to capture information such as key experimental variables, details of experimental and analysis methods, what software and pipelines were used to produce which files for the ENCODE project, and calculated quality metrics, known collectively as metadata. This data model has since been extended to additionally integrate curated data and metadata from related projects such as REMC, modENCODE, Model organism Encyclopedia of Regulatory Networks (modERN) and the Genomics of Gene Regulation (GGR). The data corpus currently encompasses over 13,000 experiments and is still growing steadily. Access is freely enabled through the ENCODE portal (https://www.encodeproject.org), which features a powerful faceted browsing interface, full-text search, and a REST API so users can easily search, filter, download and visualize the collection. Learn more about how to get started on the ENCODE portal here: https://www.encodeproject.org/help/getting-started/.

MACHINE LEARNING STRATEGIES TO IDENTIFY HIGH CONFIDENCE STRUCTURAL VARIANTS IN HUMAN GENOME REFERENCE MATERIALS

Lesley M Chapman[1], Justin Zook[1], Noah Spies[1,2], Fritz Sedlazeck[3], Peyton Greenside[2], Marc Salit[1,2], The Genome In a Bottle Consortium [1,2]

[1]National Institute of Standards and Technology, Genome Scale Measurements Group, Gaithersburg, MD, [2]Stanford University School of Medicine, The Joint Initiative for Metrology in Biology, Stanford, CA, [3]Johns Hopkins University, Computer Science, Baltimore, MD

Next generation sequencing (NGS) technologies to measure DNA sequence are rapidly evolving. Clinical decision making has advanced as a result of NGS diagnostic tools, especially for small variants in non-repetitive regions of the genome. Yet, discordance exists amongst large indel and structural variant (SV) calls as a result of variance between NGS sequencing and analysis pipelines. Improvements in the accuracy of calling these difficult structural variants is needed to enable confidence in clinical decision making. Previous work in the Genome in a Bottle Consortium has used visualization, heuristics, and exploratory machine learning to form preliminary benchmark large deletion calls, but these have been limited by inaccurate breakpoints, lack of genotype information, and/or lack of insertion calls.

The central aim of the current study is to use machine learning to integrate data from multiple sequencing technologies and generate a high confidence list of large indels and SVs within NIST human genome Reference Materials that can be used as a benchmark. We use extensive whole genome sequencing from an Ashkenazim Jewish mother-father-son trio (NIST RM 8392) to develop integration methods. Members of the Genome in a Bottle Consortium generated over 300000 candidate large indels and SVs greater than 20bp in size within this trio from 30 different informatics pipelines and 5 different sequencing technologies.

We are exploring several machine learning methods to characterize each candidate SV's genotype and breakpoint accuracy and our confidence in these assertions. First, we are using dimensionality reduction methods (t-Distributed Stochastic Neighbor Embedding [tSNE]) and clustering to select an initial set of data points for label assignment. Members of GIAB will then use svviz, IGV, and other tools to visualize these points to assign a genotype, breakpoint accuracy, and confidence level to label the selected points. We will then use these labelled data points to train machine learning models - such as the semi-supervised Label Spreading method - to assign SV genotypes, their likelihoods, and choose the candidate call with the best breakpoints at each true SV site. We will evaluate these results using visualization of our diverse data types, make the benchmark calls available, and continue to improve them as new sequencing and analysis methods are developed.

CHARACTERIZATION OF MAREK'S DISEASE TUMORS FOR DRIVER MUTATIONS IN CHICKEN, AND IMPLICATIONS ON WIDESPREAD VACCINATION AGAINST ONCOGENIC VIRUSES AND VIRAL EVOLUTION

Alec Steep[1], Hongen Xu[2], Alexis Black Pyrkosz[3], William M Muir[4], Mary E Delany[5], Dmitrij Frishman[2], Hans H Cheng[3]

[1]Michigan State U., Genetics Program, East Lansing, MI, [2]Technical U. of Munich, Genome-Oriented Bioinformatics, Munich, Germany, [3]USDA, ARS, Avian Disease and Oncology Laboratory, East Lansing, MI, [4]Purdue U., Animal Sciences, West Lafayette, IN, [5]U. of California, Animal Science, Davis, CA

A major success story in veterinary medicine was the development of vaccines in the 1970s to control Marek's disease (MD), a virally-induced disease of chickens characterized by the rapid onset of CD4 T cell lymphomas. MD vaccines act as highly effective cancer vaccines, however, are "leaky" as they do not prevent birds from being infected with, replicating, or spreading the causative Marek's disease virus (MDV). As a consequence, unpredictable yet periodic outbreaks of new and more virulent MDV field strains have repeatedly arisen driving the need for improved MD vaccines. Enhancing genetic resistance is an attractive complementary control strategy and, therefore, there is a critical need to understand how MDV induces neoplastic transformation. Meq, a bZIP transcription factor and the viral oncogene, is required but not sufficient for tumor formation. Thus, we hypothesized that additional mutations in the host genome are required. To identify these somatic mutations, experimental White Leghorn chicks were challenged with pathogenic MDV, the resulting tumors and matching normal tissues collected, and subsequently characterized using a number of genomic screens. Whole genome sequence analysis of 26 MD tumors revealed ~300 unique somatic non-synonymous single nucleotide variants and indels with ~65 non-synonymous mutations per tumor. Genes frequently mutated including IKAROS family zinc finger 1 (IKZF1), a zinc-finger transcription factor associated with T-cell development and chromatin remodeling. Like Acute Lymphocytic Leukemia (ALL), MD tumors contain dominant negative somatic mutations in IKZF1 zinc-finger binding domains suggesting its role as a tumor suppressor gene. Other significantly mutated genes are found in pathways influencing cell fate (e.g., Notch) and cell survival (e.g., STAT), which were supported by RNA sequence analysis. Collectively, these data suggest MDV Meq and host IKZF1 regulate the decision between viral replication and latency, which is perturbed by somatic mutations in IKZF1 and other genes leading to tumorigenesis. This knowledge should aid genomic selection for MD resistance as well as biomedical applications of non-sterilizing vaccine strategies directed against oncogenic viruses and the role they may play in increased pathogenicity of circulating viruses.

POPULATION-SCALE SV DETECTION AND CHARACTERIZATION USING SVTOOLS

Colby Chiang[1], Haley J Abel[1], David E Larson[1], Lei Chen[1], Alexandra J Scott[1], Allison Regier[1], Indraniel Das[1], James Eldred[1], Adam Coffman[1], Abhijit Badve[1], Liron Ganel[1], Ryan M Layer[2], Susan Dutcher[1], Nathan Stitziel[1], Aaron R Quinlan[2], Ira M Hall[1]

[1]Washington University, McDonnell Genome Institute, Saint Louis, MO,
[2]University of Utah, Department of Human Genetics, Salt Lake City, UT

Structural variation (SV), including copy number variation and balanced rearrangements, comprises a large part of human genetic diversity. The contribution of SV to human disease remains unknown, however, as the full range of SV cannot be detected from the SNP microarray or exome sequencing data collected in most large-scale genetic studies to date. The current wave of population-scale whole-genome sequencing (WGS) studies (e.g., CCDG, TOPMed) provides unique opportunities, both for the characterization of the SV landscape in the healthy population and for investigation of the role of SV in common disease. To succeed, such endeavors will require new approaches to SV detection, as current algorithms scale to at most several hundred genomes.

We present svtools (https://github.com/hall-lab/svtools), a suite of tools for SV detection and characterization, scalable to tens of thousands of genomes. The svtools pipeline first generates single-sample SV calls in parallel using LUMPY, whose output describes the uncertainty in breakpoint position via probability distributions. SV calls are then merged across samples, with shared evidence in breakpoint distributions combined to improve precision. Each sample is then genotyped and annotated with read-depth information at all variants in this discovery set; the resulting multi-sample VCF describes cohort-level SV with greater sensitivity and precision than the original single-sample calls.

We demonstrate the utility of svtools through several ongoing applications for human disease genetics. As proof of principle, we present results from an initial SV callset comprising more than 5,000 deep WGS datasets from individuals of diverse ancestry. We discuss creation of an SV catalog resource to improve the understanding of the landscape of genetic variation in the healthy population, which will prove especially valuable for interpreting 'N-of-1' type variants often encountered in clinical settings and in the study of rare disease. We further demonstrate the application of these methods to large-scale human genetics projects involving eQTL discovery, causal variant prediction at published GWAS loci, and direct ascertainment of common disease associations in WGS-based studies being conducted as part of the Centers for Common Disease Genomics program. The svtools suite enables joint SV discovery and genotyping on population cohorts, thus promising a wealth of new genetic variation to aid in the understanding of human disease.

COMPREHENSIVE LONG RANGE SEQUENCING OF FULL LENGTH LONG INTERSPERSED ELEMENT-1 IN DIVERSE HUMAN POPULATIONS

Nelson T Chuang[1,2,3], Eugene J Gardner[1,2], Emma C Scott[1,2], Scott E Devine[1,2,4,5]

[1]University of Maryland Baltimore, Graduate Program in Molecular Medicine, Baltimore, MD, [2]University of Maryland School of Medicine, Institute for Genome Science, Baltimore, MD, [3]University of Maryland Medical Center, Division of Gastroenterology, Baltimore, MD, [4]University of Maryland School of Medicine, Department of Medicine, Baltimore, MD, [5]University of Maryland Medical Center, Greenebaum Comprehensive Cancer Center, Baltimore, MD

Long interspersed element-1 (L1) is a 6 kb retrotransposon that comprises approximately 17% of the human genome. It has the ability to “copy and paste” itself from one place to another in the human genome. Although the majority of known L1s are inactive due to truncation or mutations, there is a growing portion of L1s that retain its activity. These L1s are full length with both open reading frames intact. Using our Mobile Element Locator Tool (MELT) we detected 1,213 full length L1s (FL-L1) in the ~2,500 genomes of the 1000 Genome Project. Whereas some FL-L1s are found in all populations, others are population-specific, and some are even individual-specific. In order to understand differences between each FL-L1, we developed a new method of capturing interior sequences of these elements using Pacific Bioscience long range sequencing technology. With this method we have PCR-validated and sequenced 630 FL-L1s creating the largest collection of FL-L1 elements to date. The interior sequence information revealed that over 60% of FL-L1s are potentially active and that the majority are from the Ta1d subfamily. Activity of these FL-L1s in somatic tissues also requires that these elements evade somatic repression. We showed this to be tissue specific using RNA-Seq data from the Genotype-Tissue Expression (GTEx) project. Cataloging FL-L1s in human populations will allow us to create FL-L1 profiles in individuals to predict their risk of germline and somatic mutagenesis from L1 retrotransposition. This, in turn, may allow us to identify individuals with increased susceptibility to L1-mediated diseases.

SINGLE CELL TRANSCRIPTOMICS FOR CHARACTERIZATION OF COMPLEX SYSTEMS AND BIOMARKER DETECTION

Deanna M Church, Zachary Bent, Stephane Boutet, Sofia Kyriazopoulou-Panagiotopoulou, Josephine Lee, Patrick Marks, Samuel Marrs, Elliott Meer, Jeff Mellon, Luz Montesclaros, Daniel Riordan, Paul Ryvkin, Joe Shuga, Matt Sooknah, Jessica Terry, Kevin Wu, Grace X Zheng, Tarjei Mikkelsen

10x Genomics, Research and Development, Pleasanton, CA

The advent of high throughput, droplet based systems for assaying transcriptomes at the single cell resolution has revolutionized our approach to studying complex biological systems. We recently described a fully-integrated, droplet based approach, the Chromium^{TM} single cell system, that enables 3' mRNA digital counting of up to millions of single cells. High efficiency cell capture coupled with a low doublet rate (<1% per 1000 cells), facilitates the profiling of precious and rare cell populations. We have also developed an open source analysis pipeline, Cell Ranger^{TM}, that is optimized for efficient processing of sequence data. Recently, we released an interactive analysis and visualization tool call Loupe^{TM} Cell Browser.

We demonstrate the power of this system to characterize the subpopulations of the murine embryonic brain. Starting with ~1.3 million brain cells from cortex, hippocampus and ventricular zones of 2 E18 mice we generated over 100 single cell libraries, sequencing each cell to ~18k raw reads. Major neuronal and non-neuronal cell types from different brain layers were detected. Diverse, yet rare interneurons were readily detected without FACS enrichment, demonstrating the power of the single cell system in profiling complex populations and detecting rare cell types.

Recent system additions allow for the characterization of paired T cell receptor alpha and beta chains in 10s of 1000s of T cells. This application allows comprehensive immune repertoire profiling, enabling one to determine which functional subsets of T cells have undergone clonal expansion and is invaluable in areas of infectious diseases and immuno-oncology.

We wished to expand the use of this technology to the detection of biomarkers to detect and monitor disease state. We illustrate the power of this system to track the progression of diseases through comparative analysis of AML patients undergoing hematopoietic stem cell transplant. Combining single cell transcriptional profiling with genotype analysis at the single cell level, we are able to compare the host and donor cell population changes before and after the transplant, and infer the relapse state of AML patients after the bone marrow transplant.

COMPOSITE MEASUREMENTS AND MOLECULAR COMPRESSED SENSING FOR HIGHLY EFFICIENT TRANSCRIPTOMICS.

Brian Cleary[1,2], Le Cong[2], Eric Lander[2,3], Aviv Regev[2,3]

[1]Massachusetts Institute of Technology, Computational and Systems Biology Program, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Massachusetts Institute of Technology, Biology, Cambridge, MA

Comprehensive RNA profiling provides an excellent phenotype of cellular responses and tissue states, but can be prohibitively expensive to generate at the massive scale required for studies of regulatory circuits, genetic states or perturbation screens. Here, we draw on a series of remarkable advances over the last decade in the field of mathematics to establish a rigorous link between biological structure, data compressibility, and efficient data acquisition. We begin by proposing that very few random composite measurements – in which each measurement reflects a randomly weighted linear combination of gene abundances – are needed to approximate the high-dimensional similarity between any pair of gene abundance profiles. Thus, in large-scale screens, for instance, cells that respond similarly to perturbation can be accurately clustered on the basis of a small number of random measurements. We then show how finding latent, sparse representations of gene expression data would enable us to “decompress” a small number of random composite measurements and recover unobserved, high-dimensional gene expression levels. We present a new algorithm for finding sparse, modular structure, which improves the ability to interpret samples in terms of small numbers of active modules, and show that the modular structure we find is sufficient to recover gene expression profiles from composite measurements (with ~100 times fewer composite measurements than genes). Moreover, we show that the knowledge that sparse, modular structures exist allows us to recover expression profiles from random composite measurements, even without access to any training data. Finally, we present a proof-of-concept experiment for making composite measurements in the laboratory, involving the measurement of linear combinations of RNA abundances. Overall, our results suggest new approaches for both experiments and interpretation in genomics and biology.

EVOLUTIONARY DYNAMICS OF REGULATORY CHANGES UNDERLYING GENE EXPRESSION DIVERGENCE AMONG *SACCHAROMYCES* SPECIES

Brian P Metzger[1], Patricia J Wittkopp[2], Joseph D Coolon[3]

[1]University of Chicago, Ecology and Evolution, Chicago, IL, [2]University of Michigan, Ecology and Evolutionary Biology, Ann Arbor, MI, [3]Wesleyan University, Biology, Middletown, CT

Heritable changes in gene expression are important contributors to phenotypic differences within and between species and are caused by mutations in *cis*-regulatory elements and *trans*-regulatory factors. While previous work has suggested that *cis*-regulatory differences preferentially accumulate with time, technical restrictions to closely related species and limited comparisons have made this observation difficult to test. To address this problem, we used allele-specific RNA-seq data from *Saccharomyces* species and hybrids to expand both the evolutionary timescale and number of species in which the evolution of regulatory divergence has been investigated. We find that as sequence divergence increases, *cis*-regulatory differences do indeed become the dominant type of regulatory difference between species, ultimately becoming a better predictor of expression divergence than *trans*-regulatory divergence. When both *cis*- and *trans*-regulatory differences accumulate for the same gene, they more often have effects in opposite directions than in the same direction, indicating widespread compensatory changes underlying the evolution of gene expression. The frequency of compensatory changes within and between species and the magnitude of effect for the underlying *cis*- and *trans*-regulatory differences suggests that compensatory changes accumulate primarily due to selection against divergence in gene expression as a result of weak stabilizing selection on gene expression levels. These results show that *cis*-regulatory differences and compensatory changes in regulation play increasingly important roles in the evolution of gene expression as time increases.

THE LANDSCAPE OF REGULATORY CHROMATIN CONTROLLING REGENERATION

Elena Vizcaya[1], Cecilia Klein[2], Florenci Serras[1], Roderic Guigo[2], Montserrat Corominas[1]

[1]Universitat de Barcelona, Genetics, Microbiology and Statistics, Institute of Biomedicine (IBUB), Barcelona, Spain, [2]Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra, Bioinformatics and Genomics, Barcelona, Spain

Regeneration is the ability to renew or reconstruct missing parts. A variety of mechanisms have been proposed to explain regeneration, ranging from stem cells to tissue remodeling. However, a central question is the identification of specific regulatory regions capable to trigger regeneration. Drosophila imaginal discs are epithelia that activate a regenerative response upon cell death. We used these epithelia to unveil the transcriptional program a well as the regulatory elements and reorganization of genome architecture responsible for tissue regeneration. Engineered flies allowed us to conditionally induce apoptosis, controlling the time of cell death and the zone to be removed. We report here genome-wide chromatin analyses (ATAC-Seq) and RNA-Seq at different time points after cell death induction. By integrating the maps of cis-regulatory elements and transcriptomic analyses we have constructed the gene regulatory network, with the aim to identify which regulatory regions are fundamental during the process. Comparison between control and discs immediately after cell death shows more regions of accessible chromatin, which correlates with higher number of up-regulated genes detected at the same time point. Both ATAC and RNA profiles tend to be more similar in late regeneration, when the recovery process is almost completed. Based on their presence or absence in control discs we have divided enhancers in developmental or damage-specific and, depending on their position relative to the TSS, classified them in proximal (5kb from the TSS) or distal (more than 5KB from the TSS). Distal enhancers have been linked to clusters of genes that show similar expression profiles and their position associated inside Topologically Associated Domains (TADs) meanwhile proximal enhancers tend to fall inside clusters of co-regulated genes. We have also validated several damage-induced enhancers using reporter fly lines after inducing apoptosis as well as after physical injury. Moreover, Chromatin Conformation Capture analysis (3C) has been used to confirm interacting chromatin loops between regulated genes with distal and proximal enhancers. Finally, the application of motif discovery tools searching for transcription factors that could bind to the identified regulatory regions has provided information on signaling pathways controlling regeneration. Our work provides a frame to understand gene expression regulation after damage and confirms that specific regeneration enhancers exist.

HIGH RESOLUTION EPIGENOMIC ATLAS OF EARLY HUMAN CRANIOFACIAL DEVELOPMENT

Andrea Wilderman[1,2], Jeffrey Kron[1], Justin Cotney[1,3]
[1]UConn Health, Genetics and Genome Sciences, Farmington, CT, [2]UConn Health, Grad. Program in Genetics and Dev. Bio., Farmington, CT, [3]UConn Health, Institute for Systems Genomics, Farmington, CT

Defects in embryonic patterning resulting in craniofacial abnormalities are common birth defects affecting up to 1 in 500 live births worldwide. Most of the individuals affected by congenital craniofacial abnormalities do not have defects in other tissues or organ systems, thus are considered non-syndromic. The regulatory programs that build and shape the craniofacial complex require precisely controlled spatiotemporal gene expression. Much of the information that controls these programs is thought to be encoded in the large expanses of the genome between genes and within intronic sequences. Efforts by large projects such as ENCODE and the Roadmap Epigenome Project have established associations of many biochemical features with functional portions of the genome and activity state. These efforts have revealed that enhancers are typically tissue or timepoint specific and regulate one or a small number of genes over very large genomic distances. The tissue specific nature of enhancers coupled with the enrichment of many disease phenotype associations from genome-wide association studies in such sequences suggest that defects in enhancers are at fault in many common disorders. The non-syndromic nature of many craniofacial defects suggests many may be due to defects in regulatory activity and thus likely to be “enhanceropathies”. To date early stages of human craniofacial development have not been interrogated with modern functional genomics techniques preventing systematic analysis of genetic associations and tissue-specific chromatin states. Using chromatin immunoprecipitation sequencing (ChIP-Seq), we have profiled the distribution of six post-translational modifications of histones across the genomes of craniofacial tissue from 17 human embryos spanning critical stages in human craniofacial development ranging from 4.5 post conception weeks (pcw) to 8 pcw. We have generated 91 primary ChIP-Seq datasets and have imputed these to encompass 12 epigenetic marks resulting in 252 high quality epigenomic datasets. Combining these data we have further generated 15, 18, and 25 state chromatin segmentation maps that can be directly compared to those generated by Roadmap Epigenome. Using these approaches we have identified 130435 craniofacial developmental enhancer regions. Comparisons with all chromatin states for each of the samples profiled by Roadmap Epigenome reveal over 20,000 putative enhancer regions that are only identified in our human craniofacial samples. These regions are significantly enriched near genes previously associated with craniofacial abnormalities and harbor significant numbers of common SNPs associated with orofacial clefting. We have targeted selected tissue specific enhancers for unbiased long range interaction mapping using circularized chromosome conformation capture sequencing (4C-Seq). These chromatin state and contact maps will allow identification of causative genetic changes in human patients with craniofacial abnormalities, development of better hypothesis for the etiology of craniofacial abnormalities, and generation of mouse models through precise editing of regulatory regions.

FUNCTIONAL DATA ANALYSIS INDICATES THAT THE ORAL MICROBIOME IS PREDICTIVE OF CHILD WEIGHT GAIN.

Sarah J Craig[1,2], Daniel Blankenberg[3], Alice Carla Louisa Parodi[4], Michele E Marini[5], Jennifer S Savage[5,6], Leann L Birch[7], Anton Nekrutenko[3], Ian M Paul[1,8], Matthew Reimherr[9], Francesca Chiaromonte[1,9], Kateryna D Makova[1,2]

[1]Penn State University, Center for Medical Genomics, University Park, PA, [2]Penn State University, Department of Biology, University Park, PA, [3]Penn State University, Department of Biochemistry and Molecular Biology, University Park, PA, [4]Politecnio di Milano, Department of Mathematics, Milan, Italy, [5]Penn State University, Center for Childhood Obesity Research, University Park, PA, [6]Penn State University, Department of Nutritional Sciences, University Park, PA, [7]University of Georgia, Department of Foods and Nutrition, Athens, GA, [8]Penn State College of Medicine, Department of Pediatrics, Hershey, PA, [9]Penn State University, Department of Statistics, University Park, PA

1 in 6 children in the US are overweight or obese. Obesity is a complex disease with many environmental influences; in particular, the microbiome is emerging as important. Characteristic perturbations in the gut microbiome have been shown in obese adults and adolescents when compared to their normal-weight peers. Less is known about the microbiome influence on weight gain in early childhood. This knowledge is critical because children with rapid weight gain have a greater risk for developing obesity later. Furthermore, the potential connection between oral microbiome and childhood obesity has not been studied.

We studied the relationship between infant weight gain in the first two years and gut and oral microbiome at 2 years. For this, we recruited 215 mother-child dyads. Child weight and length were measured at 6 time points during these 2 years and child mouth swabs and stool were collected at 2. We computed conditional weight gain (CWG) scores- a metric used to determine rapid infant weight gain, and also modeled growth curves using novel Functional Data Analysis (FDA) techniques. The oral and gut microbiome composition was surveyed with 16S sequencing.

Analysis of the oral microbiome revealed that it is predictive of weight gain. Children with rapid weight gain had a lower diversity and elevated Firmicutes-to-Bacteroidetes ratio in their oral microbiome community than children without rapid weight gain. Additionally, using multiple linear regression analyses, we discovered that gut microbiome is largely determined by diet (vegetables & fruit and vegetables & meats were significant predictors of the gut microbiome diversity and Firmicutes-to-Bacteroidetes ratio, respectively).

This study for the first time found significant associations between oral microbiome and child weight gain, and demonstrated the power of using growth curves as a predictor of microbiome composition.

PROBABILISTIC MODELING AND INFERENCE OF THE TRANSLATION DYNAMICS USING RIBOSOME PROFILING DATA

Khanh Dao Duc[1], Yun S Song[1,2]

[1]University of Pennsylvania, Biology, Philadelphia, PA, [2]University of California, EECS, Berkeley, CA

Translation elongation speed is quite heterogeneous along the transcript. Previous studies have shown that elongation is locally regulated by multiple factors, but the observed heterogeneity remains only partially explained. Ribosome profiling provides a detailed view into the complex dynamics of translation, but there are several challenges to using the data to characterize the translation dynamics. First, the precise relation between the observed ribosome footprint densities and the actual elongation speeds remains elusive. Second, the current experimental protocol excludes ribosomes that are positioned too close to each other (e.g., separated by < 3 codons) and does not capture the joint occupancy probability of multiple ribosomes on the same transcript. Hence, one cannot directly observe ribosomal interference and its role in limiting the elongation speed has never been quantified so far. Finally, estimating transcript- and position-specific elongation rates is a high dimensional statistical problem, and no suitable analytical tool is currently available for this task.

To address the above challenges and to dissect quantitatively the different determinants of translation speed, we have developed a realistic probabilistic model of the translation dynamics and efficient statistical inference tools. From ribosome profiling and RNA-seq data, we can estimate the initiation rate and position-specific elongation rates for each transcript. Using this approach, we infer the extent of interference between ribosomes on the same transcript and show that it varies substantially across different genes and different positions. However, we show that neither ribosomal interference nor the distribution of slow codons is sufficient to explain the observed variation in the mean elongation rate along the transcript. Surprisingly, by optimizing the fit of statistical linear models, we find that the hydropathy of the nascent polypeptide segment within the ribosome plays a major role in governing the variation of the mean elongation rate. In addition, we find that positively and negatively charged amino acid residues near the beginning and end of the ribosomal exit tunnel, respectively, are important determinants of translation speed. This result is consistent with the electrostatic and geometric properties of the exit tunnel, which we study quantitatively using cryo-EM data.

THE EFFECTS OF DEMOGRAPHIC HISTORY ON THE DETECTION OF RECOMBINATION HOTSPOTS

Amy L Dapper, Bret A Payseur

University of Wisconsin - Madison, Laboratory of Genetics, Madison, WI

In many species, meiotic recombination is concentrated in small genomic regions. These “recombination hotspots” leave signatures in fine-scale patterns of linkage disequilibrium, raising the prospect that the genomic landscape of hotspots can be characterized from sequence variation. This approach has led to the inference that recombination hotspots evolve rapidly in some species, but are conserved in others. Historic demographic events, such as population bottlenecks, are known to affect patterns of linkage disequilibrium across the genome, violating population genetic assumptions of this approach. Although such events are prevalent, demographic history is generally unaccounted for when making inferences about the evolution of recombination hotspots. To determine the effect of demography on the detection of recombination hotspots, we use the coalescent to simulate haplotypes with a known recombination landscape. We measure the ability of popular linkage disequilibrium-based programs to detect recombination hotspots across a range of demographic histories, including population bottlenecks, hidden population structure, population expansions and population contractions. We find that demographic events, and in particular, population bottlenecks and hidden population subdivision, have the potential to greatly reduce the power to discover recombination hotspots. Furthermore, demographic events can increase the false positive rate of hotspot discovery. In the worst-case scenario, demography and background recombination rate interact to more than triple the frequency of false positives, even under stringent significance cutoffs. We found that neither the power, nor the false positive rate, of hotspot detection could be predicted without also knowing the demographic history of the sample. Our results suggest that ignoring demographic history likely overestimates the power to detect recombination hotspots and therefore underestimates the degree to which recombination hotspots are shared between closely related species. As a result, our inferences about the rate of evolution of recombination hotspots may be inaccurate, especially among populations that have experienced significant deviations from an equilibrium demographic history. We argue that studies using linkage disequilibrium-based approaches to measure fine-scale variation in recombination rate should incorporate demographic history.

IDENTIFICATION AND ANALYSIS OF SOMATIC VARIANTS USING LINKED READ SEQUENCING

Charlotte A Darby[1], Ben Langmead[1,2,3], Michael Schatz[1,4,5]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Johns Hopkins Bloomberg School of Public Health, Biostatistics, Baltimore, MD, [3]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [4]Johns Hopkins University, Biology, Baltimore, MD, [5]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

In contrast to germline (inherited) variants, mutations occurring early in development are only present in some cells of the developed individual. A healthy human is thought to harbor tens to a few hundred benign "somatic mutations" throughout their body, but a single additional one can be disease-causing. Somatic mutations have been implicated in autism, rare diseases, including those where the skin has a visible "mosaic" pattern, and many forms of cancer.

These mutations do not follow typical patterns of inheritance, so exome sequencing or high-coverage whole-genome sequencing of a familial trio, paired samples, or a single sample are currently used to identify somatic variants based on statistical analysis of allele frequency. We suggest new strategies using “linked reads” to distinguish true somatic mosaicism from its confounders, such as copy number variation, sequencing error, clusters of variants, and differences between the individual sequenced and the reference genome.

A linked read is a group of short reads sequenced from the same original long DNA molecule. 10X Genomics commercializes a linked-read technology adding a library preparation protocol to standard short-read sequencing. Due to this additional locality information, linked reads enable high-quality phasing, the assignment of heterozygous variants to maternal and paternal haplotypes - a task impossible with short reads alone.

In addition to the two parental haplotypes, cells bearing a somatic mutation effectively form a third haplotype, as they are mixed in with normal cells in the biological sample used for sequencing. At these sites, the heterozygous allele fraction diverges from 50-50, or the homozygous from 0-100. Starting with simulated data, we develop a method to identify high-confidence candidate mosaic variants. Linked read-based phasing is used to assign specific reads to haplotypes, thus supplementing the allele fraction at a site with haplotype-specific allele counts. We then apply our method to linked-read sequencing data of healthy and diseased samples.

IDENTIFICATION OF DRUG EQTL INTERACTIONS FROM REPEAT TRANSCRIPTIONAL AND ENVIRONMENTAL MEASUREMENTS IN A LUPUS CLINICAL TRIAL

Emma E Davenport[1,2,3], Tiffany Amariuta[1,2,3], Maria Gutierrez-Arcelus[1,2,3], Kamil Slowikowski[1,2,3], Harm-Jan Westra[1,2,3], Ying Zhang[4], Stephen Pearson[5], David von Schack[4], Jean S Beebe[4], Nan Bing[4], Michael S Vincent[4], Baohong Zhang[4], Soumya Raychaudhuri[1,2,3,6]

[1]Brigham and Women's Hospital, Harvard Medical School, Divisions of Genetics and Rheumatology, Department of Medicine, Boston, MA, [2]Partners Center for Personalized Genetic Medicine, Boston, MA, [3]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [4]Pfizer Inc., Cambridge, MA, [5]Pfizer New Haven Clinical Research Unit, New Haven, CT, [6]University of Manchester, Faculty of Medical and Human Sciences, Manchester, United Kingdom

Environmental eQTL interactions modify the relationship between genetic variation and abundance of gene expression. Since these are genetic effects, if eQTL interactions can be found, their common driving mechanisms may represent those that are a consequence of an environmental perturbagen and therefore be informative for understanding regulatory mechanisms. This approach may be particularly powerful for pharmacogenetics where defining drug mechanism of action is critical.

In an anti-IL-6 randomized clinical trial data set of 157 patients with systemic lupus erythematosus we measured three environmental factors (cell count with flow cytometry, interferon signature, and drug exposure) at three time points alongside RNA-seq assays. Using a linear mixed model, including repeat measurements (379 RNA-seq assays from 157 patients) we identified 4,939 *cis* eQTL genes with $p<2.3x10^{-8}$ (0.05/2,184,435 tests). Repeat measurements increased our power by detecting 63% more *cis* eQTLs compared to using a linear model with a single sample. We identified 154, 182 and 128 eQTL interactions (p<0.01) with T cell count, IFN status and anti-IL-6 drug exposure respectively (more than expected by chance from 1,000 permutations). Examples of interaction genes include *NOD2* for T cell count, *SLFN5* for IFN status and *IL-10* for drug exposure.

An interaction can either magnify or dampen an original eQTL effect. We investigated the difference between magnifiers and dampeners by looking for transcription factor binding motifs interrupted by interaction SNPs. For IFN status we find an enrichment of biologically relevant transcription factors including IRF1 and IRF2.

We find 78/128 drug-eQTL interactions are consistent with free IL-6 protein interactions. Finally, we use the interactions to define a simple drug exposure score that is correlated with drug dose and can serve as a metric to define the effective drug exposure per individual. This same approach can be easily applied to larger drug trials to further our understanding of drug mechanisms.

SYSTEMATIC GENOME-WIDE ANALYSIS OF LOCALIZED HUMAN CHROMATIN STATE PLASTICITY

Jose Davila-Velderrain[1,2], Lei Hou[1,2], Manolis Kellis[1,2]

[1] Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Lab, Cambridge, MA, [2] Broad Institute of MIT and Harvard, Computational biology, Cambridge, MA

Epigenomics has a central role for understanding cellular phenotypic state transitions, for it is by definition centered at the core of a fundamental problem in development: how is the diversity of cellular and tissue types established, starting from a single, largely invariant genome? The study of the dynamic patterns of epigenomic marks provide a principled way to tackle such problem, enabling the discovery of biologically meaningful relationships between cell types, tissues, and lineages. To facilitate such large-scale and technically challenging studies, chromatin states have been introduced for genome segmentation, providing unbiased, inferred states that effectively synthesize in a single state variable the information that is otherwise collectively distributed in a large number of epigenomic tracks. The Roadmap Epigenomics project has produced whole-genome chromatin state profiles mapped to a large number of cell/tissue types (~127), providing an unprecedented opportunity to distill biological insights from a single discrete-state variable track. Interestingly, we observed a high heterogeneity across the patterns of stability of chromatin marks: some genome regions are highly plastic whereas others maintain a single chromatin state across all tissues/cell-types. In order to begin understanding the biological relevance of such heterogeneity in chromatin state plasticity, we systematically analyzed the observed dynamic patterns by computing a simple chromatin plasticity index for each 200bp bin across the human genome. Using such index, we characterize the global patterns of chromatin plasticity across the genome, discovering high heterogeneity across chromosomes, which allowed us to rank them accordingly. Importantly, we identify specific plastic hotspots that correlate with biological processes involved in the response to signaling mechanisms. We are currently dissecting such plastic hotspots, by identifying enhancer elements and transcriptional regulators preferentially associated with them; and by testing for association with disease loci. Finally, we are analysing to what extent the observed patterns of chromatin plasticity are conditional on the selection of specific subsets of cell/tissue types as reference, with a focus on brain chromatin heterogeneity. Overall, the output of our analysis will provide a reference systematic view and biological interpretation of the patterns of chromatin plasticity manifested by the dynamical changes of integrative chromatin active and repressive states across the human genome.

DECODING THE ARCHITECTURE OF HUMAN CIS-REGULATORY ELEMENTS

Jessica Davis[1], Kim Insigne[2], Eric Jones[3], Quinn Hastings[1], Sri Kosuri[1]

[1]UCLA, Chemistry & Biochemistry, Los Angeles, CA, [2]UCLA, Bioinformatics IDP, Los Angeles, CA, [3]UCLA, Molecular Biology Institute, Los Angeles, CA

It is becoming increasingly clear that variation in cis-regulatory elements (CREs) plays a large role in human disease. Yet, predicting the effects on gene expression from mutations in these regions is not always straightforward. Part of this issue is an incomplete picture of how the activities of regulatory sequences, such as transcription factor binding sites (TFBSs), combine to direct the transcription of a single gene. The differential organization of binding sites in CREs is vast and varies according to: placement, number of sites, surrounding sequence composition, combination of other binding sites and binding strength. We employ a massively parallel reporter assay to isolate the effects of TFBS variables upon driving transcription in designed synthetic regulatory elements. Using binding sites for a model transcription factor, CREB protein, we systematically test relative site placement, number, and strength of transcription factor binding sites in isolation and in combination. After integration into a single-copy landing pad in HEK293T cells, we assay our pool of 10,686 variants for their ability to drive genomic transcription of a cellular reporter. We test regulatory libraries *en masse* and use next-gen sequencing to determine transcript abundance per regulatory variant. Initial studies indicate varied transcriptional responses from designed regulatory elements to active, cellular CREB protein levels. Larger-scale studies will elucidate the contribution of these binding site variables to genomic expression. By isolating the effects of TFBS architecture in designed regulatory elements, we hope to understand how the organization of TFBSs in more complex, natural CREs culminate to drive observed transcription levels.

LOCAL REGULATORY NETWORKS ACROSS TWO TISSUES AND APPLICATIONS TO ANALYZE RARE NON-CODING VARIANTS

Olivier Delaneau[1], Konstantin Popadin[2], Marianna Zazhytska[2], Christelle Borel[1], Giovanna Ambrosini [3], Daniel Marbach[4], Sven Bergmann[4], Philipp Bucher[3], Stylianos Antonarakis[1], Alexandre Reymond[2], Emmanouil Dermitzakis [1]

[1]University of Geneva, Dpt of Genetic Medicine and Development, Geneva, Switzerland, [2]University of Lausanne, CIG, Lausanne, Switzerland, [3]EPFL, Swiss Institute for Experimental Cancer Research, Lausanne, Switzerland, [4]University of Lausanne, Dpt of Computational Biology, Lausanne, Switzerland

Population measurements of gene expression and genetic variation enable the discovery of thousands of expression Quantitative Trait Loci (eQTLs), a great resource to determine the function of non-coding variants. To describe the effects of eQTLs on regulatory elements such as enhancers and promoters, we quantified gene expression (mRNA) and three key histone modifications (H3K4me1, H3K4me3 and H3K27ac) across two cell types (Fibroblast and Lymphoblastoid Cell Lines) in 80 and 320 densely genotyped European samples, respectively.

First, we find that nearby regulatory elements form local chromatin modules spanning up to 1Mb, often comprising multiple sub-compartments and overlapping topologically associating domains. These modules bring multiple distal regulatory elements in close proximity, vary substantially across cell types and drive co-expression at multiple genes.

Next, we show that this regulation layer is under strong genetic control: we discovered ~34k chromatin QTLs (cQTLs) affecting ~30% of the chromatin peaks. In addition, we quantified chromatin modules by using principal component analysis and find QTLs for up to ~70% of them (modQTLs). These collections of cQTLs and modQTLs represent a new resource of functional variants with downstream effects on higher-order phenotypes: they often are eQTLs, cell type specific and enriched for disease associated variants (GWAS).

Then, we show how chromatin modules can increase the power of association studies of rare variants when whole genome sequencing is available. Specifically, we derived a burden test on gene expression, applied it on the Geuvadis transcriptomic data and discovered that the expression of many genes (~20%) is associated with rare non-coding variants located in modules.

Finally, we find that the coordination between regulatory elements located on distinct chromosomes (i.e. in trans) is well supported by Hi-C sequencing data and seem to drive in some cases trans eQTL effects. Of note, we replicated up to 80% of the strongest inter-chromosomal Hi-C contacts.

Overall, this large-scale study integrating gene expression, chromatin activity and genetic variation across two cell types and hundreds of samples provides key insights into the biology underlying gene regulation and eQTLs.

GENE.IOBIO - A WEB-BASED VISUALIZATION TOOL FOR REAL-TIME VARIANT ANALYSIS

Tonya Di Sera, Chase A Miller, Yi Qiao, Alistair Ward, Matt Velinder, Gabor Marth

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

Currently, identifying causative variants in genetic diseases relies on trained bioinformaticians to generate static text-based reports that are subsequently filtered on sequencing depth and quality; and prioritized based on presence in clinical variant databases, predicted variant impact, population allele frequency and a variety of other metrics. This is a computational and time-intensive workflow, not suited for the vast majority of the users who need it most -- human geneticists or research clinicians without bioinformatics training or expertise. To ameliorate these burdens we have developed a web-based tool, *gene.iobio* (http://gene.iobio.io), that allows for rapid and intuitive genetic variant visualization. *Gene.iobio* allows the analyst to select individual proband or family trio sequence alignment (BAM) and/or variant call (VCF) data for powerful real-time variant analysis within seconds, without any subsequent file uploads. The analyst may begin disease variant investigation either in a single gene or in a candidate gene list; or generate a gene list based on the patient's phenotype from within our tool. Each variant within these genes is annotated, in real time, with a variety of metrics including mode of inheritance, sequencing depth, and up-to-date variant annotations from a variety of databases such as 1000G and ExAC population frequencies and predicted functional impact by VEP, SIFT, Polyphen and Clinvar. Variants either within a gene or across all genes are prioritized according to easily configurable criteria. *Gene.iobio* can currently annotate and prioritize 200 genes from whole-exome trio data in under 60 seconds, made possible by cloud-based server parallelization. In our experience, *gene.iobio* often reveals the patient's disease-causing variant(s) within minutes of commencing analysis. In cases where no clear disease-causing variant is found, our tool allows re-calling variants in real-time with more lenient parameters, as well as intuitive examination of sequencing coverage for variants potentially missed in the primary analysis. The ease-of-use, powerful features, and real-time, highly visual analysis offered by *gene.iobio* opens sophisticated clinical variant investigation to a large community of medical and clinical researchers; and we are driving toward the application of this tool at the point of care in genetic clinics.

MAPPING TO PERSONAL GENOMES WITH STAR-DIPLOID.

Alexander Dobin[1], Carrie Davis[1], Fritz J Sedlazeck[2], Han Fang[1,3], Yunjiang Qiu[4,5], David Gorkin[4,5], Sora Chee[4,5], Anna Vlasova[6], Alessandra Breschi[6], Roderic Guigo[6], Michael C Schatz[1,2], Bing Ren[5,6], Thomas R Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]John Hopkins University, Baltimore, MD, [3]Stony Brook University, Stony Brook, NY, [4]Ludwig Institute for Cancer Research, La Jolla, CA, [5]University of California, San Diego, CA, [6]Center for Genomic Regulation, Barcelona, Spain

Personal genomics is envisaged to become an essential component of precision medicine, holding the promise for identifying genetic predispositions for common and complex diseases, diagnosis of hereditary disorders, individual treatment of cancer, and genotype guided drug research. Millions of personal genomes will be sequenced in the next few years, however, the tools are lacking for personalized processing of the functional data types such as RNA-seq and ChIP-seq, which at present are routinely mapped to the haploid reference genome. Here we present an extension of popular RNA-seq aligner STAR, the STAR-Diploid software that was developed to map RNA-seq and ChIP-seq reads to the fully phased diploid personal genomes.

In the first step, STAR-Diploid utilizes the personal variants, including single nucleotide variants, short indels, as well as large structural variants, to build the personal diploid genome sequence from the reference assembly. The reference annotations are arithmetically lifted over to each of the haplotypes. Next, the reads are mapped to both haplotypes simultaneously to produce diploid genomic alignments. Mapping to the personal diploid sequence virtually eliminates the reference bias which plagues the alignment to the haploid reference genome. The diploid alignments are then converted to the reference coordinates while preserving the haplotype information. The final output of the pipeline consists of haplotype-specific alignments and signal (wiggle) tracks in the reference coordinates which can be visualized in the standard genomic browsers, as well as allele-specific counting of reads per gene. Furthermore, STAR-Diploid converts diploid genomic alignment into diploid transcriptomic alignments that are input into RSEM for allele-specific quantification of transcripts and gene expression.

To demonstrate the effectiveness of the STAR-Diploid algorithm, we have utilized it to process a large collection of long and small RNA-seq, RAMPAGE and ChIP-seq data from the ENCODE-GTEx (EN-TEx) collaboration (~20 tissues for 4 donors). The personal diploid genomes were constructed with the variants phased by means of 10x Genomics Chromium sequencing and Hi-C short reads for chromosome-span phasing, with RNA-seq data used to supplement and resolve phasing conflicts. We have compared allelic imbalance across the multiple tissues obtained from the same donor, as well as across multiple individuals for the same tissue, for the purpose of understanding the genotypic and cell type contributions to ASE. Using long-range variant phasing information from the 10x Genomics and Hi-C data, we have identified potential causative mutations in regulatory regions responsible for the observed ASE.

TRANSCRIPTIONAL CONTROL AND CHROMATIN LANDSCAPE OF DOWNREGULATED GENES IN INNATE IMMUNE CELLS

Elisa Donnard[1], Pranitha Vangala[1], Barbara Tabak[1], Patrick McDonel[1], Anetta Nowosielska[2], Sean McCauley[2], Jeremy Luban[2], Manuel Garber[1]

[1]University of Massachusetts Medical School, Bioinformatics and Integrative Biology, Worcester, MA, [2]University of Massachusetts Medical School, Program in Molecular Medicine, Worcester, MA

Innate immune cells have dynamic responses to environmental stimulus and a coordinated regulation of transcriptional response in which thousands of genes show significant expression changes. We examined the temporal expression of genes in human dendritic cells (DCs) derived from monocytes following lipopolysaccharide (LPS) stimulation and defined a set of genes that are downregulated in this response. By using RNA-Seq reads for both intronic and exonic regions, we were able to dissect downregulation control into two main mechanisms, transcriptional and post-transcriptional. From the 1,612 genes that are downregulated, we see evidence that the majority is under transcriptional control.

We identified the cis-regulatory elements active in these cells through ChIP-Seq of the chromatin mark H3K27ac, and also determined the accessibility of these regulatory regions through ATAC-Seq. Close to 4,000 putative enhancer elements are associated to the downregulated genes, and 67% of them are already established in the progenitor monocytes from which the human DCs are derived. A comparison of these elements to regulatory regions identified in mouse dendritic cells also treated with LPS shows that the majority of them (78%) are active only in the human DCs. By clustering the post-transcriptionally downregulated genes, similar repression patterns were identified, and a de novo motif analysis for each cluster reveals enriched RNA motifs in the 5'UTR and 3'UTR of these transcripts. These de novo motifs were compared to previously annotated regulatory motifs to identify possible molecules involved in the repression control. Interestingly, comparing these results with data from mouse DCs stimulated with LPS, we detect downregulated genes with seemingly the same transcriptional or post-transcriptional control, and also genes for which the downregulation has switched from transcriptional to post-transcriptional, providing an evolutionary perspective.

MULTIPLE MOUSE REFERENCE GENOMES DEFINES SUBSPECIES SPECIFIC HAPLOTYPES AND NOVEL CODING SEQUENCES

Anthony G Doran, Jingtao Lilue, Thomas M Keane, The Mouse Genomes Project consortium

Wellcome Trust Sanger Institute, Structural variation infrastructure, Cambridge, United Kingdom

The Mouse Genomes Project consortium (Presenter: Anthony Doran, Wellcome Trust Sanger Institute)

The Mouse Genomes Project has completed the first draft assembled genome sequences and strain specific gene annotation for twelve classical laboratory and four wild-derived inbred mouse strains (WSB/EiJ, CAST/EiJ, PWK/PhJ and SPRET/EiJ). These strains represent a genetically (>1M years) and phenotypically diverse panel of inbred mouse strains, and include founders of highly used recombinant lines such as the Diversity Outbred and Collaborative Cross. Our genome sequences have increased base pair accuracy and comparable structural accuracy compared to the first release of the mouse reference genome (MGSCv3). We used a hybrid approach for genome annotation, combining evidence from the mouse reference Gencode annotation and strain-specific transcript evidence (RNA-seq and PacBio cDNA), to identify novel strain-specific gene structures and alleles. The largest number of novel gene structures were identified in the wild derived strains. As these strains are fully inbred, we used heterozygous SNP density as a marker for highly polymorphic loci, and found these loci to be enriched for genes related to immunity, olfaction and sensory function. We focus in particular on four immune related loci (IRG, Nlrp1, Schlafen and Raet1) containing novel sequence, coding alleles and differing gene structures in the wild derived strains. In mouse, anthrax lethal toxin (LT) is currently the only known activator of Nlrp1 and genetic differences in Nlrp1b are linked sensitivity to anthrax LT. For the first time, our data shows the striking allelic diversity in this locus, identifying new coding alleles shared by subsets of the susceptible and resistant strains. In another locus on Chr10 under a QTL associated with susceptibility to Aspergillus infection, we identified new allelic combinations of the H60 and Raet1 homologs that segregate with the phenotype. Of particular note was the discovery of a previously unannotated rodent specific 138 exon gene on chromosome 11. Manual annotation extended this novel gene as a combination of the human genes EFCAB3 and EFCAB13 on human chromosome 17. The genome sequences and annotation can be viewed in the UCSC and Ensembl genome browsers.

THE USE OF PREDICTED GENE EXPRESSION PROFILES TO UNCOVER PATHWAYS INVOLVED IN DRUG-INDUCED PANCREATITIS

Britt I Drogemoller[1,2], Shinya Ito[3], Bruce C Carleton[2,4], Colin J Ross[1,2], The Canadian Pharmacogenomics Network for Drug Safety Consortium[1]

[1]University of British Columbia, Faculty of Pharmaceutical Sciences, Vancouver, Canada, [2]University of British Columbia, BC Children's Hospital Research Institute, Vancouver, Canada, [3]University of Toronto, The Hospital for Sick Children, Clinical Pharmacology and Toxicology, Toronto, Canada, [4]University of British Columbia, Department of Pediatrics, Vancouver, Canada

Background

L-asparaginase is highly effective in the treatment of paediatric acute lymphoblastic leukaemia. Unfortunately, this treatment is accompanied with debilitating adverse drug reactions such as pancreatitis. As treatment dose and formulation have not been able to explain the inter-individual differences that are observed, genetic factors may play an important role in the development of pancreatitis.

Methods

A total of 123 patients who have been treated with L-asparaginase were recruited from 13 oncology units across Canada. Extensive demographic and clinical data have been collected for all patients and genotyping of 740,000 genetic variants has been performed using the Illumina HumanOmniExpress array. These data were used to impute predicted pancreatic gene expression profiles on patients using GTEx and PrediXcan. Genes that were predicted to be differentially expressed in cases and controls (P<0.01) were subsequently investigated for enrichment in specific pathways using the WEB-based GEne SeT AnaLysis Toolkit.

Results

Analyses with PrediXcan identified 29 genes that were predicted to be differentially expressed in patients experiencing L-asparaginase-induced pancreatitis when compared to patients who did not experience this adverse drug reaction (*P*<0.01). Drug association enrichment analyses using the WEB-based GEne SeT AnaLysis Toolkit revealed an enrichment in adenosine triphosphate (ATP) related genes (adjP=0.0022).

Conclusions

This study has identified a role for genes involved in the ATP pathway and the development of L-asparaginase-induced pancreatitis. These results align well with reports that pancreatitis can be prevented by supplementation with intracellular ATP and provide further insight into the development of strategies to prevent the occurrence of this adverse drug reaction.

NASCENT RNA SEQUENCING REVEALS A DYNAMIC GLOBAL TRANSCRIPTIONAL RESPONSE AT GENES AND ENHANCERS TO THE NATURAL MEDICINAL COMPOUND CELASTROL

Noah E Dukler[1,2], Gregory T Booth[3], Yi-Fei Huang[1], Nathaniel D Tippens[2,3], Charles G Danko[4], John T Lis[3], Adam Siepel[1]

[1]CSHL, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]WCMC, Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY, [3]Cornell U., Department of Molecular Biology and Genetics, Ithaca, NY, [4]Cornell U., Baker Institute, Ithaca, NY

Most studies of responses to transcriptional stimuli measure changes in cellular mRNA concentrations. By sequencing nascent RNA, it is possible to detect changes in transcription in minutes rather than hours, allowing the reconstruction of additional mechanistic information about how transcriptional changes arise. Here, we describe the use of PRO-seq to characterize the immediate transcriptional response in human cells to celastrol, a compound derived from traditional Chinese medicine that has potent anti-inflammatory, tumor-inhibitory and obesity-controlling effects. Our analysis of PRO-seq data for K562 cells reveals dramatic transcriptional effects soon after celastrol treatment at a broad collection of both coding and noncoding transcription units. This transcriptional response occurred in two major waves, one within 10 minutes, and a second 40-60 minutes after treatment. Transcriptional activity was generally repressed by celastrol, but one distinct group of genes, enriched for roles in the heat shock response, displayed strong activation. Using a regression approach, we identified key transcription factors that appear to drive these transcriptional responses, including members of the E2F and RFX families. Moreover, this approach reveals some regulatory TFs that clearly separate the celastrol and heat shock responses; for example, the loss of binding by MYC-MAX may be responsible, in part, for the broad transcriptional repression within 20 minutes of celastrol treatment. We also found sequence-based evidence that particular TFs drive the activation of enhancers. Finally, we observed increased polymerase pausing at both genes and enhancers, suggesting that pause release may be widely inhibited during the celastrol response. Our study demonstrates that a careful analysis of PRO-seq time course data can disentangle key aspects of a complex transcriptional response, and it provides new insights into the activity of a powerful pharmacological agent.

RAPID IN VITRO EVALUATION OF VARIANTS OF UNCERTAIN SIGNIFICANCE FROM PATIENTS WITH DEVELOPMENTAL DELAY AND/OR INTELLECTUAL DISABILITY

Krysta L Engel[1], Jesse N Cochran[1], Andrew A Hardigan[1], Kevin M Bowling[1], Michelle D Amaral[1], Candice R Finila[1], Susan M Hiatt[1], Michelle L Thompson[1], David E Gray[1], Neil E Lamb[1], Edward J Lose[2], Martina Bebin[2], Gregory S Barsh[1], Gregory M Cooper[1], Richard M Myers[1]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, [2]University of Alabama at Birmingham, Birmingham, AL

Developmental delay and intellectual disabilities (DD/ID) include devastating phenotypes and comprise a large fraction of rare undiagnosed conditions in children. Unfortunately, little is known about the cellular mechanisms that lead to disease, therefore therapeutic advancements have suffered. Previous studies have indicated a strong genetic component to DD/ID, and successful identification of causal genetic variants through genome sequencing sometimes leads to clinical diagnoses. However, sequencing-based diagnostic efforts typically solve only a subset of cases (~27% by our group) and a large fraction of children cannot be given a precise genetic diagnosis. Rigorous experimental evaluation of variants and genes implicated in DD/ID is needed to increase diagnostic rate, particularly for the subset of the undiagnosed cases where a potential genetic cause is identified but not confirmed; variants of uncertain significance (VUSs). VUSs typically arise as a result of a lack of information about the relevance of a given gene to disease, the impact of a given variant on gene function, or both. In our studies to date, ~15% of affected children harbor a VUS. Introduction of these VUSs, a portion of which are in gene expression regulators, into human neurons derived from neural precursor cells through gene editing technology will provide evidence for or against the association of these sequence variants on key molecular and cellular phenotypes including global gene expression at the levels of both transcription and translation, and neuronal excitability. The studies proposed here will provide critical insights regarding the biological roles of genes and variants associated with DD/ID and establish a framework for future mechanistic interrogation of genetic variation. Furthermore, the insights gained from this work will inform future studies aimed at therapeutic intervention for DD/ID disorders.

POPULATION-SCALE COLLECTION OF GENOMES, GENEALOGY, AND PHENOTYPES USING SINGLE LAB RESOURCES

Yaniv Erlich[1,2], Assaf Gordon[1], Jie Yuan[1,2], Daniel Speyer[1,2], Richard Aufrichtig[1]

[1]New York Genome Center, n/a, New York, NY, [2]Columbia University, Computer Science, New York, NY

Elucidating the genetic basis of complex traits requires substantial amount of data. In the last few years, the massive cost reduction in genomic technologies have enabled large scale studies that involve tens- to hundred-of thousands of participants. However, even with these low costs, constructing large scale collections of genetic data is a daunting task. For example, the NIH's Precision Medicine Initiative ("All of US") has allocated about $150 million just for the operations of recruiting approximately 100,000 patients using a series of procedures that mainly involve shipping biological samples without actual genotyping. As an alternative, we developed an array of tools that essentially allow to collect massive amount of data using single lab-resources. Our approach relies on the observation that prospective participants have already accumulated massive amount of digital data about themselves in multiple website and digital outlets. Therefore, we focus on digital aggregation of these datasets using crowd sourcing strategies that encourage prospective participants to donate their data. In our recent studies, we constructed population-scale family trees with 86 million individuals by collecting public data from Geni.com, the largest genealogy-driven social media website. Then, we collected over 40,000 genomes from DTC participants using a website called DNA.Land. Finally, we are now working on an approach to connect the genomes and genealogy data with phenotypic data that will be collected using Facebook profiles of individuals. Essentially, our approach creates a full stack of population-scale genomes, genealogy, and phenotypes simply using digital media. The success of these efforts offers a viable alternative to traditional cost-prohibitive efforts to collect this data.

FIDDLE: AN INTEGRATIVE DEEP LEARNING FRAMEWORK FOR FUNCTIONAL GENOMIC DATA INFERENCE

Umut Eser, Stirling Churchman

Harvard Medical School, Genetics, Boston, MA

Numerous advances in sequencing technologies have revolutionized genomics through generating many methods, such as ChIP-seq, RNA-seq, and NET-seq, that report on different aspects of the genome, epigenome and gene expression. Statistical tools have been developed to analyze individual data types, but there lack strategies to integrate disparate datasets under a unified framework. Moreover, most analysis techniques heavily rely on feature selection and data preprocessing which increase the difficulty of addressing biological questions through the integration of multiple datasets. Here, we introduce FIDDLE (Flexible Integration of Data with Deep LEarning) an open source data-agnostic flexible integrative framework that learns a unified representation of multiple data types to infer another data type. As a case study, we use multiple *Saccharomyces cerevisiae* functional genomic datasets to predict global transcription start sites (TSS) through the simulation of TSS-seq data. We demonstrate that a type of data (e.g. TSS-seq data) can be inferred from other sources of data types (e.g. NET-seq, ChIP-seq *etc.*) without manually specifying the relevant features and preprocessing. We find the average relative entropy of our model's predictions is almost equal to the relative entropy obtained by comparing two biological replicate datasets. Furthermore, we show that models built from multiple genome-wide datasets perform profoundly better than models built from individual datasets. Thus FIDDLE learns the complex epistatic relationship within individual datasets and, importantly, across datasets. Moreover, we demonstrate how FIDDLE can be used in cross-species/cell types predictions.

EXTENDING AND UPDATING THE 1000 GENOMES PROJECT DATA IN THE INTERNATIONAL GENOME SAMPLE RESOURCE (IGSR)

Susan Fairley, Peter Harrison, Ernesto Lowy, David Richardson, Ian Streeter, Galabina Yordanova, Holly Zheng-Bradley, Laura Clarke, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

In generating the largest public catalogue of human genetic variation, the 1000 Genomes Project created valuable data sets, which continue to be widely used. Following completion of the 1000 Genomes Project, the International Genome Sample Resource (IGSR) was established to ensure the continued usability of the 1000 Genomes data, to share new data generated on 1000 Genomes samples and to make new populations available.

IGSR works to meet these aims by providing ethical review for new populations, supporting work with existing samples via data coordination, improving the discoverability of data sets through the project website and by re-analysing the 1000 Genomes Project data on GRCh38.

Low-coverage and exome data from the 1000 Genomes Project were re-aligned to GRCh38 using alt-aware BWA. These publicly available alignments are being used as the basis for a 1000 Genomes call set on GRCh38. The variant callers being used in this work are SAMtools, FreeBayes and GATK, with preliminary calls on chromosome 20 available on the project FTP site.

The 1000 Genomes Project made over half a million files publicly available. To aid navigation of this resource, we developed a data portal (http://www.internationalgenome.org/data-portal/), which we continue to develop and expand. To understand the needs of users, we conducted a survey, information from which is being used to develop strategies to improve the presentation of data in IGSR and also in prioritising expansion of the data sets. The strongest interest for new data sets was in new populations and functional data.

New populations and RNA-seq data are being added to IGSR. Three new populations, each containing 100 samples, have been added from the Gambian Genome Variation Project and RNA-seq data from the GEUVADIS project has also been incorporated. In addition, over the last year, data from the Simons Diversity Project and HGDP-CEPH have joined the collections.

IGSR continues to provide data coordination for the Human Genome Structural Variation Consortium, adding data from a diverse range of technologies for trios from three populations.

IGSR will continue to seek to expand the data types, samples and populations in its collections, aiming to share high-quality, open data with the community. With this in mind, IGSR welcomes discussions relating to how IGSR can support projects sharing open data and potential collaboration.

SCIKIT-RIBO REVEALS PRECISE CODON-LEVEL TRANSLATIONAL CONTROL BY DISSECTING RIBOSOME PAUSING AND CODON ELONGATION

Han Fang[1], Yifei Huang[1], Aditya Radhakrishnan[2], Max Doerfel[1], Adam Siepel[1], Rachel Green[2], Gholson Lyon[1], Michael Schatz[1,2]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Johns Hopkins University, Baltimore, MD

Ribosome profiling (Riboseq) is a powerful technique for monitoring protein translation in vivo, analogous to RNAseq for expression profiling. However, there are very few methods available to analyze Riboseq data. Here, we present scikit-ribo, a framework for joint analysis of Riboseq and RNAseq data. We provide modules for ribosome A-site prediction, ribosome pausing site calling, joint inference of protein translation efficiency (TE) and codon elongation rate.

We show our method has high accuracy to identify A-site location for data with different mRNA digestion (0.95 for RNase I in yeast and 0.91 for RelE in bacteria). After improving the ribosome A-site resolution to 3bp, we built a negative binomial mixture model to identify and analyze ribosome pausing sites. From this we discovered the commonly used RPKM-based TE calculation is very sensitive to ribosome pausing events, thus negatively skewing the TE distributions in almost all previous studies and limiting their ability to differentiate translation efficiency and codon optimality. To solve this, we built a generalized linear model to simultaneously infer protein TE and codon elongation rates, while accounting for mRNA abundance and secondary structure. We also successfully identified nearly 100 genes with over 100 ribosome pausing sites in wild-type yeast. Subsequently we discovered mRNA with stronger secondary structure tend to have pausing ribosomes (p-value<2×10-16). Scikit-ribo almost perfectly reproduced relative codon dwell time from Weinberg et al (ρ=0.99) and found significant correlation between tRNA abundance and codon elongation rates (ρ=0.53). We also showed a balanced log2(TE) distribution after accounting for mRNA secondary structure and codon elongation rates, revealing the systematic bias in typical Riboseq analysis. Together, these results show that scikit-ribo provides robust methods for Riboseq analysis and better understanding of translational control.

K-MER BASED REFERENCE-FREE DETECTION OF FAMILY-PRIVATE VARIANTS HIGHLIGHTS THE GENETIC COMPLEXITY OF HHT

Andrew Farrell[1], Whitney Wooderchak-Donahue[2,3], Matt Velinder[1], Alistair N Ward[1], Pinar Bayrak-Toydemir[2,3], Gabor Marth[1]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]University of Utah, Department of Pathology, Salt Lake City, UT, [3]ARUP, Institute for Clinical and Experimental Pathology, Salt Lake City, UT, [4]University of Utah, Department of Radiology, Hereditary Hemorrhagic Telangiectasia Center, Salt Lake City, UT

We have previously shown that our reference free, k-mer based, variant detection method RUFUS has extremely high specificity and sensitivity for de novo variations of all types including SNPs, INDELs, and structural variations. Here we present a substantial extension of this method to identify low population-frequency, familial inherited variations, which allows us to accurately track disease-causing mutations through pedigrees, and pinpoint family-private disease-causing variants that segregate with affected/unaffected status. We applied this novel method for analyzing patients with hereditary hemorrhagic telangiectasia (HHT), an inherited disease known to be caused primarily by mutations in the genes ENG, ACVRL1, and SMAD4 (in addition to BMP9, which is associated with a phenotype similar to HHT). However, the genetic cause of the disease remains unexplained in approximately 15% of individuals identified as having HHT, despite extensive efforts to identify the causative variants with state-of-the-art existing tools. Here we present the results of our analysis of the 60X coverage Illumina whole genome sequencing data collected for 35 individuals from 13 distinct families, where previous causative variant identification methods have failed. To date, RUFUS was able to identify clear causative mutations in 7 of the 13 families: three families had a causative noncoding variant in the ENG or ACVRL1 genes that was missed by previous analyses. Two families had a deleterious variant in ACVRL1 intron 9 that ultimately disrupted splicing (confirmed by RNA sequencing), including one family with an ACVRL1 intron 9:chromosome 3 translocation (confirmed by PCR). Further confirmations are currently underway to identify additional HHT causative genes and genetic modifiers in the remaining 6 families. This means that our method was able to “solve” over half of the non-diagnostic cases, with several additional, promising hits being currently pursued, including novel mobile element insertions and small INDELs, missed by other methods, that may be disrupting splicing and gene regulation. Our methodological advances also reveal that noncoding variation plays a larger role in HHT than previously appreciated, and this is the first report to show the role of chromosomal translocation as a mechanism for the development of HHT.

INTRODUCING REFSEQ FUNCTIONAL ELEMENTS: A NEW DATASET ANNOTATED BY NCBI

Catherine M Farrell, Tamara Goldfarb, Sanjida H Rangwala, Kim D Pruitt, Terence D Murphy, RefSeq Development Team

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD

Eukaryotic genomes contain significant amounts of functional content, including conventional genes, which have long been a major focus for biomedical research and genome annotation projects. While conventional genic regions represent only a small fraction of eukaryotic genomes, other functional content is found in non-genic regions involved in processes such as gene regulation, chromosome organization, or DNA recombination, repair and replication. It is known that mutations in those regions can have functional consequences, and many genome-wide association studies show a predominance of disease-associated variation in non-genic regions. To characterize human non-genic regions, large-scale epigenomic mapping projects, including the ENCODE and Roadmap Epigenomics projects, have mapped gene regulatory elements on the reference genome, as predicted from chromatin states. However, the use of those maps can require specialized research knowledge and customized graphical displays, and thus they may not be readily apparent to all users. Furthermore, such epigenomic maps have not been reconciled with the traditional characterization of functional elements in the literature. In order to fill this gap and to provide accessible annotated data, NCBI is now introducing a new dataset of non-genic functional elements in human and mouse. This dataset includes: 1) functional elements that are experimentally validated in the literature; 2) element types that are not readily identifiable by large-scale epigenomic mapping projects, e.g., recombination hotspots; 3) NCBI annotation of each element on the human and mouse reference genome assemblies, with data available via NCBI graphical displays and FTP download; 4) richly curated RefSeq records with detailed feature annotation, including experimental evidence and publications; and 5) NCBI Gene records with detailed metadata, including publications, summaries and literature-based nomenclature. This new dataset will be highly visible and accessible to a wide array of users alongside NCBI's conventional gene annotation. It is expected to be highly useful for the interpretation of non-genic sequence variation. This presentation will provide further details on the phased release of this new dataset, and will include examples of curated non-genic functional elements and their correlation with genetic variation.

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

GENOMIC PATTERNS OF ACCELERATED EVOLUTION REVEAL NONCODING ELEMENTS THAT MAY REGULATE OVERT AND BIOMEDICALLY RELEVANT SPECIES-SPECIFIC TRAITS

Elliott C Ferris, Chris Gregg

University of Utah, Neurobiology and Anatomy, Salt Lake City, UT

ENCODE has uncovered thousands of putative noncoding regulatory elements in the human genome. In most cases however, we do not know which regulatory elements are important to which traits or disease processes, nor whether other important functional elements remain to be discovered in the genome. Here, we devise a phylogenomics approach to uncover important putative regulatory elements for various biomedically-relevant traits. Our approach involves analyzing genomic patterns of accelerated evolution in mammalian species with overt biomedically-relevant traits, including the elephant, naked mole rat, microbat, dolphin, orca and ground squirrel. We perform our analysis in genomic elements that are conserved across most mammals, including humans, which enriches for functional elements in the human genome. The elephant, orca and naked mole rat are of interest for their robust cancer resistance, the microbat and squirrel are hibernators that evolved reversible insulin resistance, and finally, the dolphin is a deep diving mammal that evolved mechanisms for breath holding and that prevent blood clot formation. By using a series of filters that refine the accelerated regions in each species to uncover regulatory elements that are most likely to shape the overt phenotypic traits of interest, our study presents a resource of putative master regulatory elements in the mammalian genome that shape traits relevant to cancer, type II diabetes, stroke and various other anatomical and physiological phenotypes.

THE EFFECT OF DECAY FACTOR KNOCKOUTS ON YEAST mRNA SYNTHESIS

Jonathan Fischer[1], Julia di Iulio[2], Mordechai Choder[3], Yun S Song[1,4,5], Nir Yosef[4], L. Stirling Churchman[2]

[1]UC Berkeley, Statistics, Berkeley, CA, [2]Harvard Medical School, Genetics, Boston, MA, [3]Technion-Israel Institute of Technology, Molecular Microbiology, Haifa, Israel, [4]UC Berkeley, EECS, Berkeley, CA, [5]University of Pennsylvania, Mathematics, Philadelphia, PA

Recent work in Saccharomyces cerevisiae has revealed that the processes of mRNA synthesis and decay are linked, meaning gene expression is circular. In particular, elements of the so-called “synthegradosome” travel between the nucleus, where they regulate transcription initiation, and the cytoplasm, where they degrade mRNAs. Using native elongating transcript sequencing (NET-seq), we explore the active transcription profile across the genome in eight distinct mutant yeast strains which lack genetic regions coding for specific mRNA decay factors. By recording the position of bound RNA polymerase II, this technology allows us to directly interrogate the effect of individual decay factors on transcription, and comparison of mutants with control samples permits inference of differentially expressed genes using a novel method we have developed. We subsequently perform enrichment analyses for a number of different biological annotations, including gene ontology (GO) and transcription regulatory complexes, among others. Of note are our observations of widespread enrichment signatures for genes preferentially associated with the SAGA transcriptional co-activator complex, and, in a subset of mutants, GO terms related to translation and cellular energy production. In addition to our gene-wide analysis, we probe transcriptional changes in gene sub-regions by exploiting the high resolution of NET-seq data, with a focus on regions adjacent the 5’ and 3’ ends of genes. Finally, we examine incongruities in antisense transcription putatively induced by the absent decay factors to determine what role, if any, these factors play in this mechanism of transcriptional regulation.

EFFECTS OF GENETIC VARIATION ON PROMOTER USAGE (pmQTL) AND ENHANCER ACTIVITY (enQTL)

Marco Garieri[1,2,3], Olivier Delaneau[1,2,3], Federico Santoni[1,4], David Mull[1], Piero Carninci[5], Emmanouil T Dermitzakis[1,2,3], Stylianos E Antonarakis[1,2,4], Alexandre Fort[1]

[1]University of Geneva, Department of Genetic Medicine and Development, Geneva, Switzerland, [2]Institute of Genetics and Genomics in Geneva, iGe3, Geneva, Switzerland, [3]Swiss Institute of Bioinformatics, SIB, Lausanne, Switzerland, [4]University Hospitals of Geneva, Service of Genetic Medicine, Geneva, Switzerland, [5]RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Japan

The identification of genetic variants affecting gene expression, splicing, chromatin states and transcription factor binding have increased our understanding of mechanisms underlying human traits and diseases. We hypothesized that some eQTLs (expression quantitative trait loci) act at the level of differential promoter usage and enhancer activity, two molecular phenotypes that can be quantified with the CAGE technology.
Transcriptomes of 154 unrelated individuals were profiled using CAGE on total nuclear RNAs extracted from EVB transformed lymphoblastoide cell lines. Sequences were mapped to promoter regions of the FANTOM atlas, yielding to the quantification of 38,759 promoters/transcripts that we tested for association against 7,508,202 genetic variants: we discovered 5,491 promoter-QTL (pmQTL, *FDR*<0.05) in *cis*. As for eQTLs, pmQTLs localize preferentially near to transcriptional start sites, within open chromatin regions and are marked with active transcriptional histone marks. Approximately 90% of the pmQTLs are found to also affect gene mRNA levels ($\pi 1$=0.904) as quantified with RNA-seq, half of them are not associated with the main but with alternative promoters and 26.3% of the associated genes have more than one promoter linked with a pmQTL. Opposite regulatory effects of pmQTL were detected on two or more promoters for 139 genes. The integration of pmQTL with eQTL allows 1) the evaluation of the relative participation of alternative promoters to mRNA abundance and 2) the detection of variants associated with promoter usage and not with mRNA levels. Taken together this gives insights into eQTL mechanisms involving differential promoter usage.
Furthermore, using the FANTOM enhancer atlas as a reference and the quantification of enhancer RNAs (eRNAs) as a proxy for enhancer activity, we mapped 108 enhancer-QTL (enQTL, *FDR*<0.05) in *cis*. For each enQTL-enhancer pair, we tested causal inference of eRNAs as molecular mediators for the expression of enhancer target genes, using causal inference testing. Causality was detected for 48 triplets (causal inference test *p-value*<0.05). This approach provides insights into eQTL mechanisms delineating effects on both distant enhancers (eRNA levels) and proximal promoters (mRNA levels).

SYSTEMS HUMAN GENOME AND METAGENOME ANALYSIS ON CIRCULATING PROTEINS IN A POPULATION COHORT

Daria V Zhernakova[1,2], Alexander Kurilshikov[1], Biljana Atanasovska[1,3], Trang Le[1], Marc Jan Bonder[1], Serena Sanna[1], Rudolf Boer[4], Folkert Kuipers[3], Lude Franke[1], Cisca Wijmenga[1], Alexandra Zhernakova[1], Jingyuan Fu[1,3]
[1]University of Groningen, Department of Genetics, Groningen, the Netherlands, [2]St. Petersburg State University, Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg, Russia, [3]University of Groningen, Department of Pediatrics, Groningen, the Netherlands, [4]University of Groningen, Department of Cardiology, Groningen, the Netherlands

Proteins circulating in blood are often measured as biomarkers for various diseases including immune diseases, cancers and cardiovascular diseases (CVD). However, the inter-individual variation of these circulating proteins in the general population is largely unknown, as are the factors underlying this variation. We have now measured serum levels of 92 CVD-relevant proteins in 1,294 individuals from a general Dutch population cohort (LifeLines-DEEP) for whom we also have data on the human genome and “the second human genome”: the metagenome.

For each protein, we performed genome-wide analysis with 8 million SNPs and metagenome-wide association analysis with 340 bacterial species and 702 functional pathways determined by metagenomics sequencing. At FDR 0.05, we identified 72 proteins that were genetically controlled and 51 proteins associated to the gut microbiome. Serum levels of 37 proteins were affected by both genetics and microbiome. C-C motif chemokine 15 (CCL15), for example, is a liver-derived chemokine involved in immunoregulatory and inflammatory processes. In addition to its strong genetic regulation (association to rs854626 at $P=2.5x10^{-136}$), an elevated serum level of CCL15 was also associated to higher bacterial capacity for fatty acid biosynthesis ($P=5.3x10^{-4}$). We further confirmed the causal effect of fatty acids on CCL15 production by stimulating hepatocytes (HepG2) with free fatty acids, observing a 40% increase in CCL15 expression 24 hours after stimulation. Fourteen proteins were more affected by the gut microbiome than by genetics. For instance, adipose-derived cytokine PAI-1 is strongly associated with obesity and its elevation is also a risk factor for atherosclerosis. While we did not detect significant associations with genetics, serum levels of PAI-1 were not only associated to a lower richness of bacterial species ($P=9.7x10^{-4}$) but also to 138 bacterial function pathways, in particular to bacterial energy metabolism.

By using 92 CVD-related circulating proteins, we demonstrate that serum proteomics are affected by both genetics and gut microbiome. Our data suggests that both the human genome and metagenome should be taken into account when using circulating proteins as potential biomarkers for disease monitoring or as therapeutic targets for personalized medicine.

ADAPTIVE EVOLUTION OF MENTAL ACTIVITY-RELATED STX GENE IN THE OUT-OF-AFRICA MIGRATION

Naoko Fujito[1], Yoko Satta[1], Masaya Hane[2], Atsushi Matsui[3], Kenta Yashima[1], Ken Kitajima[2], Chihiro Sato[2], Naoyuki Takahata[1], Toshiyuki Hayakawa[4] ‘
[1]SOKENDAI, School of Advanced Sciences, Hayama, Japan, [2]Nagoya University, Bioscience and Biotechnology Center, Nagoya, Japan, [3]Kyoto University, Primate Research Institute, Kyoto, Japan, [4]Kyushu University, The Graduate School of Systems Life Sciences, Kyusyu, Japan

It is now reported that a number of genes have undergone adaptive evolution since anatomically modern humans (AMHs) migrated out of Africa. Yet, no such evidence has been found in any gene that is involved in mental activities. It is however conceivable and even likely that AMHs faced mental challenges in the out-of-Africa migration as well as in subsequent new settlements. Here we examine this possibility focusing on a gene that encodes STX, a transferase of polysiallic acids to neural adhesion molecules. The STX gene is known to be associated with schizophrenia when overexpressed. There exist three core SNPs that can primarily alter the STX promoter activity. These core SNPs define four haplotypes in the current human populations, of which one haplotype, denoted as CGC, is prevalent only in East Asians (though to a lesser extent in South Asians and Americans as well). We first carried out the promoter assay of the four haplotypes, demonstrating significantly low promoter activity of the CGC. Furthermore, determining 63 haplotype sequences for a world-wide sample, we estimated that the CGC originated ~0.5 MYA and diverged 0.1~0.2 MYA in Africa. We also tested the 1000 genome data in terms of SFS (site frequency spectrum) and ROH (runs of homozygosity) in a 200 kb region surrounding the core SNPs. The CGC haplotype tends to carry longer ROH than the alternative, ancestral haplotypes. We evaluated the difference in ROH between the CGC and the remaining haplotypes by an analog of Welch’s t statistic and found that the observed t value is significantly larger than expected under neutrality. The SFS exhibited an excess of rare alleles in all populations examined, but it also showed an excess of intermediate frequency alleles only in the East Asian population. While the former observation could be explained by population expansion, the latter could not by bottlenecks or any other demographic causes if neutrality was assumed. Together with the result of rEHH analyses, all these results can be best explained by an ongoing soft sweep of the CGC in East Asian populations, thus providing the first evidence for positive selection on a gene associated with mental activities of AMHs. Such selection for the CGC haplotype began to act around 10 ~ 20 KYA. It appears that after the last glacial period was over, populations in the north and south routes had extended their geographic areas and encountered each other in East Asia. It is such encountering of culturally differentiated populations that had posed mental challenges for AMHs.

ASSESSMENT OF K-MER FILTERING METHOD FOR FALCON GENOME ASSEMBLY

Arkarachai Fungtammasan, Carrie Jiang, Brian Rogoff, Arina Malouka, Aleksandra Zalcman, Brett Hannigan*

DNAnexus, Science, Mountain View, CA

Recent progress in long read sequencing technology has allowed researchers to create de novo assemblies with impressive continuity. Most assembly algorithms that leverage long reads use an overlap, layout, and consensus (OLC) model where each read is compared and potentially aligned with every other read. This framework is powerful but also computationally expensive, especially in highly repetitive genomes where initial matches in repetitive regions lead to unfruitful alignment attempts for reads that do not truly align. Current approaches to minimize effort spent in aligning non-overlapping reads that share repetitive content often involve identifying putative repetitive elements by performing alignments on subsets of the full data. While these approaches have shown good results, they still rely on performing read alignments which can consume significant computational resources.

In this study, we investigate whether kmer information from the initial raw-reads can be used to make the assembly process more efficient. In particular we look at the observed k-mer frequencies in a variety of organisms, including E. coli, arabidopsis, and human. We examine the observed k-mers in three separate stages of the assembly process: 1) the raw sequences produced by a Pacific Biosciences instrument 2) the error-corrected reads formed from a consensus of overlapped raw data and 3) the final assembly. We then classify k-mers by those that are present in both the initial raw-reads and the error corrected reads / final assembly, and those k-mers that are unique to the raw-reads. We hypothesize that many of these k-mers represent errors in the initial sequencing, or are otherwise unable to be correctly assembled. Moreover, there are other k-mers which are represented much more frequently than one would expect given the coverage depth of sequencing. These k-mers likely are part of repetitive elements.

By masking these problematic k-mers early in the assembly process, we hope to avoid performing unnecessary and costly alignment operations during the overlap portion of the assembly. We therefore examine various classification schemes to detect these problematic k-mers, leveraging features such as Shannon entropy, BiEntropy, homopolymer length, edit distance, and compression ratio, and show how these features can be used to separate our k-mer classes. We then use these classifiers to mask k-mers in the raw-reads and evaluate the effect on assembly speed, resources usage, and integrity of the final assembly.

*Corresponding author

PREDICTING OFF-TARGET EFFECTS FOR END-TO-END CRISPR GUIDE DESIGN

Jennifer Listgarten[1], Michael Weinstein[2], Melih Elibol[1], Luong Hoang[1], John Doench[3], Nicolo Fusi[1]

[1]Microsoft Research, Cambridge, MA, [2]UCLA, Los Angeles, CA, [3]Broad Institute, Cambridge, MA

The CRISPR-Cas9 system provides unprecedented genome editing capabilities. However, off-target effects lead to sub-optimal usage and additionally are a bottleneck in development of therapeutic uses. Herein, we introduce the first machine learning-based approach to this problem, yielding a state-of-the-art predictive model for CRISPR-Cas9 off-target effects which outperforms all other guide design services. Our approach, Elevation, consists of two inter-related machine learning models--one for scoring individual guide-target pairs and another which aggregates guide-target scores into a single, overall guide summary score. Through systematic investigation, we demonstrate that Elevation performs substantially better than competing approaches on both of these tasks. Additionally, we are the first to systematically evaluate approaches on the guide summary score problem; we show that the most widely-used method (and one re-implemented by several other servers) performs no better than random at times, whereas Elevation consistently outperformed it, sometimes by an order of magnitude. In our analyses, we also introduce a method to balance errors on truly active guides with those which are truly inactive, encapsulating a range of practical use cases, thereby showing that Elevation is consistently superior across the entire range. We thus contribute a new evaluation metric for benchmarking off-target modeling. Finally, because of the large computational demands of our tasks, we have developed a cloud-based service for end-to-end guide design which incorporates our previously reported on-target model, Azimuth, as well as our new off-target model, Elevation.

RETROCOPIES CONTRIBUTION TO THE CANCER GENES REPERTORY.

Fernanda Orpinelli[1], Thiago A Miller[1,2], Helena B Conceicao[1], Anamaria A Camargo[1], Pedro A Galante[1]

[1]Hospital Sírio-Libanês, Centro de Oncologia Molecular, Sao Paulo, Brazil, [2]Universidade de São Paulo, Bioquímica, Sao Paulo, Brazil

mRNA-derived duplicates, also known as retrocopies, have been emerging as an important regulators of gene expression. However, studies of retrocopies in physiological and pathological states remain elusive. Here, we perform a systematic study of retrocopies of cancer associated genes in humans. Our results indicate that 66 cancer associated genes (cancer genes) have significantly more retrocopies (400) than expected by chance, being tumor suppressor genes with more retrocopies than proto-oncogenes. To further understand the cancer genes retrocopies functions, we examined their transcription potential and conservation. First, we found a set of retrocopies highly or moderately expressed in normal tissues. Next, we found expressed retrocopies under positive or negative selection. Finally, we check for retrocopies and their parental genes expression and expression dysregulation in cancer tissues. We found retrocopies only expressed and cancer tissues, as well as retrocopies under- and over-expressed in tumor tissues, suggesting a potential functional role of those retrocopies. Overall, our data reinforce that retrocopies are important players in terms of gene expression regulation and that their dysregulation may have implications in driving tumorigenesis.

AN INDUCIBLE LONG NONCODING RNA AMPLIFIES DNA DAMAGE SIGNALING

Julia T Garcia*[1], Adam M Schmitt*[1,2,7], Tiffany Hung*[1], Ryan A Flynn[1], Ying Shen[1], Kun Qu[1], Alexander Y Payumo[3,4], Ashwin Peres-da-Silva[1], Daniela Kenzelmann Broz[5], Rachel Baum[7], Shuling Guo[6], James K Chen[3,4], Laura D Attardi[2,5], Howard Y Chang[1]

[1]Stanford University School of Medicine, Center for Personal Dynamic Regulomes, Stanford, CA, [2]Stanford University School of Medicine, Department of Radiation Oncology, Stanford, CA, [3]Stanford University School of Medicine, Department of Chemical and Systems Biology, Stanford, CA, [4]Stanford University School of Medicine, Department of Developmental Biology, Stanford, CA, [5]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [6]Ionis Pharmaceuticals, Department of Antisense Drug Discovery, Carlsbad, CA, [7]Memorial Sloan Kettering Cancer Center, Department of Radiation Oncology, New York, NY

Long noncoding RNAs (lncRNAs) are extensively transcribed genes with exquisite regulation but mostly unknown functions. We demonstrate a role of lncRNAs in guiding organismal DNA damage response. DNA damage activates transcription of *DINO (Damage Induced NOncoding)* via p53. DINO is required for p53-dependent gene expression, cell cycle arrest, and apoptosis in response to DNA damage, and DINO expression suffice to activate damage signaling and cell cycle arrest in the absence of DNA damage. *Dino* knockout or promoter inactivation in mice dampens p53 signaling and ameliorates acute radiation syndrome in vivo. Thus, inducible lncRNA can create a feedback loop with its cognate transcription factor to amplify cellular signaling networks. We are investigating heterogeneous p53-Dino dependent response to DNA damage using a *Dino*^{gfp} knockin allele. This knockin reports DNA damage and shows *Cdkn1a* expression is regulated in cis and trans by *Dino*. GFP intensity is used to track DNA damage dynamics in individual cells, and signature differences in chromatin accessibility between low and high GFP expressing cells will be determined in bulk cell populations and single cells.

RECENT EVOLUTION OF THE EPIGENETIC REGULATORY LANDSCAPE IN HUMAN AND OTHER PRIMATES

Raquel Garcia-Perez[1,2], Gloria Mas-Martin[2,3], Meritxell Riera[1,2], Antoine Blancher[4], Marc Marti-Renom[2,3,5], Luciano Di Croce[2,3,5], David Juan[1,2], Jose Luis Gómez-Skarmeta[6], Tomas Marques-Bonet[1,2,5]

[1]Institut de Biologia Evolutiva, CSIC-UPF, Barcelona, Spain, [2]Universitat Pompeu Fabra, UPF, Barcelona, Spain, [3]Centro Nacional de Análisis Genómico - Centro de Regulación Genómica, CNAG-CRG, Barcelona, Spain, [4]Laboratoire d'Immunogenetique moleculaire, Faculte de Medecine Purpan, Universite Toulouse 3, Toulouse, France, [5]Catalan Institution of Research and Advanced Studies, ICREA, Barcelona, Spain, [6]Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Sevilla, Spain

Evolutionary biologists have long sought to discern the molecular basis of phenotypic and genomic variation. Changes in gene regulation are thought to play a major role in evolution and speciation, particularly in primates.. Over the last decade, the field has experienced a major shift towards inter-species comparative epigenomics in search of a conceptual step-forward in our understanding of evolution. However, the lack of coherent multi-omic datasets has hindered the integrative study of the interplay between epigenomic and genomic evolution in different species.

Our study aims to characterize the evolutionary dynamics of regulatory elements in the human and primates at a fine scale. To that end, we have comprehensively profiled lymphoblastoid cell lines (LCLs) from human, chimpanzee, gorilla, orangutan and macaque. This epigenomic characterization includes whole genome sequencing data, whole genome bisulfite data, deep-transcriptome sequencing data, chromatin accessibility data (ATAC-seq) as well as ChIP-seq data from five key histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27ac and H3K27me3), CTCF and HIC conformations. We take advantage of our data to: 1) improve the annotation of cis-regulatory elements, particularly for the non-human species; 2) investigate the contribution of inter-species differences in methylation levels and histone modifications to gene expression variation; 3) understand how the interplay between different regulatory elements evolves in the different lineages; 4) study how the combination of regulatory and sequence evolution shapes different phenotypic scenarios in a lineage-specific manner.

The results of this work will contribute to enlighten human specific regulatory innovations in the context of our closest extant relatives showing the degree of coordination of recent epigenome and genome evolution. Moreover, our data and analyses provide a unique and valuable interdisciplinary resource of major interest for the scientific community.

ANALYSIS OF MOBILE ELEMENT INSERTION (MEI) DISCOVERY IN THE GENUS CANIS PROVIDES INSIGHTS ON MEI DISTRIBUTION, EVOLUTION, AND IMPACT ON TRANSCRIPTION AND DISEASE

Eugene J Gardner[1,2], Brian W Davis[3], Jasmine B Baker[4], Cody J Steely[4], Jerilyn A Walker[4], Timothy D O'Connor[2,5,6,7], Mark A Batzer[4], Elaine A Ostrander[3], Scott E Devine[1,2,6,7]

[1]University of Maryland Baltimore, Program in Molecular Medicine, Baltimore, MD, [2]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [3]National Human Genome Research Institute, National Institutes of Health, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, [4]Louisiana State University, Department of Biological Sciences, Baton Rouge, LA, [5]University of Maryland School of Medicine, Program in Personalized and Genomic Medicine, Baltimore, MD, [6]University of Maryland School of Medicine, Greenebaum Cancer Center, Baltimore, MD, [7]University of Maryland School of Medicine, Department of Medicine, Baltimore, MD

Mobile elements (MEs) are autonomous DNA parasites that can generate new copies, or mobile element insertions (MEIs), in their host’s genomes. In the genus Canis, two families of MEs, L1 and CanSINE, have been particularly efficacious and have generated tens of thousands of polymorphic MEIs per individual. In order to discover such polymorphic sites, we adapted the Mobile Element Locator Tool (MELT) to canine genomes and performed MEI analysis on a cohort of 434 breed dogs, village dogs, wolves, and coyotes. We discovered and genotyped a total of 313,074 polymorphic MEIs, and identified 108,250 MEIs that were located within genes, including 461 insertions that directly mutagenized exons. With this collection, we analyzed recent ME activity in Canis, established active subfamilies, and assessed the genomic distributions of these elements. Additionally, we identified several thousand breed-specific variants that potentially might include MEIs that contributed to breed formation, phenotypic variation, or diseases. Since only 60 breed-specific sites directly impacted coding exons, we sought to determine how the more abundant intronic sites might play a role in canine phenotypes or diseases. As it has been shown that such intronic sites can cause transcription defects, we analyzed breed specific RNA-seq data for MEIs that triggered alternative splicing events. This analysis identified a number of both fixed and polymorphic MEIs that cause a wide range of effects on transcript structure, providing a putative mechanism for intronic MEIs in the etiology of several disorders. Overall, our study is the first large-scale analysis of MEIs in the genus Canis, and provides an extensive resource to study the impact of MEIs on genome evolution. Likewise, our collection provides a novel resource to model human diseases in canines.

DISSECTION OF CELLULAR HETEROGENEITY IN AN INNATE IMMUNE MODEL OF TRANSCRIPTIONAL REGULATION USING SINGLE CELL RNA SEQUENCING.

Kyle Gellatly[1], Pranitha Vangala[1], Sean McCauley[2], Elisa Donnard[1], Patrick McDonel[1], Alan Derr[1], Jeremy Luban[2], Manuel Garber[1]

[1]University of Massachusetts Medical School, Bioinformatics and Integrative Biology, Worcester, MA, [2]University of Massachusetts Medical School, Molecular Medicine, Worcester, MA

In vitro derived macrophages (MPs) and dendritic cells (DCs) have been used extensively as models to study the dynamics of gene regulatory networks in response to environmental stimuli; including extensive chromatin remodeling and changes in transcriptional outputs. Until recently, most studies with mouse bone marrow derived dendritic cells (BMDCs) have utilized bulk populations with the underlying assumption that these populations are homogeneous. As a result, the average gene expression level should be a faithful representation of all the cells in the culture. If this cell homogeneity is violated however, conclusions drawn from bulk studies evaluating gene regulatory networks would be flawed.

Recent publications have suggested that *in vitro* derived DCs are indeed heterogenous. Here we dissect the composition of *in vitro* derived innate immune cell cultures using a droplet barcoding single cell RNA-sequencing (scRNA-seq) method (Zilionis, et al. 2017). Our analysis suggests that this culture system contains a diverse mixture of cell types including DCs, MPs, and granulocytes, and that communication between these cell types may mold their response to environmental stimuli. For example, DCs seem to respond through paracrine signaling to the cytokines released by macrophages in the culture. Within defined cell types there is further heterogeneity. Some macrophages in these cultures lack all major histocompatibility complex (MHC) markers. Another population of macrophages displays an inverse relationship in the expression of MHC-II proteins and the master metabolic regulator fatty acid binding proteins. These contrasting expression profiles may reflect unique roles for these subtypes, such as immune surveillance compared to homeostatic functions. Proper identification of immune cell sub-populations within these cultures is essential for a full understand of DC biology and to be able to build more accurate gene regulatory networks in response to environmental stress. Single cell approaches enable a more informed view of the complex biology of gene regulation.

HUMAN PLURIPOTENT STEM CELLS RECURRENTLY ACQUIRE AND EXPAND DOMINANT NEGATIVE P53 MUTATIONS

Sulagna Ghosh*[1,2,4], Florian Merkle*[1,2,3], Nolan Kamitaki[2,4], Jana Mitchell[1,2], Yishai Avior[5], Curtis Mello[2,4], Giulio Genovese[2,4], Nissim Benvenisty[5], Steven McCarroll[2,4], Kevin Eggan[1,2]

[1]Harvard University, Stem Cell and Regenerative Biology, Cambridge, MA, [2]Broad Institute, Stanley Center for Psychiatric Research, Cambridge, MA, [3]University of Cambridge, Clinical Biochemistry, Cambridge, United Kingdom, [4]Harvard Medical School, Department of Genetics, Boston, MA, [5]Hebrew University of Jerusalem, The Azrieli Center for Stem Cells and Genetic Research, Jerusalem, Israel

* These authors contributed equally

Human pluripotent stem cells (hPSCs) hold great promise for disease modeling and regenerative medicine due to their unlimited potential for self-renewal and ability to differentiate into multiple somatic lineages. While previous studies have identified large copy number variants (CNVs) that recur during culture adaptation of hPSCs, the nature, frequency and functional impact of acquired sequence mutations remains unclear. Since mutations acquired in culture could have significant implications for clinical applications as well as experimental models, there is an urgent need to assess the genomic integrity of hPSCs at a finer scale. Here, we performed high coverage exome sequencing of 140 independent human embryonic stem cell (hESC) lines, including 26 clinical-grade lines. To uncover acquired rather than inherited variation, we applied computational strategies to identify mutations present at allelic fractions less than 50%, indicating their presence in a subset of cells. Our analyses revealed that while mosaic variants were generally rare across cell lines, five unrelated hESC lines had acquired six distinct mutations in *TP53*, a gene that encodes the tumor suppressor P53. Notably, the six *TP53* mutations identified mapped precisely to four codons in the DNA binding domain most commonly disrupted in human cancers. These mutations are known to impair the function of wild type P53 in a dominant negative manner. We observed that the *TP53* mutations conferred a strong selective advantage to hESCs and cells harboring *TP53* mutations rapidly outcompeted non-mutant cells within the same hESC line. To confirm the reproducibility of our findings, we mined 57 studies containing published RNA sequencing data from 117 hPSC lines and observed an additional nine *TP53* mutations, all resulting in coding changes in the DNA binding domain of P53. Interestingly, cell lines with P53 mutations readily differentiated into multiple cell types and in at least one case, the mutant allelic fraction increased during differentiation, suggesting that the selective advantage can persist in differentiated cells. Together, our findings indicate that hPSCs can spontaneously acquire and expand cancer-associated mutations, raising concerns for both transplantation medicine and disease modeling. Since such mutations are likely to escape detection by routine quality control measures, we suggest careful genetic screening of hPSCs and their differentiated derivatives prior to clinical and research use.

OLIGOGENIC INHERITANCE OF FAMILIAL CARDIAC DISEASE INVOLVING MKL2, MYH7 AND NKX2-5 VARIANTS

Casey Gifford, Ryan Samarakoon, Hazel Salunga, Yu Huang, Kathy Ivey, Deepak Srivastava

J. David Gladstone Institutes, Institute for Cardiovascular Disease, San Francisco, CA

The ability to parse the genetic causes of oligogenic disorders has been challenging. Recent advances in genome sequencing and editing now provide an opportunity to determine and experimentally test contributions of multiple genetic variants in human disease. Here, we report a case of familial left ventricular noncompaction (LVNC), a congenital heart defect characterized by a persistent immature myocardium, with an inherited oligogenic cause. Exome sequencing of family members identified a novel heterozygous missense mutation in the transcription factor MKL2 (MRTF-B) that segregated with the disease and exhibited reduced transcriptional activity in vitro. Mice homozygous for this mutation exhibited abnormalities in the endocardium and ventricular myocardial wall. However, heterozyosity was not lethal or sufficient to cause a LVNC-like phenotype in mice. Further exome analysis revealed a rare heterozygous missense mutation in NKX2-5 among subjects with early onset disease that was inherited from a healthy family member and had reduced DNA-binding activity in vitro. We additionally identified a novel missense mutation in MYH7, an essential protein that localizes to the sarcomere in cardiomyocytes. Mice homozygous for the Nkx2-5 and Myh7 variants were embryonic lethal, while heterozygous mice were mostly normal. Interestingly, compound heterozygous mice (Mkl2+/–Myh7+/- Nkx-2-5+/–) were born at the expected Mendelian frequency, but they exhibited symptoms of an immature and dysfunctional myocardium including enlarged papillary muscles, hypertrabeculated ventricular walls, and right ventricular expansion, mimicking the phenotypes observed in the familial case of LVNC. Transcriptome analysis from murine hearts of compound heterozygous mice identified genes associated with cellular adhesion and proliferation when compared to wild type mice. Myocardial differentiation of patient-specific human induced pluripotent stem cells revealed transcriptional dysregulation of genes involved in cardiovascular development and cellular adhesion, similar to the murine hearts. By integrating transcriptional profiles derived from in vivo and in vitro models of disease, we provide experimental evidence for complex inheritance of a human disease and reveal novel mechanisms underlying the development of a congenital heart defect.

DETECTING MUTATIONAL SIGNATURES ASSOCIATED WITH BRCA1/2 DEFICIENCY FROM DNA AND RNA SEQUENCING DATA

Dominik Glodzik[1,2], Helen Davies[2], Johan Staaf[1], Serena Nik-Zainal[2]

[1]Lund University, Faculty of Medicine, Lund, Sweden, [2]Wellcome Trust Sanger Insittute, Cancer Aging and Somatic Mutation, Hinxton, United Kingdom

Genomic instability fuels genetic diversity of cells, and is a common feature in cancer. For some patients, genomic instability is attributed to impairment of homologous recombination (HR) through mutations in BRCA1/2 genes. Tumours with inactivation of BRCA1/BRCA2 genes acquire in their cancers an excess of base substitutions and chromosomal rearrangements. The combination of mutational signatures is specific enough to HR-deficient cancers to allow us to use the data on mutational patterns in DNA to build a predictor of BRCAness (1). The predictor, named HRDetect, assigns a BRCAness score to a patient's cancer, and we hypothesize that high scores highlight tumours selectively sensitive to PARP-inhibition.

We intend to apply HRDetect to a population-based retrospective study of triple negative breast cancer in Sweden SCAN-B (ClinicalTrials.gov ID: NCT02306096). While some of the patients' samples will be characterized through whole-genome DNA sequencing, the majority has been subject to RNA sequencing only. In order to apply HRDetect to this large cohort of 7,000 patients, we are re-designing the classifier to work with data from RNA sequencing, including gene expression levels and somatic mutations detected from RNAseq.

Here we describe and compare the accuracy of classifiers that use DNA or RNA data. Based on RNA sequencing data, we identify mutational features of tumours deficient in homologous recombination. We will correlate the HRDetect scores with clinical data in the SCAN-B cohort.

References:

(1) Davies, H and Glodzik, D. et. al. HRDetect: A mutational signature based predictor of BRCA1 and BRCA2 deficiency, in press with Nature Medicine.

RRBS-SMART: A METHOD FOR STUDYING THE METHYLOME AND TRANSCRIPTOME OF LOW-INPUT TISSUE SAMPLES AND SINGLE CELLS AT ONCE

Hongcang Gu[1], Kendell Clement[1,2,3], Aleksandra Arczewska[1,2,3], Zachary Smith[1,2,3], Jiantao Shi[4], Alexander Tsankov[1,2,3], Martin Aryee[1,5], Andreas Gnirke[1], Alexander Meissner[1,2,3,6]

[1]Broad Institute, Epigenomics Program, Cambridge, MA, [2]Harvard University, Dept. of Stem Cell and Regenerative Biology, Cambridge, MA, [3]Harvard Stem Cell Institute, Cambridge, MA, [4]Dana-Farber Cancer Institute, Dept. of Biostatistics and Computational Biology, Boston, MA, [5]Massachusetts General Hospital, Dept. of Pathology, Charlestown, MA, [6]Max-Planck Institute of Molecular Genetics, Berlin, Germany

Cytosine methylation in CpG dinucleotides of mammalian genomes is associated with gene expression profiles and functional states of cells and tissues in normal development and in disease. However, the relationship between DNA methylation and transcriptional gene regulation is complex, depending on the genomic contexts, such as promoters, gene bodies and regulatory regions as well as cell or tissue type.

To investigate the DNA methylome and RNA transcriptome landscapes of precious low-input tissue and cell samples simultaneously, we developed a protocol that combines multiplex reduced representation bisulfite sequencing (RRBS) with SMART RNA-Seq: after cell lysis, mRNA is pulled-down using an biotinylated oligo(dT) primer attached to streptavidin beads and converted to a SMART-seq2 RNA-Seq library. Genomic DNA in the supernatant is cleaned up, restriction digested, ligated to sequencing adapters, bisulfite-converted and amplified. Manual RRBS-SMART allows processing 96 low-input samples such as micro-dissected tissues from early mouse embryos or flow-sorted single cells within 3-4 days.

We used this method to study the interplay between global changes of DNA methylation and RNA expression in early mouse embryos as well as in heterogeneous populations of human embryonic stem cells grown in culture. We will present detailed performance metrics of RRBS-SMART as well as biological findings from our studies.

DIVERSE NON-GENETIC ALLELE SPECIFIC EXPRESSION EFFECTS SHAPE GENETIC ARCHITECTURE AT THE CELLULAR LEVEL IN THE MAMMALIAN BRAIN

Wei-Chao Huang[1], Elliott Ferris[1], Tong Cheng[1], Cornelia Stacker Horndli[1], Kelly Gleason[2], Carol Tamminga[2], Janice Wagner[3], Kenneth Boucher[4], Jan Christian[1], Christopher Gregg[1,5]

[1]University of Utah, Neurobiology & Anatomy, Salt Lake City, UT, [2]UTSouthwestern, Department of Psychiatry, Dallas, UT, [3]Wake Forest School of Medicine, Department of Pathology, Winston-Salem, NC, [4]University Of Utah, Cancer Biostatistics, Salt Lake City, UT, [5]New York Stem Cell Foundation, NYSCF Robertson Investigator, New York, NY

Interactions between genetic and epigenetic effects shape brain function, behavior and the risk for mental illness. Random X-inactivation and genomic imprinting are epigenetic allelic effects that are well known to influence genetic architecture and disease risk. Less is known about the nature, prevalence and conservation of other potential epigenetic allelic effects in vivo in the mouse and primate brain. Here, we devise genomics, in situ hybridization and mouse genetics strategies to uncover diverse allelic effects in the brain that are not caused by imprinting or genetic variation. We found allelic effects that are developmental stage and cell-type specific, prevalent in the neonatal brain and cause mosaics of monoallelic brain cells that differentially express wildtype and mutant alleles for heterozygous mutations. Finally, we show that diverse non-genetic allelic effects exist in the macaque and human brain that impact mental illness risk genes. Our findings have potential implications for mammalian brain genetics.

COMPREHENSIVE SEQUENCING ANALYSIS OF HIGH-THROUGHPUT SINGLE T CELLS IN HUMANS: αβ CHAIN PAIRING AND ALLELIC INCLUSION

Kristina Grigaityte[1], Jason Carter [2], Adrian Briggs[3], Stephen Goldfless[3], Sonia Timberlake[3], David Koppstein[3], George Church[4], Francois Vigneault[3], Gurinder Atwal[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Stony Brook University, School of Medicine, Stony Brook, NY, [3]JUNO Therapeutics, TCR Technology, Seattle, WA, [4]Harvard Medical School, Genetics, Boston, MA

A diverse T cell repertoire is a critical component of the adaptive immune system, providing protection against invading pathogens and tumors. T cell receptors (TCRs) – the main signature of a T cell involved in antigen recognition – consist of a heterodimer of one α and one β chain. A diverse T cell repertoire arises from numerous combinations of different α and β chains, potentially generating up to 10^{15} distinct TCRs. However, identifying αβ pairs and determining full TCR sequences in a high-throughput fashion is challenging. Recent studies of the T cell repertoire have typically focused on characterizing either the α or β chain alone by bulk sequencing. Therefore, our understanding of statistical properties and selection of the T cell pool in an individual is currently limited, hampering efforts to accurately quantify T cell clonal changes in disease and immunotherapy.

We present comprehensive analyses of the paired αβ T cell repertoire in the peripheral blood of healthy humans by leveraging recent biotechnology developments in deep RNA sequencing of hundreds of thousands of lymphocytes by single-cell barcoding in emulsion. The T cells were stratified into the two major subtypes, CD4 helper cells and CD8 cytotoxic cells. The clonal distributions of the repertoire exhibited a universal power law, with longer tails for the CD8 cells. Furthermore, we report statistical associations between gene usage across α and β chains, suggesting that αβ pairing may be a significantly non-random process. These findings contradict the widely used assumption that the β chain diversity alone is an accurate representation of the T cell repertoire.

Finally, the high-throughput single-cell sequencing technology provides a unique opportunity to ascertain rates of allelic inclusion by quantifying coexpression of productive alleles in individual cells. We report a consistently higher level of allelic inclusion rates with regard to the α chain than the β chain across samples, supporting the known T cell selection mechanism in the thymus.

Together, our results highlight the critical need to assay αβ pairing and allelic inclusion in single cells to accurately profile the landscape of the T cell repertoire and to monitor changes in the clonal distribution over time.

SEX-INTERACTING eQTLs IN HUMAN SKELETAL MUSCLE

Li Guan[1], D. Leland Taylor[2,3], Ryan P Welch[4], Michael R Erdos[2], Arushi Varshney[5], Anne U Jackson[4], Peter S Chines[2], Narisu Narisu[2], Heather M Stringham[4], Lori L Bonnycastle[2], Markku Laakso[6], Jaakko Tuomilehto[7], Heikki A Koistinen[8], Michael Boehnke[4], Francis S Collins[2], Stephen C Parker[1,5], Laura J Scott[4]

[1]University of Michigan, Dept. of Computational Medicine & Bioinformatics, Ann Arbor, MI, [2]National Institutes of Health, National Human Genome Research Institute, Bethesda, MD, [3]European Bioinformatics Institute,Wellcome Trust Genome Campus, European Molecular Biology Laboratory, Cambridgeshire, United Kingdom, [4]University of Michigan, Dept. of Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, [5]University of Michigan, Dept. of Human Genetics, Ann Arbor, MI, [6]University of Eastern Finland, Dept. of Medicine, Kuopio, Finland, [7]National Institute for Health and Welfare, Chronic Disease Prevention Unit, Helsinki, Finland, [8]National Institute for Health and Welfare, Dept. of Health, Helsinki, Finland

The levels of many phenotypic traits and diseases differ by sex, and genetic effects that vary by sex can contribute to these differences. In this study we ask if effect of a SNP on skeletal muscle gene expression varies by sex. We performed high depth RNA-sequencing and array-based genotyping followed by imputation using the GOT2D panel in 267 Finnish participants (110 females and 157 males). We used an additive genetic model to test for association of inverse normalized gene expression (pre-adjusted for age, batch and other confounders) and included an interaction term between SNP genotype and sex to test for a difference in genetic effect by sex. We tested SNP-gene pair where the SNP is < 1Mb away from the transcription start site. No SNP showed differential SNP effects by sex (using FDR < 0.05 threshold). We compared our results to Kukurba et al. (2016) obtained by analyzing whole blood transcriptomes of 992 individuals. We had genotype data for 3 of the 6 variants with significant sex interacting effects; none were significant at p<.05 for the tested SNP-gene pair. Our results suggest that it is unlikely that there are very strong differences by sex in the effect of genotypes on skeletal muscle expression. We plan to combine our data with human skeletal muscle RNA-Seq data from GTEx and also with adipose RNA-Seq data from our study to increase our power to detect sex-interacting SNPs.

REVERTANT MOSAICISM REPAIRS SKIN LESIONS IN A PATIENT WITH KERATITIS-ICHTHYOSIS-DEAFNESS (KID) SYNDROME BY SECOND-SITE MUTATIONS IN CONNEXIN 26

Sanna Gudmundsson[1], Maria Wilbe[1], Sara Ekvall[1], Adam Ameur[1], Nicola Cahill[1], Ludmil B Alexandrov[2], Marie Virtanen[3], Maritta Hellström Pigg[1], Anders Vahlquist[3], Hans Törmä[3], Marie-Louise Bondeson[1]

[1]Uppsala University, Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala, Sweden, [2]Los Alamos National Laboratory, Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, NM, [3]Uppsala University, Department of Medical Sciences, Dermatology, Uppsala, Sweden

Revertant mosaicism (RM) is a naturally occurring phenomenon where the pathogenic effect of a germline mutation is corrected by a second somatic event. Development of healthy-looking skin due to RM has been observed in patients with various inherited skin disorders, but not in connexin-related disease. We report on RM in a Keratitis-ichthyosis-deafness (KID) syndrome patient and further investigate the underlying molecular mechanisms. The patient was diagnosed with KID syndrome based on characteristic skin lesions, hearing deficiency and keratitis. Investigation of the gap junction beta 2 (*GJB2*) gene, encoding the gap junction channel (GJC) protein connexin 26 (Cx26), revealed heterozygosity for the recurrent *de novo* germline mutation, c.148G>A, p.Asp50Asn. At age 20, the patient developed spots of healthy-looking skin that grew in size and number on the inside of her thighs, within her widespread erythrokeratodermic lesions. To further investigate the mechanism, single molecule real-time (SMRT) ultra-deep sequencing was performed on cDNA and gDNA from two biopsies from healthy-looking spots, generating over 10’000 sequence reads over the *GJB2* locus in all sample. We identified five somatic nonsynonymous mutations in frequencies between 2.4-12.5%. Three mutations have previously been reported to cause hearing loss, and two are novel variants predicted to be disease causing. All five mutations were independently present *in cis* with the p.Asp50Asn mutation, but absent in biopsies from lesional skin. No second-site mutations could be detected on the wild-type (wt) Cx26 allele. Functional studies of Cx26 in HeLa cells displayed co-expression of Cx26-Asp50Asn and wt Cx26 in GJC plaque. However, Cx26-Asp50Asn with second-site mutations displayed no formation of GJC plaque or co-expression with wt Cx26. Conclusively, we show that the second-site mutations inhibit the dominant negative effect of Cx26-Asn50Asp in GJCs, which results in reversion of the disease-mechanism, explaining reverted skin phenotype in our patient. To our knowledge, this is the first time RM is reported in a KID syndrome patient.

ROBUST MAPPING OF WHOLE GENOME SEQUENCING DATA

Sebastian Deorowicz[1], Agnieszka Debudaj-Grabysz[1], Adam Gudyś[1], Szymon Grabowski[2]

[1]Silesian University of Technology, Institute of Informatics, Gliwice, Poland, [2]Lodz University of Technology, Computer Engineering Department, Lodz, Poland

The DNA sequence of the human genome was published for the first time in 2003. The *Human Genome Project* took 13 years of work of research centres from all over the world and costed nearly three billion US dollars. Modern technologies allows human genome to be sequenced in a few days for approximately 1000$, and these numbers are expected to further decrease. As a result, sequencing has become crucial in life sciences, particularly in medicine.

Mapping reads generated by sequencing platforms to reference genomes is one of the most important steps in the processing of sequencing data, affecting all downstream procedures, like variant calling or expression analysis. We present a new mapping algorithm. It uses suffix arrays, unlike the most popular existing tools like Bowtie or BWA-MEM which employ Ferragina-Manzini index. This leads to an excellent mapping speed. Importantly, suffix arrays are stored in the main memory in adjustable batches which keeps RAM requirements under control. The mapping of an example human dataset (860 million of paired-end reads of length 100; coverage 28) took 1½ hours on 12-core desktop computer and required 12GB of RAM. At the same time, Bowtie and BWA-MEM analyzed the data in approximately 7 hours in 6GB of memory. The number of successful mappings was similar for all investigated packages. Presented algorithm supports reads of different lengths from few tens to few hundreds of bases. It also allows mapping results to be stored in multiple per-chromosome SAM files for easier parallelization of downstream analyses which often process chromosomes independently. Another helpful feature is direct support to gzipped input and output which decreases disk requirements.

Competitive mapping quality, superior time efficiency, reasonable memory requirements, and the presence of additional features make our algorithm an interesting alternative to existing mapping packages.

POLYMERIZATION KINETICS DECIPHERED USING PACBIO SEQUENCING: NON-B DNA AFFECTS POLYMERASE PROGRESSION AND ERROR RATE

Wilfried M Guiblet[1,2], Marzia A Cremona[3], Monika Cechova[2], Robert S Harris[2], Iva Kejnovska[6], Kristin Eckert[4,5], Eduard Kejnovsky[7], Francesca Chiaromonte[3,5], Kateryna D Makova[2,5]

[1]Penn State University, Graduate Program in Bioinformatics and Genomics, The Huck Institutes for the Life Sciences, University Park, PA, [2]Penn State University, Department of Biology, University Park, PA, [3]Penn State University, Department of Statistics, University Park, PA, [4]Hershey College of Medicine, Pathology, Hershey, PA, [5]Penn State University, Center for Medical Genomics, The Hucks Institutes for the Life Sciences, University Park, PA, [6] Institute of Biophysics, Department of CD Spectroscopy of Nucleic Acids, Brno, Czech Republic, [7]Institute of Biophysics, Department of Plant Developmental Genetics, Brno, Czech Republic

Studies of individual loci demonstrated that non-B DNA (e.g., G-quadruplexes, Z-DNA, cruciforms, and slipped structures) causes polymerase stalling and replication errors, leading to genome instability. To date, these important effects of non-B DNA have not been investigated on a genome-wide scale. Here we explore DNA polymerization kinetics using human whole-genome data resequenced with Pacific Biosciences (PacBio) technology. In addition to base calling, this technology registers the time between incorporation of two consecutive bases, or InterPulse Duration (IPD). Using novel Functional Data Analysis techniques, we demonstrate that non-B DNA motifs lead to polymerization kinetics that is significantly deviant from that observed outside of such motifs. These alterations depend on motif identity and DNA strand involved. For instance, G-quadruplexes display a strong strand-specific deceleration in polymerization. We confirm this computational result with an experimental analysis of circular dichroism and show that polymerization kinetics at these motifs depends on their thermostability. Among other notable effects of non-B DNA structures are a significant polymerization acceleration at AT-rich homopolymer motifs and a periodic change in polymerization speed at microsatellite motifs. Importantly, several non-B motifs significantly increase the rate of PacBio sequencing errors. However, this effect explains only a small proportion of variation in polymerization speed, suggesting that most of the variation is biological. Furthermore, base composition and epigenetic modifications at non-B DNA motifs cannot explain the observed variation in polymerization dynamics. Finally, we extend our in vitro observations to in vivo effects via studying the relationship between non-B DNA motifs and human genetic diversity.

FITCONS2: INTERPRETABLE CHARACTERIZATION OF THE EPIGENOMIC PROFILES FROM HOMININ SELECTIVE PRESSURE.

Brad Gulko[1], Adam Siepel[2]

[1]Cornell University, Computer Science, Ithaca, NY, [2]CSHL, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Personalized genomics, disease consortium studies, and annotation of the noncoding human genome require computational tools for the prediction of causal alleles. Popular scoring systems such as EIGEN, CADD, DeepSEA and LINSIGHT combine selective constraint with functional annotations to rank noncoding genomic loci for biological relevance. Central to these methods is the idea that primary sequence conservation is complimentary to functional assays for the assessment of regulatory activity. While effective, existing methods produce results that can be linearly constrained, under powered, training-set biased, or difficult to interpret. This limits their use to biologists interested in understanding both organismal and tissue-specific genomic properties.

To address these limitations we introduce Fitcons2, a new method for tissue-specific, whole genome segmentation based on the integration of hominid selective pressure with functional genomic assays. Fitcons2 offers base-pair resolution and models strongly non-linear relationships, while maintaining clear interpretability from both evolutionary and functional genomic perspectives.

Fitcons2 identifies a small and intelligible collection functional assay patterns (~61) that are optimally informative about sites under selective pressure, both genome-wide and jointly across tissue types. Nine functional properties are obtained for 115 karyotype normal tissue types from the Roadmap Epigenomics project and finely quantized, providing degrees of relevance. Considered properties are derived from biochemical assays and include: splice-site proximity, transcription factor binding, DNase I hypersensitivity, long and short RNA transcription, histone modification, codon position and DNA methylation. Selective pressure is assessed under the INSIGHT model, which combines human polymorphism and hominin divergence to calculate probability of selective constraint in a manner sensitive to both the adaptive and weak selective pressures present at noncoding regulatory loci.

Tissue-specific analyses are integrated to produce an organismal measure of selective pressure at each hg19 genomic position; this also generates a simple mapping from positional score to relevant tissue and tissue-specific epigenomics. FitCons2 provides discrimination of pathogenic sites and QTL approaching that of black box methods while providing superior interpretability and biological insight. Fitcons2 quantifies selective pressure at each position, identifies tissue-specific epigenomic properties that drive biological processes, and characterizes patterns of epigenomic diversity in both oncogenic and normal tissue types.

A FRAMEWORK TO INTERPRET SHORT TANDEM REPEAT VARIATIONS IN HUMANS

Melissa Gymrek[1,2,3,4], Thomas Willems[2,5], David Reich[6,7], Yaniv Erlich[2,8]

[1]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [2]New York Genome Center,, New York, NY, [3]University of California San Diego, Department of Medicine, La Jolla, CA, [4]University of California San Diego, Department of Computer Science and Engineering, La Jolla, CA, [5]Massachusetts Institute of Technology, Computational and Systems Biology Program, Cambridge, MA, [6]Harvard Medical School, Department of Genetics, Boston, MA, [7]Harvard Medical School, Howard Hughes Medical Institute, Boston, MA, [8]Columbia University, Department of Computer Science, Fu Foundation School of Engineering, New York, NY

Identifying regions of the genome that are depleted of mutations can reveal potentially deleterious variants. Short tandem repeats (STRs), comprised of repeating motifs of 1-6bp, are among the largest contributors of de novo mutations in humans and are implicated in a variety of human disorders. However, because of the challenges STRs pose to bioinformatics tools, studies of STR mutations have been limited to highly ascertained panels of several dozen loci. Here, we harnessed novel bioinformatics tools and an analytical framework to estimate mutation parameters at each STR in the human genome. We then developed a model of the STR mutation process that allows us to obtain accurate estimates of mutation parameters at each STR by correlating genotypes with local sequence heterozygosity. Finally, we applied our method to obtain robust estimates of the impact of local sequence features on mutation parameters and used this to create a framework for measuring constraint at STRs by comparing observed vs. expected mutation rates. Constraint scores identified known pathogenic variants with early onset effects. Our constraint metrics will provide a valuable tool for prioritizing pathogenic STRs in medical genetics studies.

LEVERAGING MASSIVE PARALLEL REPORTER ASSAYS FOR FUNCTIONAL REGULATORY ELEMENTS PREDICTION

Anat Kreimer[1,2], Zhongxia Yan[1], Nadav Ahituv[2], Nir Yosef[1]

[1]UC Berkeley, Electrical Engineering & Computer Science, Berkeley, CA, [2]UCSF, Bioengineering and Therapeutic Sciences, San Francisco, CA

Deciphering the functionality of the non-coding genome has been the focus of many recent studies, aiming to annotate regulatory regions and understand their specific role in disease and other phenotypes. The massively parallel reporter assay (MPRA) is a nascent technology designed to address this question in a comprehensive manner, by enabling to test thousands of sequences for their regulatory activity in a single, quantitative experiment.

Here, we examine several MPRA datasets performed by different labs using various methodologies and cell types. When applying an ensemble of machine learning methods for prediction of MPRA output, we observe an overall consistency of feature contribution, with transcription factor binding and epigenetic properties being the top predictors. We also notice improvement in MPRA prediction when it is carried out in a chromosomal rather than episomal context.

We then examine how generalizable our method is across cell types and observe a robust, although slightly reduced prediction power, when training the model on one cell type and predicting MPRA output in a different cell type.

Finally, we establish that models trained with MPRA data can distinguish developmental enhancers from genomic background in an unsupervised manner.

Our approach, which leverages experimentally measured as well as predicted chromatin properties, combined with MPRA data, can be used to highlight functional regulatory regions in any cell type or tissue throughout the genome.

SOCIAL STATUS EFFECTS ON GENE EXPRESSION ARE DECOUPLED FROM CHROMATIN ACCESSIBILITY AFTER AN ACUTE STRESS

Noah Snyder-Mackler[1], Joaquin Sanz[2], Jordan Kohn[3], Roger Pique-Regi[4], Mark E Wilson[3], Luis B Barreiro[2], Jenny Tung[1]

[1]Duke University, Evolutionary Anthropology, Durham, NC, [2]University of Montreal, CHU Sainte-Justine, Montreal, Canada, [3]Emory University, Yerkes National Primate Research Center, Atlanta, GA, [4]Wayne State University, Molecular Medicine and Genetics, Detroit, MI

Social status can strongly influence fitness in social mammals, and in humans it is one of the best predictors of disease susceptibility and mortality. These effects are thought to stem in part from dysregulation of the physiological stress response, resulting in impaired sensitivity to acute stressors and compromised immune function. Here, we used an experimental manipulation of social status in female rhesus macaques to: (i) investigate how social status influences the gene expression response to an acute stress challenge; and (ii) assess the contribution of the epigenome to this response.

To do so, we took advantage of 9 experimentally constructed social groups (n=43 females) in which earlier introduction to the group predicts higher dominance rank. We challenged peripheral blood mononuclear cells from each of these females with the synthetic glucocorticoid Dexamethasone (n=86 matched control and Dex-treated samples). Our results reveal that social status effects on PBMC gene expression are present in control cells, as expected from prior work, but are significantly attenuated after Dex treatment (Wilcoxon test: $p=1.2x10^{-116}$). Thus, in contrast to immune stimulation, which we previously showed exaggerates the effects of social status, Dex treatment (an immunosuppressant) dampens it.

To investigate the role of the chromatin accessibility epigenome in explaining this pattern, we profiled the same samples using ATAC-seq. Social status altered chromatin accessibility for 2.4% of accessible regions (n=453, FDR=10%) and the magnitude and direction of these effects significantly predicted social status effects on the nearest gene's expression ($r^2=0.032$, $p=7.8x10^{-142}$). However, unlike for gene expression, social status effects on chromatin accessibility were of similar magnitude in control and Dex-treated cells (Wilcoxon test: p=0.84). Further, status effects on chromatin accessibility were substantially less predictive of status effects on gene expression ($r^2=0.0019$, $p=7.5x10^{-8}$). Our results suggest that Dex treatment decouples the relationship between social status and chromatin accessibility from that of social status and gene expression, potentially due to changes in transcription factor binding after Dex exposure. Together, our findings indicate that social status interacts with other environmental conditions to shape gene regulation, and implicate epigenetic mechanisms as a partial basis for this observation.

CHROMATIN PROFILING REVEALS STRONG DIFFERENCES IN TRANSCRIPTION FACTOR ACTIVITY ACROSS AGEING IN MESENCHYMAL STEM CELLS

Mariana Ruiz-Velasco[1], Ximing Ding[1], Patrick Horn[2], Anthony Ho[2], Anne-Claude Gavin[1], Judith Zaugg[1]

[1]European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany, [2]Heidelberg University Hospital, Haematology, Oncology, Rheumatology, Heidelberg, Germany

Ageing in higher organisms is a complex phenomenon where diverse molecular and physiological changes result in the overall deterioration of various biological traits and an increasing susceptibility to common diseases. Many of these age-related common diseases are associated with a mis-functioning of the immune system (e.g. cancer, neurodegenerative or autoimmune diseases). Specifically, the haematopoietic stem cell (HSC) niche plays an important role in this increased disease susceptibility with age since it is a main cause of the age-dependent decline of immune system function.

In the presented study we have characterised the chromatin landscape of mesenchymal stem cells (MSCs), which are an important part of the HSC niche, derived from bone marrow in a cohort of 15 healthy individuals of different age (ranging from 20 to 60 years old). Importantly, despite the small size of the cohort, we find that age explains around 50% of the variation in chromatin accessibility among individuals. Furthermore we identified hundreds of genomic regions that behave in an age-dependent manner, many of which fall in quiescent regions, indicating that supposedly repressed chromatin domains are becoming more active again in older individuals. Strikingly, employing our novel approach for deriving transcription factor (TF) activity from chromatin measurements, we have identified many factors that are known to be involved in age-related processes, such as decreased bone mineral density and susceptibility to cancer, that we find more active in old individuals. In contrast, TFs associated with increased lifespan in *C. elegans*, and decreased mortality in mice, are more active in young individuals. Here we present a comprehensive gene regulatory network of mesenchymal stem cells and describe the TFs and their associated pathways that significantly change in activity during healthy ageing.

THE GENOME ARCHITECTURE OF BDELLOID ROTIFERS: SHAPED BY THEIR LONG-TERM AMEIOTIC EVOLUTION OR DESICCATION?

Karine Van Doninck, N. Debortoli, B. Hespeels, J.-F. Flot

Université de Namur, Department of Biology, Namur, Belgium

Loss of sex is an evolutionary dead end for metazoans, but bdelloid rotifers, micro-invertebrates abundantly found in semi-terrestrial habitats such as lichens and mosses, challenge this view having persisted asexually for millions of years. We found that the genome structure of the bdelloid lineage *Adineta vaga* is indeed incompatible with conventional meiosis. At gene scale, the genome is tetraploid and comprises anciently duplicated segments and less divergent allelic regions. However, in contrast to sexuals, the allelic regions are rearranged and sometimes found on the same chromosome. Such genomic architecture impedes meiotic pairing, confirming their ameiotic evolution. Instead, we found abundant evidence of gene conversion, limiting the accumulation of mutations in the absence of meiosis. Gene conversion may occur during mitotic recombination repair of broken DNA following cycles of desiccation and rehydration experienced by bdelloids in their temporary habitats. Indeed during desiccation the genome of *A. vaga* is broken in hundreds of DNA fragments that get repaired once rehydrated. Recently we obtained new genomic data studying the evolution of the genome of *A. vaga* following cycles of desiccation, investigating whether structural variations are apparent.

In the genome of *A. vaga* 8% of the genes are likely of non-metazoan origin and probably acquired horizontally. These genes appear to be functional and many of those involved in resistance to desiccation have been acquired through HGT. Moreover, combining nuclear and mitochondrial markers, we demonstrated recently intra- and inter-specific genetic exchanges within the lineage *A. vaga* suggesting a non-meiotic recombination mechanism of DNA exchange.

THE GENETIC BASIS OF PARENTAL CARE EVOLUTION IN *PEROMYSCUS* MICE

Andres Bendesky[1,2,3], Young-Mi Kwon[1,2], Jean-Marc Lassance[1,2,3], Caitlin L Lewarch[1,3], Shenqin Yao[1,2], Brant K Peterson[1,2], Meng X He[4], Catherine Dulac[1,3], Hopi E Hoekstra[1,2,3]

[1]Howard Hughes Medical Institute, -, Cambridge, MA, [2]Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA, [3]Harvard University, Department of Molecular and Cellular Biology, Cambridge, MA, [4]Harvard University, Graduate Program in Biophysics, Cambridge, MA

Parental care is essential for the survival of mammals, yet the mechanisms underlying its evolution remain largely unknown. Here we show that two closely related species of *Peromyscus* mice, *P. polionotus* and *P. maniculatus*, differ greatly in parental behavior and that these differences are heritable. Using a quantitative genetic approach, we identify 12 genomic regions that affect parental care, eight of which have sex-specific effects, suggesting that parental care can evolve through independent genetic mechanisms in males and females. Furthermore, some regions affect parental care broadly, whereas others affect specific aspects of parental behavior, such as nest building. Transcriptome analysis shows that, of the genes that reside in genomic regions linked to differences in nest-building behavior, vasopressin is strongly differentially expressed in the hypothalamus of the two species, with increased levels associated with less nest building. Using pharmacology in *Peromyscus* and chemogenetics in *Mus musculus*, we show that vasopressin inhibits nest building but not other parental behaviors. Together, our results define the genetic properties of parental behavior evolution in *Peromyscus* and indicate that variation in an evolutionarily ancient neuropeptide contributes to interspecific differences in parental care.

THE GENE EXPRESSION CONSEQUENCES OF MAMMALIAN REGULATORY EVOLUTION

Camille Berthelot[1,2], Diego Villar[3], Duncan T Odom[3], Paul Flicek[1]

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, [2]Institut de Biologie de l'Ecole Normale Superieure, Dynamics and Organisation of Genomes (DYOGEN), Paris, France, [3]Cambridge Institute - Cancer Research UK, University of Cambridge, Cambridge, United Kingdom

Mammalian gene expression is controlled by collections of promoter and enhancer regions. Numerous studies have documented the rapid evolution of mammalian regulatory elements, especially enhancers, and yet gene expression patterns are often highly stable between species. How stable gene expression is maintained by rapidly evolving regulatory landscapes is a fundamental question in evolutionary genetics.

To date, comparative approaches to gene regulation have largely focused on lineage-specific regulatory innovations. How evolutionarily plastic or stable regulatory elements influence gene expression remains however poorly understood.

Here, we rigorously tested the contributions of both landscape complexity (number of regulatory elements) and conservation of regulatory activity on gene expression evolution, using an integrated dataset of active promoters, enhancers and gene expression output from the same liver samples across twenty mammalian species. Our methodology captures regulatory activities that range from essential to dispensable, and from highly-conserved across mammals to present in only one species. This analysis revealed that gene expression levels and stability are reflected by the complexity of their regulatory landscape, both within a single species and across mammals. Regulatory activities conserved across placental mammals exert a powerful stabilizing effect, associating with gene expression levels that are simultaneously high and evolutionarily stable. These discoveries extend previous reports connecting evolutionary constraint on promoter and enhancer activities with conserved expression outputs, and are consistent with the proposed functional relevance of evolutionarily constrained regulatory elements. In contrast, recently-evolved enhancers contribute weakly to gene expression and transcriptional stability, consistent with a model whereby a sizable fraction of new-born enhancer elements have a neutral role on gene expression evolution. However, we further have identified a set of genes that recurrently accumulate lineage-specific enhancers across species and that display increased expression divergence, indicating that genes with flexible expression levels better tolerate regulatory plasticity.

Our results underscore how the evolutionary stability of gene expression is profoundly entwined with both the number and conservation of surrounding promoters and enhancers.

THE INTEGRATIVE HUMAN MICROBIOME PROJECT (IHMP) PROVIDES EXTENSIVE DATA RESOURCES AND TOOLS TO BETTER UNDERSTAND THE INTERACTIONS OF HOST AND MICROBIOME.

Michael Snyder, George Weinstock, Greg Buck, Curtis Huttenhower, and HMP Consortium

Recent microbiome research has brought extensive new insight into the distribution and diversity of human microbial communities, as well as their associations with human diseases. Still, relatively little is known about the biology underlying the interaction of the components of the microbiome and the human host. The Integrative Human Microbiome Project (iHMP, http://hmp2.org), the second phase of the NIH Human Microbiome Project, is studying these interactions by creating integrated data sets exploring the functional properties of the microbiome and its host, and by analyzing activities of the the microbiome and the host using longitudinal studies of disease-specific cohorts. The three projects of the iHMP focus on: 1) the impact of the microbiome on pregnancy and preterm birth in a cohort of ~1500 pregnant women sampled longitudinally throughout pregnancy for microbiome composition, metagenome and metatranscriptome profiles, immunoproteome expression, and the lipidome profile; 2) Inflammatory bowel diseases, in which 100 Crohn’s disease patients, ulcerative colitis patients, and controls are profiled longitudinally for one year each; 3) prediabetics, in which 107 subjects are profiled with more than 900 longitudinal timepoints total that span over three years. For each cohort, we are investigating the complex of interaction between the human host and the microbiome by comprehensive multi-omics profiling, creating extensive sets of microbiome, genome, transcriptome, proteome, host and microbial metabolome genomes, transcriptome, and proteome data to elucidate the host-microbiome interactions and the mechanisms by which disease is caused at the clinical, immunological, molecular, genetic, and microbial levels. We have also organized the iHMP Data Coordination Center, to provide the community with a unified data repository and resources permitting query, retrieval and analysis of data generated. Together, the iHMP serves to provide extensive data resources and tools to evaluate new models, methods, and analyses to better understand the interactions of host and microbiome.

GENOMICS AND THE ORIGINS OF SPECIES.

Nitin Phadnis

University of Utah, Biology, Salt Lake City, UT

Speciation, the process by which one species splits into two, involves the evolution of reproductive isolating barriers such as the sterility or inviability of hybrids between previously interbreeding populations. Even in his masterpiece “On The Origin of Species”, Darwin could find no satisfactory solution to the apparent paradox of why natural selection would tolerate the onset of genetic barriers such as hybrid sterility and inviability that diminish the prospect of successful reproduction and, therefore, termed this problem the "mystery of mysteries".

The key to uncovering the molecular and evolutionary basis of speciation involves the identification of genes that cause hybrid sterility and inviability and understanding the molecular mechanisms of hybrid dysfunction. The identification of the genes that drive speciation, however, represents an indispensible and rate limiting step even in today's post-genomic era. Very few such genes have been identified, and we know even less about the evolutionary forces and the molecular developmental pathways that are disrupted in hybrids.

Here, I describe the key developments with new genomics and cell biological approaches that are rapidly changing our understanding of the molecular basis of speciation. Our studies of the cellular and developmental anomalies in inter-species hybrids also provide surprising insights into the otherwise hidden evolutionary conflicts that ultimtely shape the architecture of our genomes, cells and species.

GENE ENCRYPTION IN THE MITOCHONDRIAL GENOME OF DIPLONEMIDS.

Sandrine Moreira[1,2], Matus Valach[1], Gertraud Burger[1]

[1]University of Montreal, Biochemistry and Molecular Medicine, Montreal, Canada, [2]University of Columbia, Biochemistry and Molecular Biophysics, New York, NJ

Thanks to new high throughput sequencing technologies and automatic annotation pipelines, proceeding from raw sequence reads to a GenBank file can be achieved in a single mouse click or so, for some species. Others, however, fiercely resist bioinformaticians with their confounding genomic complexity. Diplonemids are one of them.

Diplonemids are a group of poorly studied marine protists. Unexpectedly, metagenomic studies have recently ranked this group as one of the most diverse in the oceans. Yet, their most distinctive feature is a multipartite mitochondrial genome with genes in pieces. These unique features were discovered by our laboratory in the type species *Diplonema papillatum*. Specifically, the 80 mitochondrial chromosomes of *Diplonema* are composed of 95% non-coding repeats, whose distinctive arrangement allows classifying chromosomes into two classes. The remaining 5% of unique sequence includes one (or exceptionally two) gene fragments. After transcription, gene pieces are stitched together by RNA trans-splicing. Genes are composed by 1 to 11 fragments which are of small size, ranging from 50 to 350 bp.

We recently discovered that decryption of genes not only requires RNA trans-splicing, but also RNA editing of two types, (i) polyuridylation at the junction of gene pieces and (ii) substitutions of A-to-I and C-to-T. We reconstructed *in silico* the mitochondrial transcriptome from RNA-seq reads. Thereby, we have identified six new genes including one that has alternative trans-splicing isoforms. In total, RNA editing adds 237 uridines (Us) in 14 transcripts with up to 50 Us in a row, and substitutes 114 nucleotides by deamination (A-to-I or C-to-T) in seven transcripts. The latter sites are tightly clustered in a single region of the transcript.

To get insight into the evolution of these extraordinary features, we reconstructed the mitochondrial genome and transcriptome of three other diplonemids. We discovered a high plasticity of their mitochondrial chromosome architecture with variable numbers of chromosomes, chromosome classes, and gene fragment distribution. Most surprising are the numerous cases of two to three overlapping gene fragments with the extreme situation of one being completely embedded in another. The reading frames of ‘Russian doll’ gene pieces is the same, thus constituting a case of re-used DNA sequence. In contrast to this plasticity, the fragment number per gene and the position of fragment junctions are highly conserved. Also gene pieces edited by RNA substitution editing are shared in all species, although the edited nucleotide positions are variable. Among U-appendage sites, about half is species-specific.

Our study revealed the exceptional genomic architecture and the unique arsenal of post-transcriptional processes of diplonemids; a glimpse into the untapped genomic diversity of microbes.

EVOLUTION OF TISSUE-SPECIFIC REGULATORY PROGRAMS IN CICHLIDS

Tarang K Mehta[1], Sara A Knaack[2], Christopher Koch[3], Padhmanand Sudhakar[1], Luca Penso-Dolfin[1], Tomasz Wrzesinski[1], Will Nash[1], Tamas Korcsmaros[1], Wilfried Haerty[1], Sushmita Roy[2,3,4], Federica Di-Palma[1]

[1]Earlham Institute (EI), Regulatory and Systems Genomics, Norwich, United Kingdom, [2]Wisconsin Institute for Discovery (WID), Systems Biology, Madison, WI, [3]UW Madison, Dept. of Biostatistics and Medical Informatics, Madison, WI, [4]UW Madison, Dept. of Computer Sciences, Madison, WI

In vertebrates, the East African cichlid radiations represent arguably the most dramatic examples of adaptive speciation. In the great lakes Victoria, Malawi and Tanganyika and within the last few million years, one or a few ancestral lineages of haplochromine cichlid fish have given rise to over 1500 species exhibiting an unprecedented diversity of morphological and ecological adaptations. Such explosive phenotypic diversification of East African cichlids is unparalleled among vertebrates and the low protein divergence between species implies the rapid evolution of regulatory regions and networks underlying the traits under selection.

Comparative functional genomics, transcriptomics and epigenomics are powerful tools to study the evolution of tissue and species divergence. We recently developed *Arboretum*, an algorithm to identify modules of co-expressed genes across multiple species in a phylogeny. By integrating inferred modules with nucleotide variation, predicted cis regulatory elements and miRNA profiles from five East African Cichlids, we investigated the evolution of tissue-specific gene regulation. Our analyses identified modules with tissue-specific patterns for which we reconstructed the evolutionary gene regulatory networks across the five cichlids species. We report striking cases of rapid network rewiring for genes known to be involved in traits under natural and/or sexual selection such as the visual systems, and more specifically a cone opsin (*sws2a*) responsible for colour vison of selected cichlid fishes. Furthermore, in-depth analyses of regulons (transcription factor – target interactions, e.g. Egr3) show similar rapid species-specific rewiring. Investigation of these novel interactions in *Astatotilapia burtoni*, a model species for behaviour evolution, reveal significant enrichment for genes involved in neuronal and brain function. Our unique integrative approach that interrogates the evolution of regulatory networks allowed us to identify the rapid regulatory changes associated with certain traits under selection in cichlids.

REPEATED LOSSES OF PRDM9-DIRECTED RECOMBINATION DESPITE THE CONSERVATION OF PRDM9 ACROSS VERTEBRATES

Zachary Baker[1], Molly Schumer[2,3,4], Yuki Haba[5], Chris Holland[4,6], Gil G Rosenthal[4,6], Molly Przeworski[1,2]

[1]Columbia University, Systems Biology, New York, NY, [2]Columbia University, Biological Sciences, New York, NY, [3]Harvard University, Harvard Society of Fellows, Boston, MA, [4]Centro de Investigaciones Cientificas de las Huastecas "Aguazarca", Hidalgo, Mexico, [5]Columbia University, Evolution, Ecology and Environmental Biology, New York, NY, [6]Texas A&M University, Biology, College Station, TX

Studies of a handful of species reveal two mechanisms by which meiotic recombination is directed to the genome—through PRDM9 binding or by targeting promoter-like features—that lead to dramatically different evolutionary dynamics of hotspots. Here, we identified PRDM9 from genome and transcriptome data in 225 species, finding the complete PRDM9 ortholog across distantly related vertebrates. Yet, despite its broad conservation, we inferred a minimum of six partial and three complete losses. Strikingly, taxa carrying the complete ortholog of PRDM9 are precisely those with rapid evolution of its predicted binding affinity, suggesting that all its domains are necessary for directing recombination. Indeed, as we show, swordtail fish carrying a partial ortholog share recombination properties with PRDM9 knock-outs.

CHROMATIN ACCESSIBILITY PROFILES IN GENETICALLY IDENTICAL TWINS DIVERGENT FOR DISEASE REVEALS ASTHMA-ASSOCIATED DNA ELEMENTS

Ansuman T Satpathy*[1], Rebecca N Bauer*[2], Kun Qu[1], Theo Ho[2], Rachel C Miller[3], Howard Y Chang*[1], Kari C Nadeau*[2]

[1]Stanford University School of Medicine, Center for Personal Dynamic Regulomes, Stanford, CA, [2]Stanford University, Sean N. Parker Center for Asthma and Allergy Research, Department of Medicine, Stanford, CA, [3]Columbia University College of Physicians & Surgeons, Department of Medicine, New York, NY

*authors contributed equally

The interplay between genes and environment underlie the development of many common human diseases. Genome-wide association studies (GWAS) often require large numbers of participants to address inherent heterogeneity within test and control populations, and cannot directly assess the impact of environment on genes. Here we introduce Regulome-wide Association Study in twin Pairs (RASP) to identify disease-specific DNA elements from small number of well-characterized pairs. Asthma is a chronic inflammatory airway disease typified by aberrant accumulation and function of memory CD4+ T cells that produce type 2 cytokines (Th2 cells). Asthma stems from a combination of genetic and environmental factors, suggesting that altered immune function may result from dysregulated epigenetics rather than underlying differences in DNA. Using primary human T cells from twin pairs discordant for asthma to control for underlying genetics, we assessed asthma-related differences in open chromatin sites across the genome of CD4+ T cell subsets by Assay of Transposase Accessible Chromatin with sequencing (ATAC-seq). PBMC were isolated from the blood of monozygotic twins pairs discordant for asthma (n=12 pairs). CD4+ Naive (CD45RA+), Th1 (CD45RA-CXCR3+CCR4-), and Th2 (CD45RA-CXCR3-CCR4+) cells were sorted by fluorescence assistance cell sorting. Comparison of ATAC-seq profile of each asthmatic patient to his or her healthy twin revealed reproducible disease-associated chromatin signatures that spanned multiple twin pairs. The asthma-associated chromatin signature is selectively present in Th2 but not naïve or Th1 CD4+ T cells, identifies new regulatory elements and genes beyond GWAS, and is a more accurate predictor of asthma diagnosis than a clinically used airway test. These results suggest the potential utility of RASP approach to unveil the intersection of genetics and environment in human disease.

POOLED CRISPR ACTIVATOR SCREENS FOR CELLULAR REPROGRAMMING COCKTAILS BASED ON GLOBAL MODELS OF CHROMATIN REGULATION ACROSS 98 CELL TYPES

Brian Cleary[1,2], Jian Shu[1], Charles Fulco[1], Vidya Subramanian[1], Aviv Regev[1,3], Eric Lander[1,3]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]Massachusetts Institute of Technology, Computational and Systems Biology Program, Cambridge, MA, [3]Massachusetts Institute of Technology, Biology, Cambridge, MA

Using data from the Roadmap Epigenomics project we have modeled regulation of the chromatin landscape by searching for sparse modules of genomic regions that appear to have co-regulated histone modifications. For each module of genomic regions (i.e. a list of 500bp windows in the genome) we propose a set of transcription factors that have enriched motifs in these regions, and have expression levels that co-vary across cell types with the chromatin dynamics of the module. Thus, we have a generative model that takes transcription factor abundances as input, calculates the activity level of several hundred chromatin modules (with only a small proportion of modules active in any given cell type), and then, based on the individual genomic regions within each active module, outputs predicted signal intensity, genome-wide for 5 histone modifications. Using this generative model, we then consider the problem of cellular reprogramming and ask, for a given source and target cell type, which chromatin modules need to be reprogrammed, and which transcription factors can most effectively achieve this reprogramming? These predictions give us insight into why certain routes of reprogramming might be easy or difficult to achieve, what regions of the genome might be targeted by different reprogramming cocktails, and why some cocktails might be more effective than others. To test these predictions experimentally, we have developed a method for pooled screening of cellular reprogramming cocktails using CRISPR activator-based over-expression of combinations of transcription factors.

CHARACTERIZATION OF REGULATORY VARIATION IMPACTING CARDIAC TRAITS IN IPSC-DERIVED CARDIOMYOCYTES

Paola Benaglio[1], Christopher DeBoever[2,3], Matteo D'Antonio[1], He Li[4], Frauke Drees[4], Hiroko Matsui[4], Joaquin Reyna[4], Agnieszka D'Antonio-Chronowska[1], Erin N Smith[1], Kelly A Frazer[1,4]

[1]University of California, San Diego, Department of Pediatrics and Rady Children's Hospital, La Jolla, CA, [2]University of California, San Diego, Bioinformatics and Systems Biology, La Jolla, CA, [3]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [4]University of California, San Diego, Institute for Genomic Medicine, La Jolla, CA

Non-coding genetic variation in regulatory regions of the human genome is an important contributor to gene expression differences between individuals, as well as traits and disease. However, elucidating the functions of regulatory variants is challenging because they often impact cell-type specific phenotypes. Cellular systems such as induced pluripotent stem cells (iPSCs) that can be differentiated potentially into any human tissue and can be profiled in depth are a promising strategy to characterize the function of regulatory variants in cell type-specific contexts. For this purpose, we have generated a large collection of human iPSCs from 222 different individuals genotyped by whole-genome sequencing (the iPSCORE resource) and are differentiating each line into target cell types, including cardiomyocytes (iPSC-CMs). We are profiling each line by RNA-Seq, ATAC-Seq, ChIP-Seq and DNA methylation to identify common and rare genetic variants associated with these molecular phenotypes.

Here, we analyzed the iPSCs and iPSC-CMs from a pedigree of seven individuals in the iPSCORE resource by using a variety of functional genomic assays including RNA-Seq, ATAC-Seq, GRO-Seq, ChIP-Seq of histone modification H3K27ac and of key cardiac transcription factors (TFs) NKX2-5 and SRF. First, we showed that iPSCs and iPSC-CMs recapitulate known expression and epigenetic signatures of human stem cells and cardiomyocytes, respectively. Then, we identified inter-individual differences in TF binding, histone modification, and gene expression, and showed that they were correlated with each other and associated with genetic variation, suggesting that DNA variants underlie molecular differences between iPSC lines derived from different individuals. Finally, we characterized variants showing allele-specific effects (ASEs) by examining alteration of DNA motifs as well as overlap with eQTLs and GWAS SNPs. In particular, we found that ASE variants affecting NKX2-5 binding showed enrichment for altered TF motifs, heart-specific eQTLs, and variants associated with electrocardiographic traits by GWAS, suggesting that differential NKX2-5 binding may underlie some of these associations. Our findings demonstrate the utility of iPSC-CMs for studying the impact of human regulatory variation on cardiac molecular phenotypes and highlight the potential of iPSCs and derived cell types as an effective new model system to elucidate the genetic basis of human traits and diseases.

A COMPARATIVE TIMECOURSE STUDY OF ENDODERM DIFFERENTIATION IN HUMANS AND CHIMPANZEES

Lauren E Blake[1], Samantha M Thomas[1], John D Blischak[1], Chiaowen Joyce Hsiao[1], Claudia Chavarria[1], Marsha Myrthil[1], Yoav Gilad[1,2], Bryan J Pavlovic[1]

[1]University of Chicago, Department of Human Genetics, Chicago, IL,
[2]University of Chicago, Department of Medicine, Chicago, IL

There is substantial interest in the genetic regulatory framework that is established in early human development, and in the evolutionary forces that shaped early developmental processes in humans. Progress in these areas has been slow because it is difficult to obtain relevant biological samples. Recent technological developments in the generation and differentiation of inducible pluripotent stem cells (iPSCs) provide the ability to develop in vitro models of early human and non-human primates developmental stages. We have previously established matched iPSC panels from humans and chimpanzees. Using these panels, we comparatively characterized gene regulatory changes through a four-day timecourse differentiation of iPSCs (day 1) into primary streak (day 2), endoderm progenitors (day 3), and definitive endoderm (day 4).

As might be expected, we found that differentiation stage (in effect, cell type) is the major driver of variation in gene expression levels in our study, followed by species. We identified thousands of differentially expressed genes between humans and chimpanzees in each differentiation stage. Yet, when we considered gene-specific dynamic regulatory trajectories throughout the timecourse, we found that 68% of genes, including nearly all known endoderm developmental markers, have conserved trajectories in the two species. Interestingly, we observed a marked reduction of both intra- and inter-species variation in gene expression levels in primitive streak samples compared to the iPSCs, with a recovery of variation in endoderm progenitors. The reduction in variation in gene expression levels at a specific developmental stage, paired with the high degree of conservation of temporal expression across species, is consistent with the dynamics of developmental canalization. Overall, we conclude that endoderm development in iPSC-based models are highly conserved and canalized between humans and our closest evolutionary relative.

ORGANOGENOMICS – RECONSTRUCTING HUMAN ORGAN DEVELOPMENT USING SINGLE-CELL TRANSCRIPTOMICS.

J. Gray Camp[1], Keisuke Sekine[2], Tobias Gerber[1], Malgorzata Gac[1], Sabina Kanton[1], Jorge Kageyama[1], Takanori Takebe[2,3], Barbara Treutlein[1,4,5]

[1]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genomics, Leipzig, Germany, [2]Yokohama City University, School of Medicine, Yokohama, Japan, [3]University of Cincinnati, UC Department of Pediatrics, Cincinnati, OH, [4]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany, [5]Technical University Munich, School of Life Sciences, Munich, Germany

Recent advances in the field of stem cell biology have made it possible to grow in vitro three-dimensional (3-D) human tissues that model human developmental processes. We combine human stem cell based organoid systems with single-cell transcriptomics analysis to reconstruct human organ development and understand mechanisms underlying cell fate programming. I will exemplify our approach by presenting our work on a 3-D human liver organoid system that is generated by reconstituting hepatic, stromal, and endothelial cell interactions occurring during liver bud development. We use single-cell RNA-seq (scRNA-seq) to compare hepatocyte-like lineage progression from pluripotency in 2-D culture and 3-D liver bud organoids and find that organoid hepatoblasts diverge from the 2-D lineage, and express epithelial migration signatures characteristic of organ budding. We benchmark 3-D liver bud organoids against fetal and adult human liver scRNA-seq data, and find a striking correspondence between the liver bud organoid and fetal liver cells. We use network analysis to predict autocrine and paracrine signaling in LBs, and predict inter-lineage communication involved in LB vascularization and self-organization. In summary, our molecular dissection reveals interlineage communication regulating organoid development, and illuminates previously inaccessible aspects of human liver development.

ULTRACONSERVED ENHANCERS ARE REQUIRED FOR NORMAL DEVELOPMENT

Diane E Dickel[1], Athena R Ypsilanti[2], Ramón Pla[2], Yiwen Zhu[1], Brandon J Mannion[1], Yoko Fukuda-Yuzawa[1], Ingrid Plajzer-Frick[1], Catherine S Pickle[1], Elizabeth Lee[1], Anne Harrington[1], Quan Pham[1], Tyler H Garvin[1], Momoe Kato[1], Jennifer A Akiyama[1], Veena Afzal[1], John L R Rubenstein[2], Axel Visel[1,3,4], Len A Pennacchio[1,3]

[1]Lawrence Berkeley National Laboratory, Functional Genomics Department, Berkeley, CA, [2]University of California San Francisco, Department of Psychiatry, Nina Ireland Laboratory of Developmental Neurobiology, San Francisco, CA, [3]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA, [4]University of California Merced, School of Natural Sciences, Merced, CA

More than 450 regions in the human genome have perfect sequence conservation between human and rodents, and these “ultraconserved” sites have intrigued biologists in the decade since they were first described. While it is known that many of these sequences are distant-acting enhancers, the drivers of such extraordinary evolutionary constraint remain unclear. Surprisingly, initial deletion studies showed that the loss of individual ultraconserved enhancers in mice has no obvious impact on viability or fertility. To explore this apparent discrepancy between extreme evolutionary constraint and lack of obvious phenotypes in more depth, we examined the *in vivo* consequences of loss of a series of ultraconserved enhancers near the essential neuronal transcription factor *Arx*. The *Arx* locus has an unusually high density of ultraconserved sites, including the three longest perfectly human-mouse-rat conserved sequences in the genome. Four ultraconserved regions near *Arx* drive gene expression in the developing forebrain, in patterns that cumulatively recapitulate the gene’s expression domains. Single-cell transcriptome sequencing from transgenic embryos confirmed that the enhancer activity is specific to *Arx*-expressing neuronal subpopulations. We engineered mice missing these four enhancers singly, as well as pairwise for enhancers that display overlapping *in vivo* activity patterns. While the loss of any single or pair of ultraconserved enhancers resulted in viable and fertile mice, detailed phenotyping revealed neurological or growth abnormalities in all cases, including substantial deficits of cholinergic neurons, altered densities of cortical interneuron populations, and abnormalities of the dentate gyrus. Our results demonstrate the functional importance of ultraconserved enhancers and highlight that extreme sequence conservation may result from evolutionary selection against fitness deficits that appear subtle in a laboratory setting.

GENOME-WIDE CRISPR PERTURBATION OF CTCF BINDING SITES AT 3D CHROMATIN LOOPS

Oana Ursu*[1], Josh Tycko*[1], Michael Wainberg*[2], Gaelen Hess*[1], Irene Kaplow[2], Peyton Greenside[3], David Morgens[1], Maxwell Mumbach[1], Evan Boyle[1], Nasa Sinnott-Armstrong[1], Michael Snyder[1], Willian Greenleaf[1], Anshul Kundaje^[1,2], Michael Bassik^[1]

[1]Stanford University, Genetics, Palo Alto, CA, [2]Stanford University, Computer Science, Palo Alto, CA, [3]Stanford University, Biomedical Informatics, Palo Alto, CA
* equal contribution, ^ corresponding authors

The 3D organization of the genome into nuclear compartments and topologically associating domains (TADs) influences the range of target genes controlled by regulatory elements. The formation of 3D chromatin loops by CTCF determines boundaries for TADs and perturbation of CTCF binding can alter domains. Thus, modulation of CTCF binding can provide key insights how 3D structure affects genome regulation. However, the exact rules determining the strength of a TAD boundary, its robustness to perturbations and subsequent consequences on cellular phenotypes are largely unknown.

To address these issues, we performed four types of genome-wide perturbation screens for CTCF binding sites at 3D chromatin loops and measured their effects on cellular growth in K562 cells. We tested 4436 CTCF binding sites found at 3D chromatin loop anchors in K562 cells using 2-5 guides per element. A pool of cells was transfected with 16666 guides and the over- and under-representation of each guide was measured after 2 weeks by a sequencing assay. In addition to a CRISPR knockout screen, we tested these same elements in cells harboring the dCasp9, CRISPRi (inhibition) and CRISPRa (activation) systems, allowing us to query how different effectors involving both activation and repression affect the function of CTCF in relation to genome architecture. Across these four screens, we identified 229 CTCF hits, that is binding sites with a significant effect on cellular growth. The majority of hits decreased growth, and a subset of CRISPRa hits promoted growth. Overall, ~80% of CTCF functional regions are located within 1 Mb of a gene known to affect cellular growth in K562 cells based on published gene knockout CRISPR screens. We explored a large collection of genomic, regulatory, chromatin and 3D connectivity features that discriminate CTCF sites that exhibit significant effects from those do not in the different types of screens. Finally, we chose a subset of hits to study in more detail with RNA-seq and ATAC-seq experiments to measure genome-wide regulatory changes induced by the disruption of specific CTCF binding sites.Overall, our parallel screening strategy of CTCF binding sites at 3D chromatin loops advances our understanding of the role of CTCF in genome architecture and genome regulation.

CIRCULOMICS: ULTRASENSITIVE QUANTIFICATION OF EXTRACHROMOSOMAL CIRCULAR DNAS

Massa J Shoura[1], Idan Gabdank[1], Loren Hansen[1], Stephen D Levene[2], Andrew Z Fire[1]

[1]Stanford University School of Medicine, Pathology, Stanford, CA, [2]University of Texas at Dallas, Bioengineering, Biological Sciences, and Physics, Richardson, TX

Investigations aimed at defining the 3-D configuration of eukaryotic chromosomes have consistently encountered an endogenous population of chromosome-derived circular genomic DNA, referred to as extrachromosomal circular DNA (eccDNA). Although the biogenesis, distribution, and activities of eccDNAs remain understudied, eccDNA formation from specific regions of the linear genome has profound consequences for the regulatory and coding capabilities for these regions. Here we use parallel biophysical, enzymatic, and informatic approaches to develop a comprehensive eccDNA profile for C. elegans and in three human cell types, avoiding confounding features of the individual approaches and defining loci of eccDNA formation from both unique and repetitive regions. Of particular interest is a subset of eccDNAs originating from coding regions in C. elegans. Prominent among the limited number of coding regions observed to generate DNA circles are several genes known to produce a diversity of protein isoforms.

This study provides potential genomic tools for molecular and personalized medicine with the capacity to reveal allelic diversity within a genome.

TARGETED REMOVAL OF UNWANTED SEQUENCES FROM SMALL RNA SEQUENCING LIBRARIES

Andrew A Hardigan[1,2], Brian S Roberts[1], Ryne R Ramaker[1,2], Kenneth Day[1], Dianna Moore[1], Richard M Myers[1]

[1]HudsonAlpha Institute for Biotechnology, Myers Laboratory, Huntsville, AL, [2]University of Alabama at Birmingham, Department of Genetics, Birmingham, AL

In next-generation sequencing of small RNAs (smRNAs), highly abundant sequences such as adapter-dimer products and tissue-specific miRNAs can prevent accurate and reproducible identification of lowly expressed species. Previously, we developed a method wherein targeted hairpin adapters were used during library preparation to selectively deplete over-represented miRNAs and improve detection of lowly expressed species. However, this method is incapable of preventing formation of adapter-dimer ligation products that make up a considerable portion of the final library and necessitate laborious gel-separation to remove them prior to sequencing. Here, we have adapted recently described methods for CRISPR/Cas9 – based DASH (Depletion of Abundant Species by Hybridization) to smRNA-seq whereby Cas9 is complexed with sgRNAs targeting adapter-dimer ligation products and overabundant tissue-specific smRNAs for cleavage in vitro. This process dramatically reduces adapter-dimer and targeted smRNA sequences from final libraries and obviates the need for gel-separation, greatly increasing sample throughput with minimal increase in cost. This method is multiplexable, shows minimal off-target effects and improves the identification of lowly expressed miRNAs from human plasma and cell lines. Additionally, the method is fully programmable and can be adapted to remove adapter-dimer from commercial smRNA-seq preparation methods. Like CRISRPR/Cas9 DASH of ribosomal RNA in RNA-seq libraries and mitochondrial DNA in ATAC-seq libraries, our method allows for greater sequencing yield in smRNA-seq and dramatically improves throughput while improving library quality.

VALI**D**ATED **S**YSTEMATIC **I**NTEGRATI**ON**: A VISION FOR EPIGENOMICS IN HEMATOPOIETIC GENE REGULATION

Ross C Hardison[1], Cheryl A Keller[1], Amber R Miller[1], Belinda M Giardine[1], Gerd Blobel[2], David Bodine[3], Mitchell J Weiss[4], James Taylor[5], Yu Zhang[6], Feng Yue[7], Berthold Gottgens[8], Jim Hughes[9], Doug Higgs[9]

[1]The Pennsylvania State University, BMB, University Park, PA, [2]Children's Hospital of Philadelphia, Pediatrics, Philadelphia, PA, [3]National Institutes of Health, NHGRI, Bethesda, MD, [4]St Jude Children's Research Hospital, Hematology, Memphis, TN, [5]Johns Hopkins University, Biology & Comp Sci, Baltimore, MD, [6]The Pennsylvania State University, Statistics, University Park, PA, [7]The Pennsylvania State University, BMB, Coll Medicine, Hershey, PA, [8]University of Cambridge, CIMR, Cambridge, United Kingdom, [9]University of Oxford, IMM Oxford, United Kingdom

VISION is an international, multi-lab project that aims to provide a ValIdated Systematic IntegratiON of epigenomic data in mouse and human hematopoiesis. Technological advances enabling the production of large numbers of rich, genome-wide, sequence-based datasets have transformed biology. However, the volume of data is overwhelming for most investigators. We have formed an interdisciplinary, collaborative team of investigators to address the problem of how to effectively utilize the enormous amount of epigenetic data both for basic research and precision medicine.

At this point, acquisition of data is no longer the major barrier to understanding mechanisms of gene regulation during normal and pathological tissue development. The chief challenges are how to: (i) integrate epigenetic data (e.g. epigenomes and transcriptomes) in terms that are accessible and understandable to a broad community of researchers, (ii) build validated quantitative models explaining how the dynamics of gene expression relates to epigenetic features, and (iii) translate information effectively from mouse models to potential applications in human health.

The products (deliverables) from this project will provide the user community with reliable predictions of genomic regulatory regions in many blood cell lineages, backed by extensive experimental validation using genome-editing methodologies. Furthermore, these predicted regulatory regions will be incorporated into quantitative models for gene regulation, applicable genome-wide and tested in a set of reference loci. We are also building resources to facilitate translations of results of studies between human and mouse based on genomic and epigenomic conservation. The intent is for any investigator to use these publicly available resources to enhance their research. A prototype website is http://www.bx.psu.edu/~giardine/vision/.

This project is supported by the National Institute for Diabetes and Digestive Diseases of the NIH.

NON-ALLELIC GENE CONVERSION IS TEN TIMES FASTER THAN THE RATE OF POINT MUTATIONS IN HUMANS

Arbel Harpak*[1], Xun Lan*[2], Jonathan K Pritchard[1,2,3]

[1]Stanford University, Department of Biology, Stanford, CA, [2]Stanford University, Department of Genetics, Stanford, CA, [3]Howard Hughes Medical Institute, Stanford, CA

Gene conversion is the unidirectional transfer of genetic sequence from a “donor” region to an “acceptor”. In one of its modes, non-allelic gene conversion (NAGC, also known as interlocus gene conversion), the donor and the acceptor are homologous sequences on the same chromatid. Despite the implication of NAGC as the cause of various genetic diseases and its role in the concerted evolution of many human gene families, the rates and contributing factors of NAGC are not well-characterized. Recent gene duplications are of focal interest in studying NAGC as NAGC is contingent on high sequence similarity between donor and acceptor. Notably, NAGC events are expected to distort the genealogy of a gene family at an affected region. Here, we develop tools to survey duplicate gene families across primates in search of such local genealogy distortions. We identify converted regions in 44% of duplicate gene families surveyed. In addition, we find evidence that NAGC substantially increases local G/C content. We further estimate the parameters governing NAGC in humans: a mean NAGC tract length of 1,250bp and a tenfold higher rate of NAGC (per-generation probability of a nucleotide being converted) than point mutations. Despite this seemingly high rate, we show that NAGC likely has only a small average effect on the sequence divergence of duplicates—in contrast to common assumption. This work improves our understanding of the mechanisms behind NAGC and of the role NAGC plays in the evolution of gene duplicates.

* These authors contributed equally to this work

THE CATALOG OF RHESUS VARIATION AND DEVELOPMENT OF BIOMEDICAL MODELS OF HUMAN DISEASES

R. Alan Harris[1], Muthuswamy Raveendran[1], Yue Liu[1], Beth Chaffee[2], Patrick Gillespie[2,3], David Brammer[2,4], Stanton Gray[2], Lawrence Williams[2], Donna Muzny[1], Kim Worley[1], Christian Abee[2], Richard Gibbs[1], Jeffrey Rogers[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Molecular & Human Genetics, Houston, TX, [2]MD Anderson Cancer Center, ME Keeling Center for Comparative Medicine, Bastrop, TX, [3]Eli Lilly & Co, Indianapolis, IN, [4]University of Houston, Houston, TX

Rhesus macaques (*Macaca mulatta*) are among the most evolutionarily successful and widely distributed nonhuman primates. They are also the most common primate model in biomedical research. We previously published population genetics and functional analyses of SNPs discovered by WGS across 133 rhesus from 8 US research colonies. We expanded that cohort to 214 individuals thereby increasing our Catalog of Rhesus Variation to 56,335,807 SNPs. Among the potentially functional SNPs are 3617 stop gained, 396 stop lost, 177,107 missense and 38,846 splicing related variants. We also expanded the scope of analyses to include small 1-60 bp indels and larger insertions based on unmapped Illumina read assembly. We identified 10,495,498 small indels consisting of 4,045,512 insertions, 5,240,168 deletions and 1,209,818 complex alterations. Potentially functional variants among indels include 12,964 frameshifts, 2216 inframe insertions and 2899 inframe deletions. Unmapped read assembly identified 380 novel insertions.

In addition to increasing the cohort size, we included animals with phenotypes modeling human diseases such as colorectal cancer (CRC), cardiomyopathy and circadian rhythm disorders. The CRC cohort consists of 16 rhesus with colon carcinomas and histopathological features similar to those in human hereditary nonpolyposis colorectal cancer (HNPCC) patients. We used linear mixed models to identify SNP and indel associations in CRC cases compared to 198 rhesus controls. Variants in 8 DNA repair genes are associated with human HNPCC. Of 16 rhesus cases, 10 (63%) have at least one predicted damaging variant in a gene associated with human HNPCC, or in its promoter. This includes a *MSH6* missense variant in 5 rhesus and a *MLH1* stop gained in 3 rhesus.

These findings expand our knowledge of variation in the most significant primate model of human disease. Our previous results showed rhesus have ~2.5x more SNP diversity than humans, but, due to the larger effective population size, purifying selection may more efficiently remove slightly deleterious SNPs in rhesus than in humans. Despite more efficient purifying selection, rhesus have a larger absolute number of potentially functional SNPs than humans. Many of these variants have potential for development of primate genetic models of human disease.

DYNAMIC METHYLOME LANDSCAPES DURING MOUSE EMBRYONIC DEVELOPMENT

Yupeng He[1,2], Manoj Hararan[1], David Gorkin[3], Diane E Dickel[4], Chongyuan Luo[1], Rosa G Castanon[1], Joseph R Nery[1], Rongxin Fang[2,3], Huaming Chen[1], Ah Young Lee[3], Yin Shen[5], Barbara Wold[6], Axel Visel[4,7,8], Len A Pennacchio[4,7], Bing Ren[3,9], Joseph R Ecker[1,10]

[1]The Salk Institute for Biological Studies, La Jolla, CA, [2]UC San Diego, Bioinformatics Program, La Jolla, CA, [3]Ludwig Institute for Cancer Research, La Jolla, CA, [4]Lawrence Berkeley National Laboratory, Berkeley, CA, [5]UC San Francisco, San Francisco, CA, [6]California Institute of Technology, Pasadena, CA, [7]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA, [8]UC Merced, Merced, CA, [9]UC San Diego, La Jolla, CA, [10]Howard Hughes Medical Institute, La Jolla, CA

Genetic studies have revealed an essential role for cytosine DNA methylation (mC) in mammalian development. However, little is known about its spatial or temporal dynamics in the post-implantation embryo up until birth. Using deep coverage base-resolution, whole-genome bisulfite sequencing (WGBS), fetal mouse organs were profiled each day in replicate starting from embryonic day 10.5 (E10.5) to postnatal day 0 (P0) (72 tissues/organs in total). We identified 1,808,810 differentially CG methylated regions (DMRs) across the developmental landscape, 92% of which are distal to promoters wherein 36% (n=564,837) mimic the chromatin signatures of active or poised enhancers. Using the REPTILE enhancer prediction algorithm (He et al. 2017), we identified DMRs that are likely linked to active enhancers (n=415,227). The majority of them (n= 310,925) are evolutionarily conserved and enriched in GWAS SNPs associated with variety of human diseases. Strikingly, we found continuous reduction of CG methylation (mCG) in DMRs during development, establishing the tissue-specific hypomethylation patterns. In contrast, gain of methylation was initially small in early embryonic stages but dramatically increased immediately after birth. These mCG changes were anticorrelated with gene expression in specific coexpression modules, which also followed a spatiotemporal pattern. Interestingly many but not all fetal enhancer-linked DMRs are remethylated in adult tissues. These “vestigial enhancers” showed more dynamic chromatin signatures in fetal tissues compared with regions that become active enhancers in adult tissues. Unlike all other tissues, the developing liver methylome undergoes global CG demethylation and remethylation, coinciding with hematogenesis. Finally, we observed that brain tissues accumulated non-CG methylation in the bodies of gene encoding key transcription factors, such as Mef2c and Prox1, which are essential for brain development. Taken together, our study reveals a highly dynamic methylation landscape throughout postimplantation development, which can serve as a foundation for studies of human embryonic development.

Reference

He et al. *PNAS,* 2017 (PMID: 28193886)

ULTRA-FAST SEQUENCE CLASSIFICATION WITH TAXONOMER – A CASE STUDY OF *STAPHYLOCOCCUS AUREUS*

Edgar J Hernandez[1,2*], Aurélie Kapusta[1,2*], Steven Flygare[2,3], Hillary Crandall[6], Anne Blaschke[6], Carrie Byington[7], Robert Schlaberg[3,4,5], Mark Yandell[1,2,3]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]IDbyDNA Inc., San Francisco, CA, [4]University of Utah, Department of Pathology, Salt Lake City, UT, [5]ARUP Institute for Clinical and Experimental Pathology, Salt Lake City, UT, [6]University of Utah, Department of Pediatrics, Salt Lake City, UT, [7]Texas A&M, College of Medicine, College Station, TX

*Authors contributed equally

The ever-increasing volume of high-throughput sequencing data creates an urgent need for software that can rapidly and accurately process large genomic and transcriptomic datasets. Taxonomer [1,2], enables ultra-fast sequence classification applicable to both discovery and diagnostic purposes.

Staphylococcus aureus is a common infectious pathogen that causes severe disease in children and adults. These gram-positive bacteria are becoming increasingly more difficult to treat, as some strains have acquired resistance to the most effective antibiotics, such as methicillin [methicillin-resistant *S. aureus* (MRSA)]. We used Taxonomer to rapidly analyze more than 350 *S. aureus* genomes and classify genomic reads against the 2876 annotated genes of a MRSA reference strain (USA300). We identified 14 genes highly associated with MRSA samples when compared to methicillin-suseptible strains (MSSA). Included in this gene set are known methicillin resistance genes such as *mecA*, *ccrB* and *ccrA*, but also genes of unknown function. Using these 14 marker genes as a mini database, Taxonomer can be used to rapidly and precisely distinguish between patients with MRSA or MSSA. These data demonstrate the ease with which Taxonomer can be used as a diagnostic tool and as a research tool to discover bacterial markers associated with resistant strains. Further studies are underway to elucidate the physiologic role played by the genes of unknown function associated with MRSA isolates.

[1] Flygare, Simmon et al. (2016). Genome Biology
[2] http://taxonomer.iobio.io/info.html

SINGLE-CELL ANALYSIS OF CLONAL DYNAMICS OF CHILDHOOD ALL REVEALS A ROLE FOR TRANSCRIPTIONAL HETEROGENEITY IN DRIVING RESISTANCE TO CHEMOTHERAPY

Virginia A Turati[1], J. Afonso Guerra-Assunção[1], John C Ambrose[1], John Brown[1], Michael Hubank[2], Mark Lynch[3], Bernadett Gaal[1], Lucia Conde[1], Javier Herrero[1], Sten E Jacobsen[4], Tariq Enver[1]

[1]UCL Cancer Institute, University College London, London, United Kingdom, [2]Molecular Diagnostics, Royal Marsden Hospital, London, United Kingdom, [3]Fluidigm, San Francisco, CA, [4]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom

While intratumor heterogeneity has been long recognized, its biological and clinical significance is not well understood. To determine the relative contribution of different sources of heterogeneity to therapeutic resistance in childhood acute lymphoblastic leukemia (cALL), we have developed a mouse model that affords longitudinal analysis of subclonal dynamics. By assessing the reproducibility of independent outcomes, one can distinguish between deterministic and stochastic mechanisms of selection during therapy.

We use a combination of single cell assays to track the fate of individual genetic subclones. Using multicolor-FISH, we show that, while treatment of sensitive ALLs results in a striking reduction in leukemic burden, the overall extent of genetic diversity is largely unaffected. We have adapted PicoPLEX whole-genome amplification to the Fluidigm C1 platform to produce higher resolution single-cell copy number maps. While FISH captures all the major clones, WGS data dissect heterogeneity in greater detail, revealing the coexistence of clones carrying the same lesion but distinct breakpoints, formally proving convergent evolution of cALL.

Functional analysis based on limiting dilution secondary transplantation assays shows that chemotherapy enriches for cells with tumor propagating potential. Bulk transcriptome analysis of treated and untreated cells reveals that the former have distinct signatures shared amongst metastatic sites. Resistant cells display differences in cell cycle, differentiation status, transcription rate and activation of key signaling pathways. At the single cell level, the treatment naïve population displays higher variance both across cells and genes. These observations highlight a high degree of heterogeneity, which disappears as a result of a chemotherapy induced bottleneck selection process. Importantly, a small population of untreated cells appears to have resistance potential as they exhibit the same transcriptional program.

All together the data suggest that genetic heterogeneity in cALL arises through independent acquisition of copy number alterations affecting the same gene or genes within a pathway, thereby resulting in convergent phenotypic evolution; resistance is instead likely driven by a pre-existing population of immature cells with a distinct global gene expression signature and higher tumor propagating potential.

COLLECTING ROADMAP AND ENCODE DATA INTO REFERENCE EPIGENOMES

Jason A Hilton, Kathrina C Onate, Cricket A Sloan, Esther T Chan, Idan Gabdank, Aditi K Narayanan, J. Seth Strattan, Marcus Ho, Ulugbek Baymuradov, Forrest Tanaka, Christopher Thomas, Tim Dreszer, Benjamin C Hitz, J. Michael Cherry

Stanford Unversity, Department of Genetics, Palo Alto, CA

The Encyclopedia of DNA elements (ENCODE) project has produced data from more than 8,000 experiments using a variety of techniques to study the structure, regulation, and transcription profiles of human and mouse genomes. The data from these experiments first pass through the ENCODE Data Coordination Center (DCC) for basic validation and metadata standardization before they are openly available at the ENCODE site (https://www.encodeproject.org/). The ENCODE portal also hosts data from external projects, including the NIH Roadmap Epigenomics Mapping Consortium. In order to align with the existing structure of the ENCODE portal, Roadmap experimental metadata were carefully curated, and raw data were collected from production labs and public repositories. The import of these experiments allows for Roadmap data to be searched, downloaded, and analyzed alongside ENCODE data. Furthermore, the ENCODE DCC is currently processing Roadmap data using the uniform processing pipelines designed by the ENCODE consortium. This provides high-quality and consistent data with cross-project compatibility. One of the aims of the Roadmap project was to provide to the scientific community a set of reference epigenomes for human cells and tissues. Using data produced from the Roadmap project, 111 reference epigenomes were assembled, plus an additional 16 reference epigenomes from ENCODE data, each of which have been modeled on the ENCODE portal. The ENCODE DCC has also curated data produced by the ENCODE consortium to compile reference epigenomes for human and mouse biosamples, following guidelines set forth by the International Human Epigenome Consortium. These reference epigenomes, combined with the open access to ENCODE's uniform processing pipeline code, provide the framework for ease of comparison and integration within future studies from the scientific community.

INVESTIGATING THE REGULATORY RELATIONSHIPS BETWEEN TRANSCRIPTION FACTORS AND THEIR TARGETS IN *SACCHAROMYCES CEREVISIAE* USING RNA-SEQ

Josephine Ho[1], Joseph D Coolon[2]

[1]Wesleyan University, Molecular Biology and Biochemistry, Middletown, CT, [2]Wesleyan University, Biology, Middletown, CT

Gene expression is regulated at the level of transcription by the combinatorial control of numerous transcription factor (TF) proteins that bind to TF binding sites in each gene's cis-regulatory elements. Cooperatively, these TFs act as activators or repressors to modulate the expression of the focal gene. Because the expression of each TF is also regulated this way, each gene in the genome is typically controlled by complex, multi-level and highly interconnected regulatory networks. The study of regulatory networks typically defines the edges in these networks by chromatin immunoprecipitation of DNA bound by a TF followed by high-throughput sequencing (ChIP-seq). Using this approach, researchers have begun to define regulatory network structure by the association of TFs with the sites on each gene's cis-regulatory sequences to which they bind. While much progress has been made in defining genome-wide gene regulatory network structure in many model systems using ChIP-seq and other similar methodologies, very little is known about how information flows through these networks. We are now investigating TF-target gene expression relationships quantitatively in the yeast *Saccharomyces cerevisiae* using Tet-titratable TF expression modification followed by RNA-seq. Using these RNA-seq analyses, we have investigated gene expression patterns that emerge downstream in a regulatory network as a consequence of altered TF expression levels. Assessment of these patterns will serve to establish a foundation for understanding properties of information flow in these biological networks.

ASSEMBLY-BASED CHARACTERIZATION OF LONGITUDINAL METAGENOMIC SAMPLES

Larson J Hogstrom[1], Moran Yassour[1,2], Mikael Knip[3], Ramnik J Xavier[1,2], Eric Lander[1]

[1]The Broad Institute of MIT and Harvard, Cambridge, MA, [2]Mass. General Hospital, Center for Computational and Integrative Biology, Boston, MA, [3]National Institute for Health and Welfare, Dept. of Health, Turku, Finland

Many important health states are characterized by reduced complexity in the gut's microbial diversity. In early life, for example, microbial communities in the gut often display a low-complexity composition, comprised of 1-3 abundant species in the first week following birth. Here, we demonstrate a computational method that evaluates metagenomic complexity transitions in an individual's time course using assembly graphs created from short, paired-end metagenomic sequencing samples. Our approach compares the complexity of DNA assembly graphs corresponding to the genomic regions associated with two classes of marker genes. First, de Bruijn graphs are computed for genomic sequences corresponding to universal, single-copy genes that are abundant across many taxa of bacterial reference genomes. Second, assembly graphs are created for sequences associated with clade-specific marker genes. We introduce a measure called universal-to-specific assembly complexity ratio (USACR) to compare the complexity of de Bruijn graphs associated with the two gene classes. Evaluating both clade-specific and universal gene assemblies, our method enables a robust measure of metagenomic complexity transition events. Furthermore, we provide a path to better formalize genetic variation found in genomic regions are that uniquely informative in the assessment of taxa abundance. The approach is applied to longitudinal gut metagenomic samples from 48 infants. These time courses typically had a sample within 24 hours of birth which was characterized by a dominant abundance of a single species (e.g., Escherichia coli or Bifidobacterium longum). Our approach can be generalized to other metagenomic time-course data to quantify microbiome complexity transition events in early life or disease-related states.

THE IDENTIFICATION OF RECENT FINE-SCALE POPULATION STRUCTURE AND IMPACT ON GENOME-WIDE ASSOCIATION STUDIES.

Eurie L Hong[1], Julie M Granka[1], Eunjung Han[1], Ross E Curtis[2], Peter Carbonetto[1], Jake Byrnes[1], AncestryDNA Team[1,2], Kenneth G Chahine[2], Catherine A Ball[1]

[1]AncestryDNA, San Francisco, CA, [2]AncestryDNA, Lehi, UT

Identifying the geographic or historical origins of individuals using genetic data has broad applications in understanding human evolution and demography as well as interpreting human health and disease. Numerous approaches based on sequence variations that differentiate populations have been developed to achieve these goals. Additionally, methods to associate single nucleotide polymorphisms (SNPs) with diseases and traits account for such population structure. Current association approaches, however, are limited in identifying and understanding the impact of recent fine-scale population structure that represents recent demography. To address this, we identify population structure using a community detection algorithm across a network of identity-by-descent from nearly 1 million AncestryDNA customers who have consented to research. Identified sub-populations, such as individuals from the U.S. Appalachian region or French Canadians in the US, correspond to historical trends, geographic isolation, and cultural forces that shaped genetic variation within the last several hundred years. We suggest that current GWAS approaches could be susceptible to such fine-scale population structure but when properly addressed, may mitigate unknown confounders.

INTEGRATIVE BIOBANKING TO INVESTIGATE GENETICS, METAGENOMICS, AND SOCIAL BEHAVIOR IN CAYO SANTIAGO RHESUS MACAQUES

Julie E Horvath[1,2,3], Michael Montague[4], Noah Snyder-Mackler[3], Athy Robinson[5], Karli Watson[6], Lauren Brent[7], JH Pate Skene[8], Michael Platt[4,9,10]

[1]North Carolina Central University, Biological and Biomedical Sciences, Durham, NC, [2]North Carolina Museum of Natural Sciences, Genomics & Microbiology, Raleigh, NC, [3]Duke University, Evolutionary Anthropology, Durham, NC, [4]University of Pennsylvannia, Neuroscience, Philadelphia, PA, [5]Duke University, Neuroscience, Durham, NC, [6]University of Colorado Boulder, Institute of Cognitive Science, Boulder, CO, [7]University of Exeter, Centre for Research in Animal Behavior, Exeter, United Kingdom, [8]Duke University, Institute for Brain Sciences, Durham, NC, [9]University of Pennsylvannia, Marketing, Philadelphia, PA, [10]University of Pennsylvannia, Psychology, Philadelphia, PA

Understanding the intricate connections between social behavior and underlying genetics requires integrative datasets and multi-faceted approaches. Toward this end, for more than eight consecutive years we have compiled genetic and behavioral data and biological samples from approximately 1000 free-ranging Indian origin rhesus macaques (Macaca mulatta) living on Cayo Santiago in Puerto Rico. These highly social animals are excellent models for understanding variation in human social behavior, including individuals with autism spectrum disorder (ASD). We have amassed extensive behavioral and cognitive datasets for the monkeys through assays and focal follows. We have extracted DNA from whole blood for more than 1000 animals and have conducted whole genome sequencing on over 250 individuals identifying predicted amino acid changes in key neuromodulatory pathway genes. More focused genetic analyses from more than 500 animals include VNTR (*AVPR1A, TPH2, 5HTTLPR* and *MAOA*) and re-sequencing efforts of more than 90 genes linked to ASD and other cognitive pathways. Additional datasets include skin, lung, fecal and digestive tract microbiota, as well as twenty peripheral tissues and organs sampled from hundreds of animals. Whole blood and an unprecedented set of flash frozen brain sections from over 100 animals are currently being used for gene expression (RNAseq) and epigenomic (ATAC-seq) analyses. Here we assess the skin microbiome from more than 50 animals to illustrate the power of how our integrative dataset can correlate microbiota variation with behavioral metrics, immunity and rhesus macaque genetic variation.

IS YOUR ASSEMBLY GOOD ENOUGH? EVALUATING AND IMPROVING REFERENCE GENOME ASSEMBLIES WITH GEVAL

Kerstin Howe, William Chow, Richard Durbin

Wellcome Trust Sanger Institute, Human Genetics, Hinxton, Cambridge, United Kingdom

The advent of long-range sequencing and mapping technologies is driving the generation of an ever increasing range of high quality vertebrate genome assemblies in a time- and cost-effective manner. However, it is not always trivial to decide which one of a multitude of possible assemblies for a certain dataset is the best one and whether it is good enough to be considered the reference for the respective species. Being a member of the Genome Reference Consortium (GRC), our group has developed pipelines and tools to assess and compare assembly quality and to take active measures to improve it, both through automated approaches and through detailed manual curation.

Our initial assessment pipeline combines publicly available tools for gathering length metrics and the computation of the assembly likelihood score with contamination screening and assessment of core gene presence. This is complemented by further analysis in our bespoke genome assembly evaluation browser, gEVAL (geval.sanger.ac.uk). gEVAL features a wide variety of aligned data, typically including optical/genome maps, clone end sequences and transcript sequence as well as whole genome alignments to other assemblies of the same species. It detects any lack of concordance and allows easy and intuitive visual access to assembly issues, both through issue lists and through colour coding in the genome browser. It facilitates developing assembly improvement strategies, but also allows any user to easily assess whether a certain region of interest is of sufficient quality and therefore a reliable basis for their research findings.

gEVAL is one of the central tools of the GRC, providing the basics for reference genome curation, but has also been used successfully in supporting the Mouse Genomes Project in creating assemblies of 16 mouse strains and the Pig and Chicken Genome Projects in their latest releases. We are now working with the Avian Genomes Project to assess hummingbird assemblies and with Genome10K and the Vertebrate Genomes Project to evaluate the increasing numbers of assemblies of other vertebrates, with a primary focus at the Sanger Institute on fish and mice.

A MODEL-BASED APPROACH FOR POPULATION GENOMIC INFERENCE OF TUMOR GROWTH DYNAMICS

Zheng Hu, Ruping Sun, Christina Curtis

Stanford University, Departments of Medicine and Genetics, Palo Alto, CA

Cancer results from the acquisition of somatic alterations in a microevolutionary process that typically occurs over many years, much of which is occult. Understanding the evolutionary dynamics that are operative at different stages of progression in individual tumors may inform the earlier detection, diagnosis, and treatment of cancer. Although these processes cannot be directly observed, the resultant spatiotemporal patterns of genetic variation amongst tumor cells encode their evolutionary histories. Here we describe an agent-based modeling framework that enables the efficient simulation of 'virtual' tumors (~10^9 cells) growing with explicit spatial constraints and under different modes of evolution (neutral or varying levels of selection). By integrating this spatial computational model with a statistical inference framework based on Approximate Bayesian Computation (ABC), to analyze patterns of genetic variation based on tumor sequencing data, we demonstrate the robust inference of clinically relevant patient-specific parameters including the mutation rate, clone mixing and strength of selection. We further show how application of this tumor evolutionary dynamic framework yields quantitative insights into human tumor progression.

WHOLE METAGENOMIC SHOTGUN SEQUENCING ANALYSIS OF ANTIBIOTIC RESISTANCE GENES IN WOMEN WITH SYMPTOMATIC BACTERIAL VAGINOSIS

Bernice Huang, Vaginal Microbiome Consortium at VCU

Virginia Commonwealth University, Center for the Study of Biological Complexity, Richmond, VA

Vaginal complaints are one of the most common reasons for women seeking medical attention worldwide and clinical symptoms are often misdiagnosed. New insights into the diversity of human microbial flora, including vaginal flora, have been made possible through DNA sequencing via culture-independent approaches. As part of the Vaginal Human Microbiome Project at Virginia Commonwealth University (Vahmp), we sampled thousands of women from the clinical setting to explore the relationship of the vaginal microbiome with various physiological and infectious conditions. Here, we report an analysis of vaginal microbiome profiles of close to 5000 women using deep sequencing of 16S rRNA partial sequences. Species-level analysis of vaginal microbiome profiles reveals how the samples further cluster into distinct ‘vagitypes,’ driven by the predominant bacterial species in the sample. The vagitypes include several *Lactobacillus* types (*L. crispatus*, *L. iners*, *L. gasseri*, *L. jensenii*, *L. delbruecki*), a *Gardnerella vaginalis* type, a type dominated by bacterial vaginosis-associated bacterium (BVAB1), and more than fifty additional rare and minor vagitypes. A complete species-level analysis reveals more than 200 species-level taxa in the vaginal microbiome including ~35 novel or uncharacterized taxa.

Recurrent bacterial vaginosis (BV) infection is one of the biggest clinical challenges in women's urogenital health. A likely cause of treatment failure is the presence of antimicrobial resistance genes. With whole metagenome sequencing it is possible to analyze antibiotic resistance (AR) and virulence markers in entire microbial communities. A recent study reported an overall increase in abundance and number of antimicrobial resistance genes in women with BV compared with women without. To assess AR gene analysis in BV, we used ShortBRED to generate short unique markers for all AR gene families in an AR-specific gene database. Shotgun sequences were mapped to the resulting AR-specific markers and normalized across samples to generate AR profiles for the metagenomes of 64 women with clincially diagnosed BV.

This work was supported by grants from the NIH Common Fund Human Microbiome Project program (1UH2/UH3AI08326 and 8U54HD080784 and to G. Buck, K. Jefferson and J. Strauss). We thank the members of the Vaginal Microbiome Consortium at VCU (vmc.vcu.edu) for their hard work on this project, and the Global Alliance for the Prevention of Prematurity and Stillbirth for their participation in launching the MOMS-PI project.

TRACKING SUBCLONAL METASTATIC EXPANSION IN TRIPLE NEGATIVE BREAST CANCER

Xiaomeng Huang[1], Yi Qiao[1], Samuel Brady[2], Andrea Bild[2], Gabor Marth[1]

[1]University of Utah, Eccles Institute of Human Genetics, Salt Lake City, UT, [2]University of Utah, Department of Pharmacology and Toxicology, Salt Lake City, UT

Metastatic breast cancer is an advanced-stage disease in which the cancer cells have spread to distant organs, e.g. bones, liver, brain and lung. This type of breast cancer accounts for approximately 6%-10% of all breast cancer diagnoses, with a dramatically lower 5-year survival rate of 22%. The goal of this study is to dissect metastatic tumor expansion at a subclonal level, in order to identify its genomic drivers, as well as the aggressive colonizing subclones seeding new metastatic sites. As our driving dataset, we have collected two primary tumor biopsies, one at initial diagnosis and one at mastectomy necessitated by the patient's relapse; 26 metastatic tumors across seven organs via a rapid autopsy procedure hours after the patient's decease; as well as two skin biopsies to be utilized as normal control tissues. All samples were subjected to exome-enriched whole genome sequencing with an average genomic coverage of 60X, and higher in exonic regions. Variant calling revealed inherited variants in genes involved in estrogen regulatory networks, hormone receptor mediated transactivation pathways and cell migration. BRCA2, but not BRCA1, harbored inherited missense variants. Somatically acquired homozygous TP53 missense variants and RB loss were present in all tumor samples, explaining the widespread chromosomal aberrations we observed in all tumor samples, including both copy number variations (CNVs), and large regions with loss of heterozygosity (LOH). We have developed computational techniques to determine the time ordering of the chromosomal events, the likely order in which the various metastatic lesions were established, and reconstructing the subclonal evolution of tumor across the multiple primary and metastatic sites. Surprisingly, but consistently across chromosomal events and somatic SNVs, our analysis revealed that the site of origin for the patient's tumor was in the lung, and that the presumed primary (and the mastectomy) breast sample in fact were the result of metastatic colonization from this site. These finding suggest that either an early, undetected primary breast tumor migrated to the lung, and later returned to the primary site; or that the primary lesion was a lung tumor, which metastasized first into the breast, then into other organs. Our method reveals four metastatic "waves": after the initial breast metastasis, the tumor invades abdominal organs (liver, pancreas), lymph nodes, and finally, moves to the brain and bones. Trained on a large (perhaps currently the largest) metastatic biopsy dataset from a single patient, our method provides a novel framework to simultaneously analyze CNV, LOH, and SNV data to reconstruct metastatic tumor expansion at subclonal resolution. Dissection of the genetic changes that define the most aggressive colonizing subclones forms a starting point to identify and target the genetic/genomic drivers, in the hopes of disrupting the metastatic process in the patient.

DIET NETWORKS FOR PERSONAL GENOMICS: A DEEP LEARNING APPROACH

Julie Hussin[1,2], Adriana Romero[3], Pierre Luc Carrier[3], Maxime Barakatt[2], Akram Erraqabi[3], Tristan Sylvain[3], Alex Auvolat[3], Etienne Dejoie[3], Marc-André Legault[1], Marie-Pierre Dubé[1,2,4], Jean-Claude Tardif[1,2], Yoshua Bengio[3]

[1]University of Montreal, Faculty of Medicine, Montreal, Canada, [2]Montreal Heart Institute, Research Centre, Montreal, Canada, [3]University of Montreal, Montreal Institute of Learning Algorithms, Montreal, Canada, [4]Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada

In the last decades, advances in genomic technologies resulted in an explosion of data, propelling human genomics into the Big Data era and its challenges. Key methodological advances, such as genome-wide association studies, have helped define the genetic components of both human disease and drug response, and opened the door to personal genetics. Here, we use large genomic datasets to assess the ability of deep learning techniques to extract useful information in the context of precision medicine. Machine learning problems in genomics pose an important challenge, loosely called “the fat data problem”: the number of input features can be orders of magnitude larger than the number of training examples, making it difficult to avoid overfitting. We propose a novel neural network architecture, that we call *Diet Networks*, which considerably reduces the number of free parameters in the fat layers of the model. The *Diet Networks* parametrization is based on the idea that we can first learn a distributed representation for each input feature (feature embedding) and then learn how to map a feature’s representation to the parameters linking feature's values to the hidden units of the network. This embedding can be learned using other datasets and prior knowledge about the features, enabling experts’ input into this deep learning approach. We show experimentally on a population stratification task that the proposed approach considerably lowers the number of free parameters in the fat layers of the model and significantly reduces the error rate of the classifier. We also describe how we can extract useful information to understand learned features and help interpretability of the approach. We further extend the model to incorporate imputation tasks, to deal with missing data. In the next steps, we will use simulated and well-established datasets in human genomics (eg. The Wellcome Trust Case Control Consortium, the 1000 Genomes, the UK Biobank), as well as data from the Montreal Heart Institute biobank and pharmacogenomic cohorts available at the Beaulieu-Saucier Pharmacogenomic Center, to investigate the performance of this deep learning approach to predict phenotypes across a spectrum of health-related traits. Improving the applicability of deep learning techniques in handling such datasets, as well as making deep learning models interpretable, can have an important impact in medical research, more specifically in precision medicine, where high-dimensional data regarding a particular patient is used to make predictions on health outcomes.

NONCODING INDEL HOTSPOTS TARGET LINEAGE-DEFINING GENES IN HUMAN CANCER

Marcin Imielinski[1,2], Guangwu Guo[3,4], Matthew Meyerson[3,4]

[1]Weill Cornell Medicine, Pathology, New York, NY, [2]New York Genome Center, Cancer program, New York, NY, [3]Broad Institute, Cancer program, Cambridge, MA, [4]Dana Farber Cancer Institute, Medical Oncology, Cambridge, MA

Whole genome sequencing analysis of lung adenocarcinomas revealed indel hotspots in surfactant protein genes (SFTPA1, SFTPB, and SFTPC). Extrapolation to other solid cancers demonstrated highly recurrent and tumor-type-specific indel hotspots targeting the noncoding regions of highly expressed genes defining certain secretory cellular lineages: albumin (ALB) in liver carcinoma, gastric lipase (LIPF) in stomach carcinoma, and thyroglobulin (TG) in thyroid carcinoma. The sequence contexts of indels targeting lineage-defining genes were significantly enriched in the AATAATD DNA motif and specific chromatin contexts, including H3K27ac and H3K36me3.

To identify regions of the genome under positive somatic mutational selection in lung cancer, we analyzed whole genome sequencing reads from 79 lung adenocarcinoma tumor-normal pairs. The three most significantly altered loci in the indel analysis of lung adenocarcinomas overlapped the genes SFTPB (P = 1.8 x 10-14), SFTPA1 (P = 4.8 x 10-10), and SFTPC / BMP1 (P = 1.3 x 10-7). SFTPB, SFTPA1, and SFTPC encode surfactant proteins that are specific markers of type II pneumocytes in the normal lung, where they help generate the surface tension required to maintain open air spaces. SFTPB, SFTPA1, and SFTPC demonstrate striking lung-specific expression in Genotype-Tissue Expression (GTEx) data (3), obtained from healthy human tissues.

Scanning over 500 genomes in 12 cancer types for somatic variants in the three SFTP loci, we found a 25-fold enrichment (95% CI: [13.2, 47.4]) in lung adenocarcinoma vs. other tumor types, even after correcting for sample-specific variations in indel density (P=5.6 x 10-23, Wald test, logistic regression). Given the specificity of SFTP gene expression to lung tissue and SFTP somatic indels to lung adenocarcinoma, we hypothesized that other tumor types might harbor similar noncoding indel enrichment at highly expressed and lineage-specific genes. Through analysis of 2917 GTEx samples spanning 30 normal tissues and 487 whole genome sequenced samples spanning 12 cancer types (other than lung adenocarcinoma), we found that lineage-specific gene territories were 14.3 fold (95% CI: [10.7, 19.2], P = 9.5 x 10-70, Wald test) enriched in indels in the expression native vs. foreign tumor context.

Our findings illuminate a prevalent and hitherto unrecognized mutational process linking cellular lineage and cancer. The findings may also represent a novel noncoding driver phenomenon that is under selection in multiple cancer types.

USING THE UNIQUE BREEDING HISTORY OF COLDBLOODED TROTTERS TO IDENTIFY GENES THAT INFLUENCE ATHLETIC PERFORMANCE

Kim Jäderkvist Fegraeus[1], Brandon D Velie[1], Jennifer R S Meadows[2], Leif Andersson[2,3], Gabriella Lindgren[1]

[1]Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Uppsala, Sweden, [2]Uppsala University, Department of Medical Biochemistry and Microbiology, Uppsala, Sweden, [3]Texas A&M University, Department of Veterinary Integrative Biosciences, College Station, TX

The origin of the Coldblooded trotter (CBT) provides a unique opportunity to identify genes influencing racing performance. The CBT originates from the North Swedish draught horse (NSH) and these two breeds retain high levels of genetic similarity (Fst = 0.08). However, prior to the introduction of paternity testing in 1969, crossbreeding with Standardbreds (SB) was used to improve CBT performance. We hypothesize that the gains in CBT performance over the last 50 years may in part be explained by the maintenance of favorable genetic variants originating from the SB. As such, the aim of the current study was to compare the genetic makeup of these three breeds and to identify genetic footprints of athletic performance. A sliding window Delta Fst analysis was performed across all breeds using data generated from the equine SNP50K array (CBT, n=11; NSH, n=19; SB, n=12). Five key regions were revealed where the CBT and SB were genetically similar, but together differed from the NSH. These regions included the tsukushi (TSKU) and F-box and leucine rich repeat protein 4 (FBXL4) genes. Based on the relationship between the CBT and the NSH as well as the intensive selection for racing performance in CBT, we believe that these genes are important for harness racing performance. This proposed association is currently being followed up in 475 randomly selected CBTs, and the results for these genes and any other potential candidate genes will be presented at the meeting. The results from this study will provide important information to the horse industry and assist in the selection of better racehorses. To further understand the function of the genes, it would also be interesting to perform comparative studies in other species, for example humans.

GENETIC REGULATION OF THE HUMAN CORTEX TRANSCRIPTOME IN SCHIZOPHRENIA

Andrew E Jaffe[1,2]

[1]Lieber Institute for Brain Development, Translational Sciences, Baltimore, MD, [2]Johns Hopkins Bloomberg School of Public Health, Mental Health, Baltimore, MD

Genome-wide association studies (GWAS) have identified over 108 loci that confer risk for schizophrenia, but risk mechanisms for individual loci are largely unknown. In order to more fully characterize the transcriptional correlates of genetic risk, we performed genotyping and RNA sequencing of dorsolateral prefrontal cortex (DLPFC) tissue from 412 subjects, including 175 subjects with schizophrenia. RNA-seq data was summarized to five convergent transcript features – assembled transcript isoforms, annotation-dependent gene and exon counts, and annotation agnostic exon-exon splice junction counts and contiguously expressed regions.

We performed expression quantitative trait loci (eQTL) analyses across these five feature summarizations and identified widespread genetic regulation of nearby expression. We subsequently integrated two additional independent DLPFC datasets – the CommonMind Consortium (N=285) and GTEx (N=99) to form a significant and replicated set of high-confidence eQTLs ("core eQTLs"). The majority of these core eQTLs were largely gene-specific, consistent across both interrogated ethnic groups, very proximal to the TSS of genes, and expression of the majority of genes was associated with more than 1 linkage disequilibrium-independent SNP. We further found the largest effect sizes on average were for the junction and ER eQTLs, and interestingly a large proportion of these core eQTLs corresponded to unannotated transcriptional activity. Furthermore, while the majority of genes with features as eQTLs have multiple possible Ensembl transcript isoforms, 67.0% of exon eQTLs and 31.1% of junction eQTLs were specific to a single Ensembl transcript.

We then explored the landscape of eQTLs associated with genetic risk for schizophrenia using the set of core eQTLs among the GWAS-significant risk SNPs, and found that 46 (42.5%) risk SNPs significantly associated with expression levels of 764 nearby features at FDR < 1% significance. Integrating schizophrenia-control expression differences further identified core eQTL signal across a subset of 34 GWAS loci with convergence of allele and illness state expression directionality. These loci often point to individual "risk" genes or even more specific "risk" transcripts that can represent targetable entry points for more focused cellular assays and model organism work to better characterize schizophrenia risk mechanisms.

By combining genetic risk at the population-level with eQTLs and case-control differences, we identify putative human frontal cortex mechanisms underlying risk for schizophrenia and replicable molecular features of the illness state.

REVEALING THE CAUSATIVE VARIANT IN MENDELIAN PATIENT GENOMES WITHOUT REVEALING PATIENT GENOMES

Karthik A Jagadeesh*[1], David J Wu*[1], Johannes Birgmeier[1], Dan Boneh[1,2], Gill Bejerano[1,3,4]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Electrical Engineering, Stanford, CA, [3]Stanford University, Developmental Biology, Stanford, CA, [4]Stanford University, Pediatrics (Medical Genetics), Stanford, CA

*Authors contributed equally

Rare diseases affect 1 in 33 babies. Exome and genome sequencing have revolutionized the diagnosis of thousands of rare Mendelian diseases to thousands of different human genes. Thousands of additional rare Mendelian diseases and human genes await discovery. Frequency-based filters have proven extremely effective in providing diagnosis in such cases. In essence, variants found in a control population (common variants) are likely to be benign while functional rare variants not found in the control population but seen in multiple affected individuals are likely to be disease causing. These filters seek the gene or variant present in most affected individuals but in very few unaffected individuals.

Frequency-based computation highlights the fundamental “serve or protect” dilemma of genomic data: “Serve:” to find the root cause of a patient’s disease, one wishes to compare a patient genome to as many other genomes as possible, both affected and unaffected, related and unrelated. Thus, to advance modern medicine, all sequenced genomes should be shared. “Protect:” one’s genome continues to reveal more and more about oneself, including critical health information and susceptibility to a variety of diseases. Sharing it with others can lead to discrimination and bias. To protect its owner and next of kin, no sequenced genome should be shared.

We introduce here a modern, proof-of-concept cryptographic implementation which both serves and protects. The secure computation can be run on entire genomes (Serve), while no party involved in the computation learns anything about the inputs of the other participants except for the output which is computed together (Protect). We use real patient data to show that our secure implementation reveals minimal information while diagnosing patient genomes through 3 different strategies (small patient cohorts, trio analysis, two hospital collaboration) using practical amounts of compute time and memory. The causal variant is discovered jointly, while keeping up to 99.7% of all participants’ most sensitive genomic information private. All similar frequency-based operations performed today to diagnose such cases are done openly, keeping 0% of participants’ genomic information private. This work will help usher in an era where genomes can be both utilized and truly protected.

TRANSCRIPTOME AND OPEN CHROMATIN PROFILING OF EMBRYONIC STEM CELLS DIFFERENTIATION TO DEFINITIVE ENDODERM IN MULTIPLE MAMMALIAN SPECIES

Shan Jiang[1,2], Christina R Wilcox[1,2], Ali Mortazavi[1,2]

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

The pluripotency of embryonic stem cells (ESCs) to differentiate into different germ layers makes it an attractive model to study cell-fate commitment and differentiation. However, the underlying gene regulation processes and their level of conservation across different mammals are not fully understood. We focus on differentiating ESCs to definitive endoderm (DE) in human, mouse, rat and marmoset in vitro with the presence of Activin A. We applied RNA-seq and ATAC-seq to observe the transcriptome and open chromatin regions profiles during the time-course of human and mouse DE differentiation. We observed dynamic expression changes of DE markers, such as CXCR4, SOX17 and FOXA2, and lowly expression or silencing of mesoderm and ectoderm specific markers during DE differentiation for human and mouse. We integrated dynamic gene expression and chromatin accessibility profiles to construct gene regulatory networks during DE differentiation. We also used single-cell RNA-seq (scRNA-seq) to examine individual cells pluripotency at ESCs stage and we will continue to apply scRNA-seq to understand how individual cells transit to DE lineage during differentiation to identify the conserved and species-specific portions of the DE program.

SINGLE CELL RNA-SEQUENCING OF CELL LINE WITH CHROMIUM AND OTHER PLATFORMS

Yukie Kashima, Yutaka Suzuki

The University of Tokyo, Graduate school of Frontier Sciences, Kashiwa, Japan

Single cell sequencing is a powerful tool to investigate the clonal evolution, cellular diversity, and to understand the characteristics of rare cells. Although single cell RNA-seq (scRNA-seq) is currently most advanced among those single-cell sequencing technologies, the data handling of scRNA-seq can be difficult, due to its complexity in the data. In order to investigate and solve these problems, we evaluated, compared some scRNA-seq platforms.

First, we evaluated the performance of Chromium system of 10X Genomics, and then we compared the obtained data with from another single cell analytical platforms. Using Chromium, we measured the cell doublet rates for the mixture of mouse NIH/3T3 and human HEK293 cells as materials. Next, we applied Chromium single cell technology to several cancer cell lines. Chromium can process larger number of cells per sample and even multiple samples at the same time, however the number of sequence tags allocated for each single cell were inherently smaller. In the case of single cell sequencing of a cancer cell line PC-9, Chromium and Fluidigm C1 system yielded 5,780 cells and 46 cells, respectively. For the sequence depth, Chromium and C1 had ~22,000 reads/cell and ~22,000,000 reads/cell, respectively and scatter plot comparing the results from the two platforms showed reasonable correlation coefficient.

This study highlights the advantages and disadvantages of various scRNA-seq platforms. Chromium is superior to other platforms in its prominent process-ability regarding the number of cells, but due to the lack of sequencing depth per given cell, it can be difficult to infer the detailed mechanism of biological problem. In contrast to Chromium, C1 has advantage in its sequence depth per cell and flexible control of the sequence read depth; however, due to the lack of number of cells under investigation, it is difficult to study small population of cells. Our study shows that different single-cell platforms need to be chosen to fit its purpose depending on the research question.

SEX-STRATIFIED ANALYSIS OF OBSESSIVE-COMPULSIVE DISORDER REVEALS MINOR DIFFERENCES IN GENETIC ARCHITECTURE

Ekaterina A Khramtsova[1], Lea K Davis[2], Barbara E Stranger[1]

[1]University of Chicago, Department of Medicine, Chicago, IL, [2]Vanderbilt University, Vanderbilt Genetics Institute, Nashville, TN

Obsessive-compulsive disorder (OCD) demonstrates sexual dimorphism in age of onset and clinical presentation, especially in comorbidities, suggesting a possible sex difference in genetic architecture. Here, we present the first genome-wide assessment of sex-specific genetic architecture of OCD. First, we performed a sex-stratified meta-GWAS (N=9870, 1:1.4 male/female ratio) to identify specific autosomal and sex chromosome risk factors with different effects in sexes. There were no genome-wide significant associations in either sex. However, there were several suggestive associations in each sex not present in the opposite sex. Second, we assessed whether heterogeneous OCD risk alleles are involved in gene regulation to elucidate the biological mechanisms by which those variants may impact dimorphism. On a genome-wide level, heterogenous SNPs (including and excluding SNPs in the HLA region) were strongly enriched for immune expression quantitative trait loci ($p<0.001$). Third, we used heritability (h^2) analysis to test for evidence of variable liability threshold for OCD between sexes and to assess the proportion of overall OCD h^2 explained by the X chromosome. There were no differences between sexes in OCD h^2. The X chromosome contributed 6.7% to total h^2 which is not statistically different from expectation. The genetic correlation between sexes is high: GCTA GREML (1.0, se=0.27) and LDSC (1.04, se= 0.51, p=0.0405). Since the lower bounds of genetic correlation estimate could range from 0.49-0.73, we explored whether males and females demonstrate differential genetic correlations between OCD and other traits which may play a role in OCD development (e.g. brain volumes), show sexual dimorphisms (e.g. autism, Tourette's syndrome, etc.) or are known to be differential clinical symptoms in OCD (e.g. smoking, alcohol consumption, etc.). Although OCD demonstrated significant genetic correlation with many traits, there were no statistically significant differences between sexes when corrected for multiple testing. We observed that the genetic correlation between OCD and alcohol consumption in males was not significantly different from zero (Rg=0.35, SE=0.42, p=0.40) but was strongly negative in females (Rg=-0.91, SE=0.47, p=0.056). The difference is statistically significant (p=0.024), providing support for observed sex differences in clinical presentation of OCD. We identified minor differences in genetic architecture of OCD, and although the sex-stratified sample size is likely too small to identify variants with small effect, we have developed a pipeline for sex-stratified genetic analysis which will be applied to OCD and other sex-biased phenotypes as larger cohorts become available.

INTEGRATION OF GENETIC AND EPIGENETIC ALTERATIONS WITH TISSUE-SPECIFIC NETWORK REVEALS REGULATORY DRIVERS OF PROSTATE CANCER

Priyanka Dhingra[1,2], Alexander Martinez-Fundichely[1,2], Andre N Forbes[1,2], Eric M Liu[1,2], Deli Liu[2], Andrea Sboner[2,3,4], David S Rickman[3,4,5], Mark A Rubin[3,4,5], Ekta Khurana[1,2,4,5]

[1]Weill Cornell Medicine, Department of Physiology and Biophysics, New York, NY, [2]Weill Cornell Medicine, Institute for Computational Biomedicine, New York, NY, [3]Weill Cornell Medicine, Department of Pathology and Laboratory Medicine, New York, NY, [4]Weill Cornell Medicine, Englander Institute for Precision Medicine, New York, NY, [5]Weill Cornell Medicine, Meyer Cancer Center, New York, NY

Cancer genomes exhibit multiple genomic and epigenomic alterations in their coding and non-coding regions. Understanding the global effects of these alterations in tissue-specific regulatory network can provide useful insights about their role in dysregulating gene expression and transforming normal cells to tumorigenic state. We report a novel computational method to investigate the combined effects of single nucleotide variants (SNVs), structural variants (SVs) and DNA methylation changes in the tissue-specific regulatory machine. The method involves (a) construction of tissue-specific regulatory network using DNase I hypersensitive data, (b) identification of significantly mutated, rearranged and differentially methylated coding and non-coding regions and (c) interpretation of the impact of genetic and epigenetic alterations on the regulatory network. In particular, we integrated whole-genome sequencing, RNA-Seq and DNA methylation data from 521 primary prostate tumor samples. We observe SNVs, SVs and methylation changes tend to target different genes in prostate cancer (PCa). Our results indicate a stronger regulatory impact of SVs in PCa, as they affect more transcription factor (TF) hubs in comparison to SNVs and methylation changes. Moreover, we observe the altered expression of the TF hubs due to SVs is correlated with global DNA methylation changes. Consistent with this, functional validation using Enhanced Reduced Representation Bisulphite Sequencing showed global methylation changes caused by ERG overexpression, an oncogenic TF frequently overexpressed in PCa due to gene fusion. Thus, we find SVs impact more TF hubs in the tissue-specific regulatory network in PCa. A crosstalk between TF hub expression modulated by SVs and DNA methylation levels likely leads to differential expression of target genes. Overall, our work identified known TFs (ERG and TP53) and nominates novel TFs (ERF, CREB3L1 and POU2F2) as regulatory drivers of prostate tumorigenesis. The proposed computational method can be used to interpret the global regulatory effects of genetic and epigenetic alterations in diverse tumor types.

AUGMENTING SUBNETWORK INFERENCE WITH INFORMATION EXTRACTED FROM THE SCIENTIFIC LITERATURE.

Sid Kiblawi[1,2], Deborah Chasman[3], Amanda Henning[4], Eunju Park[5,6], Hoifung Poon[7], Michael Gould[4], Paul Ahlquist[5,6,8], Mark Craven[2,1]

[1]University of Wisconsin, Department of Computer Sciences, Madison, WI, [2]University of Wisconsin, Department of Biostatistics and Medical Informatics, Madison, WI, [3]University of Wisconsin, Wisconsin Institute of Discovery, Madison, WI, [4]University of Wisconsin, Department of Oncology, Madison, WI, [5]University of Wisconsin, Institute of Molecular Virology, Madison, WI, [6]University of Wisconsin, Howard Hughes Medical Institute, Madison, WI, [7]Microsoft Research, Redmond, WA, [8]Mordridge Institute for Research, Madison, WI

Motivation: Many biological studies commonly involve either (i) manipulating some aspect of a cell or its environment and then simultaneously measuring the effect on thousands of genes, or (ii) systematically manipulating each gene and then measuring the effect on some response of interest. A common challenge that arises in these studies is to explain how a set of genes identified as relevant in the given experiment are organized into a subnetwork that accounts for the response of interest. Subnetwork inference is typically dependent on the information available in publicly available databases, which suffer from incompleteness. A wealth of information resides within the free form text of scientific literature, such as information about genes relevant for certain concepts as well as interactions that occur between various biological elements. We contend that by exploiting this information, we can improve the explanatory power and accuracy of subnetwork inference in multiple applications.

Results: We show that we can use literature-extracted information to (i) augment the set of nodes identified as being relevant in a subnetwork inference task, (ii) augment the set of interactions used in the process, and (iii) support targeted browsing of a large inferred subnetwork by identifying nodes and edges that are closely related to concepts of interest. We use this approach to uncover the pathways involved in interactions between virus and host cell, and the pathways that involve a transcription factor associated with breast cancer.

CURATED GENOMIC-BASED 16S RIBOSOMAL RNA DATABASE AND PLATFORM FOR SEAMLESS UPDATING

Seok-Won Kim, Todd D Taylor

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

Data growth in DDBJ/EMBL-EBI/NCBI is rising exponentially due to the increase of novel bacteria isolation and metagenomic studies. To manipulate these data, the primary data, including its associated metadata and sequences, needs to be checked in the next update stage against its own secondary database. This may result in a bottleneck when updating such massive datasets. Here, we present a massive sequence tracking and management platform for solving this issue. We constructed a manually edited 16S ribosomal RNA (rRNA) gene database called GRD. In GRD, both the 5' regions and 3' regions, including the anti-SD sites, have been carefully checked and contaminating sequences have been removed. Because of this careful manual checking of the 16S rRNA sequences, our database can be considered the most reliable reference source for downstream analyses. In addition, we are including PCR-based sequences which are published in public primary databases. We developed this platform for continuous updating and maintenance of the sequences and taxonomy information in this database. In particular, recently changed taxonomic names are updated according to the NCBI Taxonomy database. As with the genomic-based sequences, we confirm all amplified sequences which have 5’ and 3’ regions by manual curation. Our platform can be applied not only for rRNA genes, but also for other marker genes.

SPATIAL AND TEMPORAL TRANSCRIPTIONAL LANDSCAPE DURING FLY DEVELOPMENT

Cecilia C Klein[1,2,3], Marina Ruiz-Romero[1,2,3], Sílvia Pérez-Lluch[1,2,3], Alessandra Breschi[1,2,3], Amaya Abad[1,2,3], Emilio Palumbo[1,2,3], Roderic Guigó[1,2,3]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Institut Hospital del Mar d'Investigacions Mediques (IMIM), Barcelona, Spain

During development most tissues undergo fast changes in order to develop into a functional organs. In this scenario regulation of gene expression turns to be essential to determine cell and tissue specificity. Although many studies aim to decipher tissue signature through the analysis of their transcriptome profiles most lack temporal information during tissue differentiation. In this study we track down the transcriptome of pre-determined cells throughout differentiation to identify the dynamic transcriptional profile from the antenna, eye, leg, genitalia and wing of *Drosophila melanogaster* development. We identified three main sets of genes: commonly expressed genes that change across time, tissue-specific genes and time-tissue specific genes. Our analyses suggest that although differences among tissues increase over time a conserved gene regulatory network is leading the differentiation process. At the early stages of differentiation, the genes contributing to tissue morphogenesis are related to cell cycle and proliferation while later to cuticle formation and neural development. A comparison of the splicing patterns shows that there are fewer differences in splicing when compared to gene expression. Nevertheless, the differences in isoform usage are mainly associated to genes known to play a role in neural fate and are found in late stage, suggesting that splicing may play a role during differentiation. Overall we observe that the transcriptome diverge as tissues become more specified. Finally to further characterize cell sub-populations in tissue development we analyzed the four compartments of wing primordia and identified genes that are essential to maintain organ structure and compartment formation.

METAGENOMIC INSIGHT OF RELATIONSHIPS BETWEEN CARBON AND NITROGEN METABOLISMS IN A TANNERY WASTEWATER TREATMENT PLANT BIOAUGMENTED WITH THE MICROBIAL CONSORTIUM BMS-1

Woo-Jun Sul[1], S. Aalfin Emmanuel [2], HoonJe Seong[1], Jae-Soo Chang[2], Sung-Cheol Koh[2]

[1]Chung-Ang University, Systems Biotechnology, Anseong, South Korea, [2]Korea Maritime and Ocean University, Environmental Engineering, Busan, South Korea

The metagenomic insight into nitrogen metabolisms such as denitrification, dissimilatory nitrate reduction to ammonia (DNRA) and ammonium assimilation has been elucidated through a metagenome sequencing of the microbial communities in the 5 different stages of treatment system of tannery wastewater. It has been assumed that an efficient removal of nitrogen and COD (and hence sludge reduction) occurs through a combined work of the inorganic nitrogen metabolism and the carbon metabolism of amino acid, fatty acid and organic acids which are likely dominant in the tannery wastewater. The goal of this study was to elucidate relationships between the inorganic nitrogen removal and the carbon source utilization in the metagenomic perspective. Metagenomic taxonomic and funcitonal analysis were profiled using HUMAnN2 (http://huttenhower.sph.harvard.edu/humann2). Shotgun metagenomic reads were mapped to 'ChocoPhlAn' pan-genome database and MetaPhlAn2 database for organism specific functional profiling. We then analyzed the gene families and pathways using the extended databases UniProt Reference Clusters (UniRef90, http:/www.uniprot.org) and MetaCycmetabolic pathway database (www.metacyc.org). During the functional analysis, 40,181 gene families and 196 pathways were revealed for the five stages. Of these, 32 metabolic pathways were involved in amino acid production whose pathways were dominantly found in the stage I and PA, and most of amino acid degradation pathways were also dominant in I, indicating that NH4+ could be mostly released in these stages. However, nosZ gene involved in reduction of N_2O to N_2 was highly dominant in the stage B where a significant removal of nitrogen and COD was observed while nrfA gene involved in DNRA was highly dominant in the stage PA. It was also revealed through a pathway correlation analysis that L−asparagine degradation is specifically linked to DNRA pathway. A linkage analysis between denitrification/DNRA and degradation of fatty acid and other organic acids will be also discussed in association with an efficient removal of nitrogen and COD (and hence sludge reduction) in the eco-friendly treatment of tannery wastewater. These metagenomic insights will contribute to a successful monitoring and operation of the eco-friendly tannery wastewater treatment system which is a highly important issue in the tannery business.

COMPLETE GENOME SEQUENCE OF *PAENIBACILLUSYONGINENSIS*, A NOVEL PLANT SYMBIONT THAT PROMOTES GROWTH VIA INDUCED SYSTEMIC RESISTANCE IN *PANAX GINSENG* MEYER AND *ORYZA SATIVA* JAPONICA UNDER SALINITY STRESS

Yeon-Ju Kim[1], Sung-Cheol Koh[2], Johan Sukweenadhi[1], Deok-Chun Yang[1]

[1]Kyung Hee University, Graduate School of Biotechnology, Suwon, South Korea, [2]Korea Maritime and Ocean University, Environmental Engineering, Busan, South Korea

The completed *Paenibacillus yonginensis* DCY84T genome consists of a single circular chromosome of 4,985,901 bp, with a GC content of 51.01%, which is similar to most *Paenibacillus* strains (45 to 54%) as reported previously. A total of 4,498 genes were predicted for the genome, including 4,233 coding sequences (94.1% of total genes) and 147 pseudo genes. "*Paenibacillus yonginensis*"DCY84T(=KCTC33428=JCM19885), which is a gram-positive rod-shaped bacterium displays plant growth promotion via induced systemic resistance of abiotic stresses. Accordingly, this study aimed to utilize PGPB to enhance the tolerance of the *Panax ginseng* and Rice plant against salt stress. The strain *Paenibacillus yonginensis* DCY84T was registered as one novel bacteria species and found to have IAA production, siderophores production, phosphate solubilization traits and antibacterial activity. In planta test using arabidopsis and rice showed the potential use of strain DCY84T to alleviate salinity stress. Expression of AtRSA1, AtVQ9 and AtWRKY8 which determining saline stress condition was higher on DCY84T treated arabidopsis. Strain DCY84T(=PGPB) can be used to prime ginseng seedlings and induced its salinity tolerance. Moreover, PGPB treatment can induce plant defense mechanism mediated by ABA signal under salinity stress. Also, it can promote the root hair formation,based on these results, strain DCY84T is suggested to be used as "plant strengthener" type of microbial inoculant. The application of this strain to improve other kind of crop/ plant growth is widely possible.

POPULATION AND PHYLOGENETICS OF BABOONS THROUGH THE LENS OF THE *ALU* MOBILOME

Cody J Steely[1], Vallmer E Jordan[1], Jerilyn A Walker[1], Thomas O Beckstrom[1], Cullen L McDaniel[1], Corey P St-Romain[1], Emily C Bennett[1], Arianna Robichaux[1], Brooke N Clement[1], The Baboon Genome Analysis Consortium[2], Clifford J Jolly[3], Jane Phillips-Conroy[4], Kim C Worley[2], Jeffrey Rogers[2], Mark A Batzer[1], Miriam K Konkel[1,5]

[1]Louisiana State University, Dept. of Biological Sciences, Baton Rouge, LA, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [3]New York University, Dept. of Anthropology, New York, NY, [4]Washington University, Dept. of Anatomy and Neurobiology, St. Louis, MO, [5]Clemson University, Dept. of Genetics & Biochemistry, Clemson, SC

The habitat of baboons (genus *Papio*) ranges across sub-Saharan Africa and the southern Arabian Peninsula. *Papio* baboons began to radiate ~2 million years ago and are now recognized as 6 distinct species: olive, yellow, Guinea, hamadryas, chacma, and kinda. However, the evolutionary relationships among baboons continue to be debated, likely in part because of hybrid zones. The divergence of baboon-macaque occurred at a comparable time to the human-chimpanzee-gorilla divergence, allowing for direct comparisons of mobile element evolution. Notably, analysis of the *Papio anubis* (olive baboon) [Panu_2.0] draft genome assembly supports a more rapid expansion of *Alu* elements compared to most other primates except the most recent rhesus macaque assembly [Mmul8.0.1]. To investigate the population and phylogenetic relationships within the genus Papio, we analyzed a panel of 79 baboons (representing all 6 species) using 494 polymorphic *Alu* insertions. To reduce ascertainment bias, we selected candidate insertion loci from the baboon reference genome assembly and from high-throughput sequencing data of all 6 baboon species. The unexpectedly high homoplasy rate due to *Alu* insertion polymorphism across different baboon species suggests extensive incomplete lineage sorting, recent speciation, and/or ongoing introgressive hybridization. Depending on marker selection, our data reflect previously reported evolutionary discrepancies. Analysis of older *Alu* insertion data supports an early north- (olive, Guinea, and hamadryas) south (yellow, chacma, and kinda baboon) split. Inclusion of all polymorphic *Alu* loci suggests a more complicated evolutionary history, in agreement with continuous introgression. Our Structure analyses clearly distinguish the 6 baboon species with yellow baboons from the Southwest primate center showing varying degrees of admixture with olive baboons. Furthermore, 2 distinct population clusters each are supported within kinda and yellow baboons. In summary, our analyses support the complex relationships among baboons, and provide evidence for rapid mobile element expansion possibly linked to interspecies hybridization.

THE DISTRIBUTION OF FITNESS EFFECTS OF NEW MUTATIONS IN CODING AND NON-CODING DNA IN HUMAN POPULATIONS

Athanasios Kousathanas, Lluis Quintana-Murci

Pasteur Institute, Department of Genomes and Genetics, Unit of Human Evolutionary Genetics, Paris, France

Characterizing the distribution of fitness effects of new mutations (DFE) is of fundamental importance to understand how deleterious variation is maintained in human populations. However, previous estimates of the DFE in humans have been obtained only for a limited set of populations (i.e., African Americans and European Americans), using low sample sizes (i.e., 10-20 individuals), and focusing mostly on coding mutations. Here, we analyzed the large dataset of low-coverage whole genome sequences from the 1000 Genomes project, and employed methods that infer an unbiased site frequency spectrum (SFS) from low-coverage data by analyzing the sequencing reads directly. We inferred the DFE for nonsynonymous sites and other functional classes of noncoding DNA, such as UTRs and promoter regions, by jointly fitting a 3-epoch demographic model and several DFE models to the SFSs of a presumably neutral site class (synonymous or intronic) and the focal site class. We first observed that the use of large sample sizes allows to estimate the parameters of the DFE, such as the mean effect of a new mutation, with much higher precision. In doing so, we detected significant differences in the DFE between different site classes and between different human populations, these differences being particularly pronounced for strongly deleterious mutations. We found that by utilising intronic sites as a neutral standard, we can detect a significant fraction of new mutations (10-30%) being deleterious in UTRs and that 5’UTRs experience significantly weaker purifying selection than the 3’UTRs. More generally, our results not only confirmed previous reports that selection is less efficient in non-African than in African populations but also identified a strong positive correlation between effective population size (Ne) and the average strength of purifying selection when using coding and non-coding DNA. Finally, using simulations that consider the inferred demographic models and selection parameters, we show that the detected difference in selection strength is not expected to have substantially affected the mean number of deleterious alleles per individual in the short time elapsed since human populations started to diverge.

IMPROVING STRINGTIE TRANSCRIPTOME ASSEMBLY USING SUPER-READS

Samuel Kovaka[1], Aleksey Zimin[1,2], Michael Schatz[1,3], Mihaela Pertea[1]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]University of Maryland, Institute for Physical Sciences and Technology, College Park, MD, [3]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Super-reads, first used for genome assembly in the MaSuRCA assembler, are generated using a k-mer index by extending short reads as long as unique k-mers exist on either end. Super-reads were proposed for use in transcriptome assembly in the first release of StringTie, where they improved assembly in two ways: simplifying splice-site chains by spanning multiple exons, and reducing the number of multi-mapping reads during read alignment. To simplify coverage estimates and to reduce the likelihood of super-reads not reflecting true transcript structures, super-reads were only used to fill in the region between paired-end reads, turning many paired-end reads into long single-ended fragments. Despite this limitation, the shortened super-reads still improved transcript assembly and quantification. Here we present a package for the creation of full-length super-reads which can be used for input to StringTie. This package is based on the MaSuRCA assembler, with error correction parameters adjusted to better suit transcriptome assembly. Using simulated data we found that more than 99.5% of super-reads preserve alternate splice-sites and do not produce chimeric transcripts, and the few exceptions receive low coverage and therefore have little effect on the final assembly. We designed an expectation-maximization algorithm to distribute coverage from reads assigned to multiple super-reads, and altered StringTie to read these coverage estimates from SAM/BAM file tags. We ran StringTie on multiple simulated human RNA-seq datasets using full-length super-reads and only short reads, and found that the super-read assemblies have higher sensitivity and precision than the short read assemblies, as well as better correlations of abundance estimates. We also tested on real human and plant RNA-seq datasets from multiple sources and found similar sensitivity and precision improvements for known genes. This super-read package will be released as part of the next version of StringTie.

DEBROWSER: INTERACTIVE DIFFERENTIAL EXPRESSION ANALYSIS TOOL

Alper Kucukural[1,2], Berk Sarioz[1], Melissa J Moore[3], Manuel Garber[1,2]

[1]Bioinformatics Core, University of Massachusetts Medical School, Worcester, MA, [2]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA, [3]Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA

cDNA sequencing (RNA-Seq) has rapidly become the standard method for expression analysis. RNA-Seq data posed many computationally and algorithmically challenges. Today there exist a large number of methods to deal with specific problems but very few end-to-end solutions that take data from its raw form onto its final analysis. The heterogeneity of software implementations makes it difficult for those with little or no computational expertise to analyze RNA-Seq data. Data visualization in particular is especially difficult without knowledge of statistical software packages or programming. To remedy the void in powerful, user-friendly tools for RNA-Seq data visualization we created DEBrowser, a web application for interactive exploration of differential gene expression data. DEBrowser implements standard approaches for quality control as well as for visualization of the results.

The key features of DEBrowser allow a) the user to make complex comparisons by building custom tables that include results of any supported DE algorithms on any combination of replicates of interest b) batch effect detection and outlier identification by All-to-All, IQR, denstiy, and PCA plots. Hence even users have no programming and statistical knowl edge to easily identify any unexpected behavior arising from technical or biological problems. c) allow the users to graphically select from any of the genome wide visualiztions (Scatter, Heatmap, Scatter, Volcano and MA plots) any subset of genes for further in-depth exploration such as ontology, pathway and disease analysis.

DEBrowser is designed as a bioconductor package that can be installed to R or RStudio under the GNU General Public License. The open-source code can be downloaded from https://github.com/UMMS-Biocore/DEBrowser. The documentation and user manual can be obtained from http://debrowser.readthedocs.org.

NOVEL EVIDENCE OF COMPLEX PATTERNS OF GENE FLOW IN CHIMPANZEES AND BONOBOS

Martin Kuhlwilm[1], So Jung Han[1], Marc de Manuel[1], Tomas Marques-Bonet[1,2,3]

[1]CSIC-Universitat Pompeu Fabra, Institut de Biologia Evolutiva, Barcelona, Spain, [2]National Centre for Genomic Analysis–Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain, [3]Institucio Catalana de Recerca i Estudis Avançats, (ICREA), Barcelona, Spain

Our closest living relatives, chimpanzees and bonobos, have a highly dynamic demographic history including episodes of ancestral interbreeding among them. We have presented multiple lines of evidence for a signature of gene flow from bonobos into the ancestors of non-western chimpanzees, possibly during extended periods of time up to 550 thousand years ago. This suggests that admixture appears to have been widespread during hominid evolution. However, it seems possible that additional events of gene flow have taken place, in particular, admixture from lineages outside the known *Pan* clade. Similar observations of "ghost introgression" have been made in Denisovans.

Here, we analyzed the high-coverage whole genomes of 69 wild-born chimpanzees and bonobos from ten countries in Africa to investigate the genetic traces of such archaic gene flow events by applying the S* statistic. Our findings suggest additional gene flow events into chimpanzee populations. While possibly 1% of central chimpanzee genomes might carry bonobo haplotypes, we find that chimpanzees might carry a smaller amount of haplotypes falling outside the common *Pan* clade. We conduct a detailed investigation of demographic scenarios causing these observations, and characterize these haplotypes regarding their age and their divergence patterns and impact on possible functional consequences and adaptation, exposing for a first time partial genome information from an extinct ape lineage.

PASSENGER MUTATIONS IN >2500 CANCER GENOMES: OVERALL BURDENING & SELECTIVE EFFECTS

Sushant Kumar[1,2], Patrick Mcgillivray[4], William Meyerson[1,2,4], Leonidas Salichos[1,2], Shantao Li[1], Arif Harmanci[1,2], Xiaotong Li[1], Mark Gerstein[1,2,3]

[1]Yale University, Program in Computational Biology and Bioinformatics, New Haven, CT, [2]Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT, [3]Yale University, Department of Computer Science, New Haven, CT, [4]Yale University, Yale School of Medicine, New Haven, CT

The classic view of cancer claims that very few driver variants play an important role in tumor progression, whereas majority of variants often labeled as passengers are neutral and inconsequential for tumorigenesis. A growing number of studies challenge this classic view and have proposed that passengers can be further classified as latent driver, mini driver and deleterious passengers based on their fitness affect on cancer cell. While latent driver and mini driver are argued to be under weak positive selection and promote tumor growth, deleterious passengers undergo negative selection and are thought to inhibit tumor progression. In this work we utilize comprehensive variant data set from pan-cancer analysis of whole genome (PCAWG) project to test the underlying hypothesis of the classic view that passenger variants are under neutral selection and play no role in cancer progression. We find various evidences suggesting passenger variants undergo weak positive and negative selection. We observe a multimodal functional impact distribution with significant enrichment of variants with medium impact score in certain cancer types. Furthermore, these intermediate impact passenger variants have different signature profile and are enriched among essential genes. In addition, we provide further evidences based on variant allele frequencies, co-mutation frequencies and clinical data to contradict the classic view that all passenger variants are neutral. In brief this study further reiterates the need to extend the canonical dichotomy of passenger and driver as proposed in the classic view.

ELUCIDATING COMPLEX GENOMIC REGIONS THROUGH DE NOVO ASSEMBLY OF 10X LINKED-READS

Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, David B Jaffe

10x Genomics, Computational Biology, Pleasanton, CA

Knowledge of the complete diploid genome sequence for an organism will enable more accurate predictions and descriptions of its biology. Recently, we have demonstrated that high quality diploid genome sequences can be obtained through the assembly of short reads that are organized into barcoded groups encoding long-range genomic context. Such data can be generated at low cost using the 10x Genomics Linked-Read technology from a single short-read library created from 1 ng of high molecular weight DNA.

Here we exhibit a new computational approach that allows the assembly of complex genomic regions using 10x Linked-Reads. Utilizing a K=48 de Bruijn graph assembly as a starting point, we construct sets of barcodes whose reads belong to a genomic locus and create a "local assembly" for each set. These local assemblies and the initial graph assembly are merged based on shared sequence while utilizing the barcodes to keep distant loci separate. To further enhance our sensitivity we create “micro assemblies” generated by gluing reads using overlaps as short as 16 bases while maintaining specificity through the barcodes. This combination of techniques produces a genome assembly with excellent contiguity even over regions where the read coverage is low. For a typical human genome, we perform approximately 100K local assemblies and about 1K micro assemblies and generate a final merged assembly with multi-megabase scaffold and phase block sizes.

We generated several diploid assemblies using this approach, including a synthetic diploid assembly constructed by mixing 1:1 DNA from two hydatidiform moles. We combined two publicly available high-coverage PacBio assemblies (GCA_001297185.1, GCA_000983465.1) for each mole to form a “diploid reference”. We assay the quality of our assembly using the 4-megabase MHC region, which appears in a single scaffold that is phased into two haplotypes and is 98 % complete relative to the PacBio reference. We demonstrate similarly high quality assemblies with multi-megabase phased scaffolds in additional human samples including a family trio. The low cost, simplicity and power of our approach suggest its applicability to a wide range of genomes.

SCALABLE, EXTENSIBLE AND MODULAR RNA-SEQ EXPRESSION PROFILING USING THE OPEN-SCIENCE KBASE CYBERINFRASTRUCTURE

Vivek Kumar[1], Sunita Kumari[1], Jim Thomason[1], Doreen Ware[1,2], Shinjae Yoo[3], Sean McCorkle[3], Priya Ranjan[4], Nomi Harris[5], Chris Henry[6], Robert Cottingham[4], Adam Arkin[5]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]USDA-ARS NEA, Ithaca, NY, [3]Brookhaven National Laboratory, Upton, NY, [4]Oak Ridge National Laboratory, Knoxville, TN, [5]Lawrence Berkeley National Laboratory, Berkeley, CA, [6]Argonne National Laboratory, Argonne, IL

The U.S Department of Energy Systems Biology Knowledgebase (KBase, http://kbase.us) provides an open, web-accessible system for systems biology research focused on microbes, plants and their communities. It provides a range of integrated biological data types and a variety of analysis tools that include modeling, simulation methods and visualizations.

KBase includes a rich set of computational methods and curated datasets for gene expression analysis based on RNA-seq such as a selection of preprocessed high-quality reference genomes and a wide variety of algorithms for short-read mapping, identification of splice junctions, differential expression analysis, and visualization. More specifically, KBase supports both the original and new Tuxedo tool suites, including Bowtie2, TopHat2, HISAT2, Cufflinks, StringTie, Cuffdiff, CummeRbund and Ballgown.

KBase also provides services that are compatible with gene expression profiles for downstream analysis including clustering of expression profiles based on different algorithms. RNA-seq analysis services are available from within a highly interactive, Jupyter-based user interface that supports the creation of dynamic workflow documents called Narratives that enable experimental and computational biologists to work together to share and publish data, approaches, workflows, and conclusions, leading to transparent and reproducible computational experiments. Within a Narrative, short reads from an RNA-seq experiment can be uploaded into KBase to perform gene expression analysis and the results can be shared such that the research can be reproduced and extended by others in the KBase community. We demonstrate the utility of Narratives by performing a point-and-click, yet detailed analysis of public RNA-seq data from several species and tissue types, including experiments in A. thaliana, P. trichocarpa, and E. coli.

KBase analysis apps include genome assembly and annotation, metabolic modeling, phylogenetics, comparative genomics and microbial community analysis, and using the KBase Software Development Kit, developers in the KBase user community are able to add new apps to the system.

GERMLINE COPY NUMBER VARIATIONS ASSOCIATED WITH BREAST CANCER SUSCEPTIBILITY AND THE ROLE OF EMBEDDED miRNA GENES

M Kumaran[1], R Hubaux[2], W Lam[2], P Krishnan[1], Y Yasui[1], C Cass[1], S Damaraju[1]

[1]University of Alberta, Lab Med & Pathol, Edmonton, Canada, [2]BC Cancer Research Centre, Dept of Integrative Oncology, Vancouver, Canada

Breast cancer (BC) is one of the most commonly diagnosed cancers among women. Copy Number Variants (CNVs) from germline DNA and their role in genetic predisposition to disease risk of many cancer types, including BC is still evolving. Embedded within the germline CNVs are cis-regulatory elements, coding and non-coding genes (e.g., miRNAs). We address potential roles of miRNAs originating from the germline CNVs contributing to BC risk.
Hypothesis: Germline CNVs harboring miRNA genes show expression in breast tumors, affect regulation of downstream target genes and associated pathways contributing to the phenotype of BC.
Objectives: 1) To identify miRNA regions within germline CNVs (miRNA-CNVs) associated with BC; 2) To demonstrate breast tumor specific expression of CNV-miRNAs; 3) To identify miRNA regulated target genes (mRNAs) in tumor tissues and 4) To identify the pathways affected by the target genes.
Methods: We utilized buffy coat DNA (germline) from 422 BC cases and 348 healthy controls to conduct a CNV based GWAS using Affymetrix Human SNP 6 array genotype data (discovery dataset). Partek Genomics Suite was used for copy number analysis. Associated CNV regions were annotated with miRNA genes based on miRBase ver. 20. These CNV-miRNA regions were mapped to TCGA germline and BC primary tumor copy number segments in 84 samples (validation dataset). NGS data (Illumina HiSeq platform) was available for TCGA samples (miRNA and mRNA expressions). Expressed miRNAs with > 5 read counts in 50% of the samples were retained for analysis. miRNA target genes were predicted using Target Scan. Expressed target genes (mRNAs) were correlated with miRNA ($r < -0.3$, $p\text{-value} < 0.05$). Correlated target genes were used for Ingenuity Pathway Analysis (IPA) to identify associated pathways.
Results/conclusions: We have identified 21 germline CNV regions encompassing 64 miRNAs, to be associated with BC risk in the discovery dataset as well as in validation set. 52 miRNAs were expressed in the TCGA BC tumor samples and 10 miRNAs passed the filtering criteria. Of these nine miRNAs mapped to CNVs in the locus, 14q32.31 and one miRNA at 19p13.3. In total, we identified 156 correlated target genes potentially regulated by miRNAs and enriched in pathways such as mismatch repair, protein ubiquitination, BRCA1, hereditary breast cancer signaling, and telomerase extension. This study supports the concept that germline CNVs are functional, and exerts their actions through the expressed miRNAs and down-stream regulation of mRNA targets in BC tissues.

TRANSCRIPTIONAL PROFILES ARE ACCURATE PREDICTORS OF PAN-CANCERS

Fabien C Lamaze, Mawusse Agbessi, Jean-Christophe Grenier, Vanessa Bruat, Philip Awadalla

Ontario Institute for Cancer Research, Informatics and Bio-computing, Toronto, Canada

Extensive genetic and phenotypic variations exist within and between cancers representing a challenge for the discovery of biomarkers and personalised oncology. While some loci are now “canonical” or common, and despite some evidence for functional convergence in cancer development (e.g. metabolic pathways and cell cycle), common biomarkers identification have largely been unexplored to define similar phenotypes in lesions that are yet genetically divergent.

Paired tumour and normal tissues biopsied from over 1700 patients, including whole transcriptomes and genomes, across 31 different cancers from the Pan Cancer and TCGA consortium were used to define a common set of transcripts and pathways across all of these cancers. The value of these tools is that we can now predict with high confidence the status of any given sample (tumor vs. normal), regardless of cancer type and tissue. First we used a 1000x resampling strategy to identify a common set of 322 deregulated genes. These genes are largely functionally enriched for cell cycle, signalling and recombination pathways or processes. We used a machine learning strategy over 20,000 RNAseq data set from normal and tumor biopsies. The entire data set was split into training (60%) and comparable prediction (40%) sets. We were able to further refine our gene-set to only 162 genes which segregate with high predictive accuracy (98%), specificity (99%) and sensitivity (93%) tumors from normal tissue biopsies across all cancers. For each cancer, accuracy, sensitivity and specificity also remained high - all above 97%, 93% and 100% (1st quartile), respectively. We were also able to show that this gene set was highly predictive beyond humans, correctly identifying tumor vs. normal tissues in other mammalian cancers including Tasmanian devils.

The selective differential expression of these biomarkers in tumor compared with paired normal tissues in mammals suggests that they resume common differentiation pathways during “mammalian” carcinogenesis. This stable expression in tumor can be used in un-described, or rare cancers, or when the differentiation markers are unreliable. Finally, these biomarkers could be used for increasing the accuracy of the early detection diagnosis and stratification of tumor types in the clinic.

QUANTIFYING GENETIC REGULATORY VARIATION AFFECTING EACH GENE IN HUMAN POPULATIONS

Pejman Mohammadi[1,2], Stephane E Castel[1,2], Tuuli Lappalainen[1,2]

[1]New York Genome Center, -, New York, NY, [2]Columbia University, Department of Systems Biology, New York, NY

Estimating the amount of genetic variation affecting genes has become an essential metric for understanding the spectrum of genome variation, selective constraint on different genes, and prioritizing rare disease-causing mutations. While several existing approaches quantify the amount of genetic variation in coding regions, estimating the amount of regulatory variation has remained challenging.

In this study, we introduce a novel method, Analysis of Expression Variation (ANEVA) to quantify the total amount of cis genetic variation affecting the expression of each gene in different tissues. This method uses allele-specific expression data from a population sample, leveraging on its unique sensitivity to capture the net effect of all cis-regulatory variants across the frequency spectrum. We used a probabilistic model to describe the data as net outcome of haplotypes harboring a set of unobserved regulatory variants, and also estimate the total expression variation that includes environmental and noise effects. Our simulations demonstrated that the model is robust and accurate across a spectrum of allele frequencies, regulatory complexity, and gene expression levels.

Applying the method to the GTEx project v6 data with 8555 RNA-seq samples from 455 individuals, 12,000 genes had sufficient data for ANEVA analysis. We first benchmarked ANEVA's estimates of genetic regulatory variation with traditional heritability estimates, with highly consistent results from the two unrelated approaches corroborating the accuracy of ANEVA and indicating that it can be used as an orthogonal approach for estimating expression heritability without using genotype data. The amount of genetic regulatory variation per gene was correlated with coding constraint and haploinsufficiency, indicating that it captures of selective constraint on genes. With multi-tissue RNA-seq data enabling tissue-specific analysis, we showed that genes expressed in the brain show depletion of regulatory variation. Genes associated to rare diseases such as autism and congenital heart disease were depleted of genetic regulatory variation, while GWAS genes for most diseases are not, with signs of tissue-specific patterns consistent with disease etiology. Finally, in order to prioritize potential disease-causing regulatory variants from patient RNA-sequencing data, we demonstrate how ANEVA analysis identifies genes where patients harbor extreme genetic regulatory effects. Altogether, our work provides novel insights into tissue-specific architecture and selective constraint on cis-regulatory genetic variation, and enables rigorous analysis of rare regulatory variants in human disease diseases.

EVOLUTION OF THE OLFACTORY GENOME IN WILD RODENTS.

Jean-Marc Lassance, Hopi E Hoekstra

Harvard University/HHMI, Organismic and Evolutionary Biology;Molecular and Cell Biology, Cambridge, MA

Deer mice (genus *Peromyscus*) have been the subject of ecological study by natural historians from the early 1900's and recently have become a model for the study of the genetic basis of adaptation – from morphological, to physiological, and behavioral traits. The availability of high-quality *Peromyscus* genome assemblies will play an instrumental role in furthering the integration of ecological, evolutionary, and genomic information. To this end, we sequenced the genome of two sister species, *Peromyscus maniculatus* and *P. polionotus*, to carry out comparative genomic analyses of these wild rodents, which diverged about 0.6 MYA and differ from each other in a number of traits. The assemblies yielded 2.5Gb with an N50 scaffold size of 13 Mb. We could anchor 97% of the *de novo* assembled bases into 23 autosomes plus X chromosome using our high-density genetic linkage maps. The assignment of chromosomal locations to the majority of the assembled bases resulted in very high quality reference genomes. Here, we focus on the identification and characterization of the chemosensory gene repertoires, which we identified by comparative-genomic techniques together with deep RNA sequencing of the olfactory mucosa and vomeronasal organ, in both male and female adult mice from both species. The peripheral sensory organs of olfaction, and the sensory receptor neurons they house, play a major role in modulating specific behavioral responses as they are the first element in the processing of olfactory information by the nervous system. In total, we identified ~1200 olfactory receptors (ORs) and 240 vomeronasal receptors (VRs), respectively, in the *Peromyscus* genome. While the OR repertoires are similar to that of *Mus*, the VR gene family appears reduced in size in *Peromyscus* (359 in *Mus*). Perhaps more interestingly, although clades previously identified in *Mus* have representatives in Peromyscus, our phylogenetic reconstructions indicate that most gene duplications took place after the split between the *Mus* and *Peromyscus*; several clades show sign of lineage-specific expansion or contraction, revealing an extremely dynamic evolution for these genes which are responsible for the detection of pheromones and predator cues. Also, we found that the transcriptome of the two organs differ only minimally between males and females of both species, supporting the idea that the sexual dimorphism seen in olfactory-mediated behaviors is not driven by alteration at the peripheral level but rather have their origin in divergence at higher processing centers. By contrast, the two species show a high degree of differential regulation in the transcriptome of both organs. Taken together, these differences in the olfactory repertoires likely reflect the unique habitats as well as the highly divergent social and mating behavior of these two species of deer mice, and provide insights into the evolution of the olfactory system of mammals.

STIX: A POPULATION-SCALE STRUCTURAL VARIANT INDEX FOR RAPID ALLELE FREQUENCY INTERROGATION

Ryan M Layer, Brent S Pedersen, Aaron R Quinlan

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

When considering the effect of an SNV in a Mendelian disorder, the frequency of that allele in ExAC weighs heavily on our assessment of pathogenicity. Any allele that is not in ExAC is given priority because we trust that if it is not seen among 60K exomes it is likely rare. Unfortunately, there is no equivalent dataset for structural variants (SV). This gap is due to both the absence of a similarly sized SV dataset (the largest is 2,504 genomes from the 1000 Genomes Project (1KG)), and the complexity inherent to identifying and representing SVs. Unlike SNVs, for which we can list the frequency of every allele, the extent of possible SV size and type combinations (about equal to the number of grains of sand on Earth) make it intractable to list the support for every possible variant. Because of this, projects like 1KG publish only a subset of high-quality calls. Filtering to high confidence SV calls suffers from an unknown false negative rate, making it difficult to draw conclusions about the prevalence of unreported SVs. What is needed is a method that can provide a full accounting of SV occurrence that also scales to the forthcoming deep SV data from TOPMED and the Centers for Common Disease Genetics (CCDG).

Here we propose the structural variant index STIX, which can search every discordant alignment (paired-end and split) across thousands of samples. For a given SV, STIX quickly reports a per-sample count of all concurring evidence. From these counts we can, for example, conclude that an SV with high-level evidence is a polymorphism. By representing the raw signal, we can avoid many false negatives, with the knowledge that STIX results can be filtered based on experimental needs. As a proof of concept, we indexed 2,504 genomes from 1KG and quantified the frequency of 15,555 deletions reported by TCGA in 20 minutes. We find that 2% of variants appear to be polymorphisms in 1KG, and 1% occur in over 10% of samples. Half the common SVs were false negatives in the 1KG call set.

STIX retains all of the alternate evidence for thousands of samples, making it useful for other analysis such as large-scale SV genotyping. In conjunction with a compact representation of the reference evidence, STIX can quickly genotype SVs across all indexed samples. This will be vital to sequencing projects such as TOPMED and the CCDG, which sequence cohorts in batches. For each batch, variants are genotyped across existing samples and existing variants across new samples. We demonstrate the scalability of this approach by indexing and genotyping a cohort of 100 50X genomes. STIX also eases the identification of SV hotspots and deserts and empowers population scale variant detection.

GENOME-WIDE QUANTIFICATION AND PREDICTION OF DNA METHYLATION-DEPENDENT REGULATORY ACTIVITY

Amanda J Lea[1], Christopher M Vockely[2,3], Christina Del Carpio[6], Luis B Barreiro[4], Timothy E Reddy[2,3], Jenny Tung[5,6,7]

[1]Duke University, Biology, Durham, NC, [2]Duke University, Genomic and Computational Biology, Durham, NC, [3]Duke University, Computational Biology and Bioinformatics, Durham, NC, [4]University of Montreal, Sainte-Justine Hospital Research Centre, Montreal, Canada, [5]National Museums of Kenya, Institute of Primate Research, Nairobi, Kenya, [6]Duke University, Evolutionary Anthropology, Durham, NC, [7]Duke University, Population Research Institute, Durham, NC

DNA methylation plays a central role in development, disease, and the response to environmental conditions. However, when and how changes in DNA methylation alter gene regulation remains poorly understood, in part because tests of this relationship rely on low-throughput or *in vitro* methods. We developed a new tool, ‘mSTARR-seq,’ that overcomes these limitations by integrating high-throughput reporter assays with manipulations of CpG methylation. To demonstrate our approach, we cloned sheared and *MspI*-digested DNA from GM12878 cells into CpG-free mSTARR-seq vectors, designed so that fragments with regulatory activity would self-transcribe. We methylated introduced fragments and transfected into the K562 cell line (n=6 unmethylated and 6 methylated replicates). By measuring transcripts derived from mSTARR-seq vectors, we tested for DNA methylation-dependent activity in 18% and 52% of enhancers and promoters, respectively. 10% of all assayed regions had significant regulatory activity (n=24945 of 277896 regions, FDR<10%). As expected, these regions were highly enriched for K562 enhancers and promoters, and depleted for heterochromatin, insulators, and repressed regions (all $p<^{10-15}$). Using these data, we investigated the causal relationship between DNA methylation and regulatory activity. DNA methylation suppresses activity in most cases: of 2143 regions with methylation-dependent activity, 88% are more active when unmethylated. Using random forests, we predicted methylation-suppressed regulatory regions with 77% accuracy. These regions are more likely to occur in K562 active promoters near highly expressed genes, and to have higher CpG densities and lower endogenous methylation levels than regions unaffected by methylation (all $p<10^{-15}$). In contrast, regulatory elements that are *more* active when methylated often contain binding sites for the Homeobox, bZIP, and bHLH TF subfamilies, which are known to bind methylated DNA *in vitro*. Finally, by taking advantage of genetic differences between GM12878 and the human reference genome, we identified 4 cases in which a novel CpG site created unexpected methylation-dependent activity. Together, our results demonstrate the utility of mSTARR-seq for identifying regions where differential methylation ‘matters’ for gene regulation.

THE COMPLEMENT CONTROL PROTEIN CSMD1 IS A BROAD REGULATOR OF ADULT DEVELOPMENT AND FERTILITY IN BOTH SEXES

Arthur S Lee[1], Jannette Rusch[1], Abul Usmani[1], Wendy S Wong[2], Ni Huang[1], Ronald E Worthington[3], Xiaobo Wu[4], John P Atkinson[4], Rex A Hess[5], Donald F Conrad[1]

[1]Washington University, Department of Genetics, St. Louis, MO, [2]Inova Health Systems, Translational Medicine Institute, Falls Church, VA, [3]University of Southern Illinois, Pharmaceutical Sciences, Edwardsville, IL, [4]Washington University, Medicine, St. Louis, MO, [5]University of Illinois, Comparative Biosciences, Urbana-Champaign, IL

Although male and female infertility have historically been classified as clinically distinct disease entities, much of the molecular apparatus governing proper gametogenesis is shared between both sexes. Our search for loci that modulate gonadal function in men and women led us to the gene *CSMD1* which encodes a largely unstudied complement control protein. We identified *CSMD1* in a rare-variant GWAS for human gonadal function in the two sexes consisting of nearly 15,000 total males and females. Rare deletions over *CSMD1* are associated with early idiopathic menopause in women (OR = 16; nominal P = 4.0 x 10^{-4}; genome-wide P = 0.015) and spermatogenic impairment in men (OR = 3.3; nominal P = 6.5 x 10^{-3}). Rare deleterious SNVs in *CSMD1* are also associated with earlier age at menopause in 1,500 women (P < 5 x 10^{-3}).

CSMD1 directly overlaps the region of greatest human-chimp nucleotide divergence of all the autosomes (3.2% over 1Mb). Whole-genome sequencing of over 1,300 trios showed a massive increase in germline *de novo* mutation rate over *CSMD1*. This increase is not well explained by primary sequence context, and we implicate structural variation as a potential cis-modifier of mutation rate in this region.

We performed a battery of physiological and histological assays to dissect this pathology in *Csmd1* mutant mice. In male mice, *Csmd1* deficiency leads to near-complete histologic destruction of the testes as early as 34 days of age, and *Csmd1* genotype segregates significantly with the extent of testis degeneration (P = 7.69 x 10^{-3}). In female mice, *Csmd1* deficient ovaries are significantly smaller than wildtype ovaries (P = 0.022). Remarkably, half of Csmd1 knockout females lose their entire litter within 2 days of delivery (P = 7.93 x 10^{-7}), which we show is likely due to defects in mammary gland development. Finally, neither male nor female *Csmd1/C3* double knockout mice produced any offspring over 5 months of breeding. Double knockout females suffer from severe inflammatory and necrotic changes of the ovary and oviduct, as well as failure of ovulation. Prior work has implicated complement proteins in synaptic pruning of the neurons. Our results raise the possibility that *CSMD1* and complement play an important role in cellular regulation across diverse tissues and cell types.

A GRAPH REMAPPING FRAMEWORK FOR IN SILICO ADJUDICATION OF SNVs, INDELs, AND STRUCTURAL GENETIC VARIANTS FROM GENETIC SEQUENCING DATA

Dillon H Lee, Alistair N Ward, Gabor T Marth

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

Several state-of-the-art, easy to use tools are available both for short-variant detection (e.g. GATK, FREEBAYES), and structural variant (SV) detection (e.g. LUMPY, MANTRA, DELLY), but 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 result in call sets that require expert manual review to reduce false positive detection rate.

Our algorithm, GRAPHITE (https://github.com/dillonl/graphite) requires as input a collection of variant calls, made by one or more short-variant or SV detection tools. Typically, this starting set is high sensitivity (i.e. inclusive), but low specificity (i.e. have a high false discovery rate). We then apply a novel “variant adjudication” procedure to discard false positives, while keeping true positive calls. This is accomplished by constructing a graph from these variants (the Variant Graph) representing allelic variants as graph branches, in addition to the branches formed by the current, linear genome reference sequence. Using a graph mapping algorithm (GSSW, 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 retain 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 highly specific callset that also maintains the high sensitivity of the inclusive starting callset constructed by multiple primary variant calling methods. Because the graph construction and mapping approach works for most types of SVs in addition to all short variants, variants of all different types can be integrated in a single step.

Here we present the application of this method for cross-validating structural variants calls from Pacific Biosciences data by remapping deep Illumina WGS read sets to Variant Graphs constructed using the candidate Pacific Biosciences variants, as part of the Human Genome Structural Variation Consortium (HGSVC) data analysis project. We also present GRAPHITE’s application to improving the accuracy of allele frequency measurement in tumor sequencing data, which is essential for the accurate reconstruction of subclonal evolution in longitudinal tumor samples.

A NOVEL MACHINE LEARNING FRAMEWORK FOR INTEGRATING MULTIPLE DATA SOURCES TO IDENTIFY ROBUST BIOMARKERS FOR 160 ANTI-CANCER DRUGS IN LEUKEMIA

Su-In Lee[1,2], Safiye Celik[1], Benjamin A Logsdon[3], Scott M Lundberg[1], Timothy J Martins[4], Vivian M Oehler[4], Elihu H Estey[4], Chris P Miller[4], Sylvia Chien[4], Akanksha Saxena[4], Anthony C Blau[4], Pamela S Becker[4]

[1]University of Washington, Computer Science & Engineering, Seattle, WA, [2]University of Washington, Genome Sciences, Seattle, WA, [3]Sage Bionetworks, Seattle, WA, [4]University of Washington, Division of Hematology, Department of Medicine, Seattle, WA

The growing availability of gene expression data and in-vitro drug sensitivity data from cancer cell lines enabled us to identify biomarkers for drugs by finding *statistical associations* between genes and drugs. However, there are many reasons why drug response is associated with gene expression that do not reflect the underlying biological mechanism of drug response. False positive associations could be driven by biological confounders (e.g., disease sub-types), experimental confounders (sample ascertainment), or even technical confounders (batch effects) and often do not replicate in another data set. The *high-dimensionality* of data increases the multiple hypothesis testing burden and the chance of false positive.

We present a unique resource to identify biomarkers for 160 anti-cancer drugs in acute myeloid leukemia (AML), which consists of new data from 42 primary patient samples and a novel computational method. The key idea of the method is to reduce the dimensionality of data based on pre-existing data. We developed a machine learning method to incorporate a set of *driver features* that characterize each gene's potential to drive AML: 1) mutation frequency from TCGA, 2) 'hubness' in a gene network inferred from publicly available expression datasets, 3) whether the gene is known to be a regulator based on gene annotation databases, 4) CNV from TCGA, and 5) methylation from TCGA. We name these MERGE features (Mutation, Expression hubs, known Regulators, Genomic CNV, and mEthylation). The MERGE algorithm models the marker potential of each gene (i.e., prior probability that the gene is a true biomarker) as a weighted combination of the gene's MERGE features and jointly learns these weights and the impact of the *marker potentials* of genes on the observed gene-drug associations.

We demonstrate that MERGE outperforms four previous methods including the DREAM challenge winning method based on the statistical robustness and biological relevance of the identified biomarkers, and drug response prediction. Finally, we identify SMARCA4 as a biomarker and driver of the increased sensitivity to topoisomerase II inhibitors, mitoxantrone and etoposide, in AML by showing that cell lines transduced to have high SMARCA4 expression show dramatically increased sensitivity to these agents.

A STUDY ON ADMIXTURE GRAPH FITTING.

Kalle Leppälä

Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

The phylogenetic history of populations can be approximated by admixture graphs, i.e., connected directed acyclic graphs with no indegrees more than two . Edge lengths represent genetic drift and nodes with indegree two represent admixture events. Such nodes are equipped with a proportion describing the amount of gene flow from the two parental nodes.
The graphs can be tested against observed f2, f3 or f4 -statistics, using for example the *admixturegraph* package [Kalle Leppälä, Svend V. Nielsen & Thomas Mailund; *admixturegraph: an R package for admixture graph manipulation and fitting*; Bioinformatics (in press)].

The number of different trees with labeled leaves already grows factorially with the number of leaves, and allowing additional admixture events increases the count evenfurther. Thus, treating the edge lengths and admixture proportions as variables and only considering the shape of the graph, determining the admixture graph best fitting the observed f-statistics is a non-trivial task. We show that there exists an upper bound on the number of admixture events as a function of the number of populations, after which all the admixture graphs fit the observed data at most as well as some simpler graph. This reduces the infinite problem to a finite one.

We also compare f2, f3 and f4 -statistics with simulations to learn which kind of data most reliably recovers the original graph as the best fitting admixture graph after adding more or less error on the observations.

CIRCULOMICS: COMPREHENSIVE DETERMINATION OF THE CIRCULAR COMPONENT OF EUKARYOTIC GENOMES

Massa J Shoura[1], Idan Gabdank[1], Loren Hansen[1], Stephen D Levene[2], Andrew Z Fire[1]

[1]Stanford University School of Medicine, Pathology, Stanford, CA, [2]University of Texas at Dallas, Bioengineering, Biological Sciences, and Physics, Richardson, TX

Understanding how linear chromatin is organized within the nucleus and how its 3D architecture influences gene regulation and evolution are major questions in cell biology. Despite spectacular progress in this field, we still know remarkably little about the fraction of the eukaryotic genomic DNA having a circular topology. Extrachromosomal-circular DNAs, or eccDNA, have been anecdotally associated with many human phenotypes and diseases, yet we still lack rigorous whole-genome studies focused on determining the distribution of eccDNAs and molecular mechanisms that underlie the connections between eccDNA repertoire and physiology. The absence of a rigorous approach to profile eccDNAs is a major hindrance in advancing the field. Towards obtaining a comprehensive understanding of the genome (both the linear and the circular components) and its regulatory machineries, we developed an unbiased whole-genome approach to profile eccDNAs. We successfully applied our approach to the whole organism, C. elegans and other systems. Our findings suggest a dynamic interplay between linear chromatin (sequence and accessibility) and the corresponding eccDNA distribution (surprisingly, even at multiple coding regions). Our study presents an emerging understanding of how genome organization, as a whole, interacts dynamically with cell fate and function.

This study highlights an understudied, mysterious fraction of eukaryotic genomes. It puts forth a model of genome dynamics that will be of general interest to scientists studying genome architecture, features, accessibility, and structural variation.

A NOVEL APPROACH TO DETECT VIROME-WIDE INTEGRATIONS IN HUMAN CANCER GENOME USING WHOLE-GENOME SEQUENCING

Dawei Li[1,2,3]

[1]University of Vermont, Department of Microbiology and Molecular Genetics, Burlington, VT, [2]University of Vermont, Department of Computer Science, Burlington, VT, [3]University of Vermont, Neuroscience, Behavior, and Health Initiative, Burlington, VT

The WHO estimates that 18% of the human cancer cases world-wide are caused by viruses. There are seven known human oncogenic viruses, such as HPV triggering cervical cancer and HBV causing liver carcinoma. In all cases, the virus integrates with numerous integration sites in the cancer genome. For other cancers with speculated viral etiology, the specific oncovirus(es) has not been determined. Whole-genome sequencing (WGS) has the potential to capture all viral integrations in the human genome. However, the existing WGS-based bioinformatics methods for viral integration identification only screen one or a set of viruses, preventing the discovery of novel oncovirus(es). Here, we are presenting the first approach to accurately identify novel viral integrations in the human genome on a virome-wide scale. This approach has been applied to a large number of tumors of multiple cancer types, and the identified viral integrations have been experimentally validated with an overall >95% accuracy. We compared our methods with the existing methods. In 88 liver tumors, we found all those HBV integrations detected by the existing methods and confirmed by Sanger sequencing. We also found a large number of novel HBV/HCV integrations, and many of the integration hotspots are located in oncogenes. We compared tumors with and without integrations, and found the *TERT* gene expression level was highly enhanced in the tumors with HBV integrations ($P < 10^{-9}$). We also screened RNA-seq data and identified many novel HBV-human fusion genes. We validated seven randomly-selected fusion transcripts, and all of them were confirmed by PCR/Sanger sequencing. We also screened 11 HIV positive lung cancers, and found two of them (18%) have HPV-16 or HPV-33 integrations. In conclusion, we demonstrate that WGS allows for identifying novel viral integrations, and our approach can be used to identify novel viral integrations and oncovirus(es) in tumors with speculated viral etiology.

TECHNICALLY CHALLENGING BUT MEDICALLY IMPORT SEQUENCE AND COPY-NUMBER VARIANTS IN 80,000 PATIENTS: IMPLICATIONS FOR LABORATORY TECHNOLOGIES AND VALIDATION.

Stephen Lincoln[1], Rebecca Truty[1], Justin Zook[2], Brian Shirts[3], Matthew Ferber[4], Catherine Huang[5], Russell Garlick[5], Swaroop Aradhya[1], Mark Salit[2,6], Robert Nussbaum[1,7]

[1]Invitae, San Francisco, CA, [2]National Institute of Standards and Technology, Gaithersburg, MD, [3]University of Washington, Seattle, WA, [4]Mayo Clinic, Rochester, MN, [5]Seracare, Gaithersburg, MD, [6]Stanford University, Palo Alto, CA, [7]University of California, San Francisco, CA

Many medically important genes are located in or contain technically challenging sequence contexts, and complex but highly relevant mutations are well-known. The impact of these facts on diagnostic yield and on technologies for routine clinical genome/exome sequencing has not yet been thoroughly described.

We examined over 80,000 patients clinically tested for physician-specified genes underlying a hereditary cancer, cardiovascular, neurological or pediatric condition. Sensitive methods using NGS, long read single molecule sequencing, long range PCR, MLPA, and arrays were used to detect and orthogonally confirm a broad spectrum of variants in these patients.

Of the 12,489 pathogenic and potentially actionable true positives, approximately 10% of belong to a technically challenging class not easily addressed by short-read NGS. Approximately 3% are CNVs affecting only a single exon, 2% are either large indels or complex variants, and 5% lie in low-complexity, segmentally duplicated, or extreme-GC regions. This general observation was consistent across clinical areas, although specifics varied. Very large tri-nucleotide expansions and certain cytogenetic abnormalities were not included and would increase this total.

Such challenging variants are often under-represented or are entirely absent from published validation studies, despite their high prevalence. This may, in part, be due to difficulty obtaining positive controls. We developed synthetic controls which contain a diverse set of technically challenging mutations in the coding regions of commonly tested genes. These controls were evaluated using multiple NGS protocols and bioinformatics pipelines. We find that the synthetic variants mimic their endogenous counterparts, presenting the same technical challenges and artifacts.

In summary, technically challenging variants are a substantial fraction of medically relevant germline findings. To reduce false negatives, laboratories may implement methods that accurately detect these variants, and synthetic controls are one approach to validating such methods. The controls we developed may be of interest to other labs and will be available by the time of the meeting.

EXPERIMENTALLY VALIDATED COMPUTATIONAL METHODS TO INFER XCI ESCAPE LANDSCAPE IN POPULATION SCALE RNA-SEQ DATA

Renan Sauteraud[1], Jesica James[2], Laura Carrel[2], Dajiang Liu[1]

[1]Penn State University, Public Health Sciences, Hershey, PA, [2]Penn State University, Molecular Biology and Biochemistry, Hershey, PA

X chromosome inactivation (XCI) is an epigenetic mechanism that silences the gene expression from one copy of the X chromosome in females and hence balances the gene dosage between males and females. Up to 30% of the X-linked genes escape from XCI and get expressed from both Xs. The escape of XCI shows substantial inter-individual differences, where a given gene may escape XCI in a subset of the sample, and remain inactivated in others. The identification of XCI escape genes were hampered by the XCI mosaicism, where the assignment of active X (Xa) and inactive X (Xi) varies between cells. As a result of this biological complexity, despite being enriched in disease-relevant genes, the functional genomics and disease association for the X chromosome have often been understudied.

To fill in the methodology gap, we present a new empirical Bayes based method to infer XCI escape status from population scale bulk RNA-Seq data. The method first calculates XCI skewing in each sample using read ratios for a training set of genes known to be X inactivated. Genes with allelic expression ratios that significantly deviate from this XCI skewing estimate are likely to escape XCI. In order to improve power, we also borrow strength across samples by assuming realistic parametric distribution on the read depths and allelic expression levels and estimate the posterior probability for each gene to escape XCI.

To evaluate the proposed methods, we provided an experimental system for validating the inferred XCI genes using single cell derived clones. Specifically, we selected a mosaic LCL and isolated single-cell derived lines with complete non-random XCI of the maternal or paternal X. In these lines, the number of reads from one X directly reflects XCI escape without the need to infer XCI skewing. We compared the inferred XCI states from the original cell line and the observed XCI states in the single-cell clones. We showed that the type I errors and power were well controlled for the proposed methods.

Applying these methods, we analyzed the 272 female lymphoblast (LCL) samples from GEUVADIS. After rigorous quality control and the correction of reference bias, 670 genes are informative with >1 heterozygous well-expressed SNP. We identified 375 genes to be X inactivated, while 207 genes are classified as variable escape genes and bi-allelically expressed in >15% of the samples. The method was implemented in an R package and is readily available for use by the community.

FASTER GROWTH WITH SHORTER ANTIGENS EXPLAINS A VSG HIERARCHY DURING AFRICAN TRYPANOSOME INFECTIONS: A FEINT ATTACK BY PARASITES

Dianbo Liu[1,2], Luca Albergante[1,3], David Horn[1], Tim Newman[1]

[1]School of Life Sciences, University of Dundee, Dundee, UK, Computational Biology, Dundee, United Kingdom, [2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [3]U900, Institut Curie, 75005, France, Paris, France

The parasitic African trypanosome, Trypanosoma brucei, evades the adaptive host immune response by a process of antigenic variation that involves the clonal switching of variant surface glycoproteins (VSGs). The VSGs that periodically come to dominate in vivo display a hierarchy, but how this hierarchy arises is not well-understood. Combining publicly available genetic data with mathematical modelling, we report a VSG-length-dependent hierarchical timing of clonal VSG dominance in a mouse model, revealing an inverse correlation between VSG length and trypanosome growth-rate. Our analysis indicates that, among parasites switching to new VSGs, those expressing shorter VSGs preferentially accumulate to a detectable level that is sufficient to trigger an effective immune response. Subsequent elimination of faster-growing parasites then allows slower parasites with longer VSGs to accumulate. This interaction between the host and parasite is able by itself to explain the temporal distribution of VSGs observed in vivo. Thus, our findings reveal a length-dependent hierarchy that operates during T. brucei infection, representing a 'feint attack' diversion tactic utilised during infection by these persistent parasites to out-maneuver the host immune system.

TRANSLATIONAL DYNAMICS REVEALED BY RIBOSOME OCCUPANCY AND TIME-RESOLVED PROTEOMICS DURING CHEMOTHERAPEUTIC RESPONSE

Tzu-Yu Liu*[1], Hector H Huang*[2], Diamond Wheeler[2,4], Yichen Xu[3], James A Wells[4], Yun S Song+[1,5], Arun P Wiita+[2]

[1]University of Pennsylvania, Dept. of Mathematics and Dept. of Biology, Philadelphia, PA, [2]University of California, San Francisco, Dept. of Laboratory Medicine, San Francisco, CA, [3]University of California, San Francisco, Dept. of Urology, San Francisco, CA, [4]University of California, San Francisco, Dept. of Pharmaceutical Chemistry, San Francisco, CA, [5]University of California, Berkeley, Computer Science Division, Dept. of Statistics, and Dept. of Integrative Biology, Berkeley, CA

*authors contributed equally
+to whom correspondence should be addressed

Dynamic changes in the cancer proteome control tumor growth, proliferation, metastasis, and response to the therapy. Hence, it is important to understand the translational regulation in cancer cells during chemotherapeutic response. We studied multiple myeloma cells exposed to a low dose of bortezomib, designed to elicit a drug-induced stress response but not lead to widespread translational shutdown and cell death. Longitudinal mRNA-seq, ribosome profiling, and pulsed-stable isotope labeling (pSILAC) mass spectrometry-based proteomics were integrated to directly monitor the synthesis of new proteins and degradation of existing proteins across a time course. We modeled the dynamic changes with a system of differential equations and solved the equations using functional data analysis.

We directly compared measurements of translational efficiency from ribosome profiling and the estimates of translational rate parameter using our proposed quantitative model. Results confirmed that ribosome footprints reflected protein synthesis before the onset of bortezomib-mediated translational repression. Under conditions of translational inhibition, we found that pSILAC methods were able to directly detect global alterations of translation not identified by standard ribosome profiling approaches. We further demonstrated that our model, incorporating experimental protein synthetic and degradation rates, could predict protein-level dynamics in response to different levels of stress-induced translational inhibition. These findings underscore the utility of pSILAC proteomics as a complementary method in studies of translational regulation to ribosome profiling, particularly under conditions of cellular stress.

NICHE-ASSOCIATION AND RETENTION DYNAMICS IN THE HUMAN MICROBIOME

Jason Lloyd-Price[1,2], Anup Mahurkar[3], Gholamali Rahnavard[1,2], Jonathan Crabtree[3], Joshua Orvis[3], A. Brantley Hall[2], Arthur Brady[3], Heather H Creasy[3], Carrie McCracken[3], Michelle G Giglio[3], Daniel McDonald[4], Eric A Franzosa[1,2], Rob Knight[4,5], Owen White[3], Curtis Huttenhower[1,2]

[1]Harvard T. H. Chan School of Public Health, Biostatistics Department, Boston, MA, [2]The Broad Institute, Cambridge, MA, [3]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [4]University of California San Diego, Department of Pediatrics, La Jolla, CA, [5]University of California San Diego, Department of Computer Science & Engineering, La Jolla, CA

The trillions of microbes that live on or in us, our microbiome, have been identified as key to the establishment and maintenance of health. To understand these complex microbial ecologies, the NIH Human Microbiome Project (HMP) established the characteristics of the baseline "healthy" microbiome across body sites in a population-scale North American cohort. To date, this remains the largest body-wide survey of the human microbiome. Here, we present new findings and a dramatic expansion of shotgun metagenomes (now ~2,400 samples) from the HMP, termed HMP1-II, with three new characterizations of microbial health: functional specialization, strain distributions, and temporal dynamics. Species-level functional characterization showed different taxa contributing common functions at different body sites, as well as newly defining niche- and human-specific microbial metabolism and signaling. Strain identification revealed distinct subspecies clades for some species, some of which showed evidence of strain-level specialization to specific body sites. Temporal analysis using Gaussian processes decomposed microbial and functional variation by their characteristic rates of change within and among individuals. Species dynamics in the gut were most individualized, for example, while pathway abundances rarely were. By identifying features which structurally and dynamically maintain differences between individuals, we have achieved an improved understanding of personalized human microbiome strain-level structure and function.

IMPROVING DRAFT GENOMES OF HUMAN PATHOGENS FOR USE IN METAGENOMIC STUDIES

Jennifer Lu[1,2], Florian P Breitwieser[2], Steven L Salzberg[1,2,3,4]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins School of Medicine, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [3]Johns Hopkins University, Computer Science, Baltimore, MD, [4]Johns Hopkins School of Public Health, Biostatistics, Baltimore, MD

In theory, metagenomic sequencing of human patient samples has the potential for improving patient diagnoses and treatment. Sequencing of the infectious tissue or bodily fluids can provide insight into the bacterial, viral, and/or eukaryotic pathogens potentially causing the infection. As many patient symptoms are non-specific, the sequencing results can narrow the diagnosis and thereby improve the probability of effectively treating the patient [1]. However, while this has the ability to enhance the diagnostic process, the quality of the existing bacterial, viral, and eukaryotic pathogen genomes has yet to be sufficient for diagnostic purposes.

This project focuses on the many eukaryotic pathogen genomes that are still in their draft forms. These genomes contain significant contaminants (including human DNA) and low complexity sequences that cloud the sequencing results. As a result of human DNA contamination, the human DNA contained in any patient sample matches portions of the draft genomes and thereby produce false positives.

In order to investigate the extent of this contamination, we used Kraken to classify DNA reads from 20 brain samples against the eukaryotic pathogen genomes [2]. Most eukaryotic pathogens are unable to infect the sterile, contained human brain. Based on this prior knowledge, we categorized any eukaryotic pathogen that matched as a false positive and then analyzed the exact sequences causing the false positives. We have developed a new method based on this protocol that can automatically clean draft genomes and produce much-improved results when subsequently used in a metagenomics pipeline for diagnosis of infections.

[1] Salzberg SL, Breitwieser FP, Kumar A, Hao H, Burger P, Rodriguez FJ, Lim M, Quinones-Hinojosa A, Gallia GL, Tornheim JA, Melia MT, Sears CL, Pardo CA. 2016 Next-generation sequencing in neuropathologic diagnosis of infection of the nervous system. Neurol Neuroimmunol Neuroinflamm 3:4 Article e251
[2] Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15 Article R46

GENE-ENVIRONMENT INTERACTIONS, GENE EXPRESSION AND SPLICING IN 250 CELLULAR ENVIRONMENTS

F. Luca[1,4], A. L Richards[1], G. Moyrbrailean[1], A. Pai[2], D. Kurtz[1], C. Kalita[1], O. Davis[1], C. Harvey[1], A. Alazizi[1], D. Watza[1], Y. Sorokin[4], N. Hauff[4], X. Zhou[3], X. Wen[3], R. Pique-Regi[1,4]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]MIT, Department of Biology, Cambridge, MA, [3]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [4]Wayne State University, Department of Obstetrics and Gynecology, Detroit, MI

Most human traits result from a complex interaction between environmental exposures and an individual's genotypes. While genome-wide association studies (GWAS) have identified a large number of loci associated with complex traits variation, these loci are generally located in non-coding regions and explain a small portion of the variation. We hypothesized that an in depth characterization of the response to environmental perturbations can elucidate the genetic architecture and molecular mechanisms underlying complex trait variation. We exposed 5 cell types to 50 treatments that represent common environmental exposures. We used a high throughput approach to identify 89 environmental conditions with large gene expression changes and we deeply sequenced the RNA from each of these conditions in three individuals. We identified 32,838 genes differentially expressed (10% FDR) in any of the 89 environments and constructed a gene coexpression network comprised of 87 modules. Modules associated with >12 treatment conditions contained genes involved in N-linked glycosylation, cellular homeostasis, and response to endoplasmic reticulum stress, which may represent common response mechanisms across several treatments. Across all conditions, we identified 15,628 RNA processing event shifts (15% FDR) in 4,567 genes. We found that changes in gene expression are correlated with differences in splicing.The highest proportion of changes were retained introns (RI), and alternative first (AFE) and last (ALE) exons. Many of these changes occurred consistently in the same direction across conditions, indicating global regulation by trans factors. Accordingly, differential expression of transcription factors (TFs) was significantly associated with changes in AFEs genome-wide. Furthermore, using a generalized linear model, we identified TF binding sites that predict which genes have shifts in AFE usage in response to a specific treatment, suggesting that AFE usage is regulated through changes in TF binding. Using ATAC-seq, we demonstrated that AFEs preferred following selenium treatment show differential TF footprints for over 50 TF motifs. Finally, we identified 215 genes with GxE, of which 49% were associated with complex traits in GWAS. Furthermore, genes with a transcriptional response to environmental perturbations showed 7-fold higher odds of being found in GWAS. For example, caffeine GxE at PPP3CA, a hub gene for a caffeine-response module, may contribute to variation in blood pressure. Our results demonstrate that response to environmental perturbations is mediated by both gene expression and splicing changes. Mechanistic understanding of GxE in these processes is indispensable to thoroughly annotate genes and bridge epidemiological and genome-wide association studies.

INVESTIGATING INTRAGENOMIC CONFLICTS IN THE AMPLICONIC GENES IN HUMAN POPULATIONS

Elise A Lucotte, Kasper Munch, Moisès Coll Macià, Mikkel H Schierup

Aarhus University, Bioinformatic Research Center, Aarhus, Denmark

It has been known for decades that the sex chromosomes play a disproportionately large role in the formation of new species, notably through hybrid incompatibilities. However, the mechanisms in play remain mostly unknown. A growing hypothesis presents the emergence of selfish elements on sex chromosomes as a driver of speciation. Indeed, it would lead to an intragenomic conflict between the X and the Y and drive rapid divergence between isolated population. Interesting candidate regions for such conflict to arise are the Ampliconic Genes (AG): they are multicopy, enriched in testis-expressed genes, X or Y-linked and known to be subject to highly dynamic evolution. Because of their repetitive nature, they are largely understudied. In this study, we explored the evolution of the AG by investigating copy number variation (CNV) between human populations using the Simons Genome Diversity Project. We developed a method to assess CNVs using the read-depth on artificial X and Y chromosomes composed of one repetition of each AG. Our results indicate that there are differences in copy number between human populations for AG including genes expressed in testis. Moreover, the CNV for Y-linked AG cannot be explained by haplogroups, suggesting independent amplifications of these genes within human population.

LRSIM: A LINKED READS SIMULATOR GENERATING INSIGHTS FOR BETTER GENOME PARTITIONING

Ruibang Luo[1,2], Fritz J Sdelazeck[1], Charlotte A Darby[1], Stephen M Kelly[3], Michael C Schatz[1,2,4]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Johns Hopkins University School of Mecine, Center for Computational Biology, Baltimore, MD, [3]New York University School of Medicine, Center for Health Informatics and Bioinformatics, New York, NY, [4]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, New York, NY

Haplotype-resolved or phased genomes are essential for studying allele-specific regulation and expression, and other important genomic features. However, most of the genomes assembled to date are only a single haploid ‘consensus’ sequence with parental alleles merged arbitrarily. 10X Genomics recently invented a low-cost, labor-efficient library preparation protocol to obtain phased genomes. The protocol ligates adapters to reads which will be sequenced on an Illumina instrument; the adapters allow each read to be traced back to its progenitor molecule. The so-called linked reads, spanning tens to hundreds of kilobases, offer an alternative to long-read sequencing for de novo assembly, haplotype phasing and other applications. However, there is no available simulator, making it difficult to measure their capability or develop new informatics tools for linked reads.

Our analysis of 13 real human genome datasets of 10X Chromium linked reads revealed their characteristics of barcodes, molecules and partitions. The first “linked read” simulator we wrote, named LRSim, generates simulated linked reads by emulating the library preparation and sequencing process with highly customizable parameter settings. We compared the simulated results from LRSim to NA12878 to illustrate a high concordance between the simulated and real data. We concluded that from the phasing results of 6 simulated datasets with different mean molecule lengths and a real dataset of NA12878 that if constrained at a certain sequencing depth, the best molecule size to achieve the best phase block size needs to be meticulously chosen. This can be done by wet-lab experiments, but would be more efficient with a simulator in silico. We also performed experiments on 6 simulated A. thaliana datasets with a different number of partitions and demonstrated a substantial degradation in assembly performance with an improper number of partitions, which leads to insufficient cover-age per molecule. Finally, we concluded an appropriate sequencing depth needs to be chosen for different applications and species before sequencing to achieve the best performance out of linked-reads.

PARALLEL SEASONAL SELECTION ACROSS *DROSOPHILA MELANOGASTER* POPULATIONS

Heather E Machado[1], Alan O Bergland[2], Paul S Schmidt[3], Dmitri A Petrov[1]

[1]Stanford University, Biology, Stanford, CA, [2]University of Virginia, Biology, Charlottesville, VA, [3]University of Pennsylvania, Biology, Philadelphia, PA

Fluctuating selection can play a role in the maintenance of genetic variation. For species with several generations per year, the seasonal changes in climate and resource availability can result in fluctuating selection. We performed population genomic sequencing of 56 seasonal (spring/fall) samples collected from 19 North American and European *Drosophila melanogaster* populations. We find a significant enrichment of seasonally varying SNPs across populations. We find high concordance of allele frequency change across seasons and with latitude for SNPs that are both strongly seasonal and strongly latitudinal. However, we estimate that seasonal sites tend to be of relatively small effect size and that they are variable in their identity across populations, resulting in low predictability across populations and years. Nonetheless, we demonstrate that seasonal selection is a general phenomenon in *D. melanogaster*, implicating it in the maintenance of genetic variation.

CONTRIBUTION OF RNA DECAY IN MEDIATING CHANGES IN THE TRANSCRIPTOME IN CELLULAR STRESS

Sho Maekawa[1], Sumio Sugano[1], Nobuyoshi Akimitsu[2], Yutaka Suzuki[1]

[1]The University of Tokyo, Graduate School of Frontier Sciences, Kashiwa, Chiba, Japan, [2]The University of Tokyo, Isotope Science Center, Tokyo, Japan

The eventual RNA expression is determined by rates of RNA generation and RNA decay and the balance between the rates determine the expression levels. The recent large-scale genomics projects have predominantly focused on epigenomics as a means to understand the regulatory mechanisms behind gene expression; however, the understanding of RNA decay mediated regulation of gene expression is far from exhaustive. We have previously reported the extent of the role of RNA decay in controlling the eventual RNA expression and the effects of RNA decay factors in regulating those genes with shorter RNA half-lives.

In order to focus on the biological relevance behind RNA decay mediated RNA expression regulation, we focused on the response to hypoxia, a condition under low oxygen potential, in a human colorectal cancer cell line. We profiled RNA decay using BRIC-seq, a 5'-bromouridine based pulse chase assay, and we compared RNA decay against transcription and RNA expression to try and understand the mechanistic roles that both transcription and RNA decay play in mediating changes in RNA expression. We found that 2381 and 1037 genes had their RNA decay elongated, or shortened by two-fold, respectively. For those genes that showed elongated RNA decay, 61 genes had their RNA expression up-regulated by two-fold, and 341 genes showed up-regulation in RNA expression without changes in RNA decay. Interestingly, for those genes that are targeted by HIF-1, a key transcription factor that regulate of hypoxia response, we only observe 4 genes that changes their RNA decay, in comparison to 17 genes that had their RNA expression up-regulated, suggesting that the control of RNA decay is independent of HIF-1. Taken together, our approach in adding RNA decay information will be useful to understand the mechanisms of gene expression, which will be a useful add-on to the genomic datasets.

CHARACTERISATION OF THE PAN-GENOME OF *VITIS VINIFERA* USING NEXT GENERATION SEQUENCING

Gabriele Magris[1,2], Fabio Marroni[1,2], Michele Vidotto[1,2], Sara Pinosio[1,2], Gabriele Di Gaspero[2], Michele Morgante[1,2]

[1]Universita degli Studi di Udine, D4A, UDINE, Italy, [2]Institute of Applied Genomics, UDINE, Italy

Analyses of structural variations in plants have shown that a single genome might not reflect the complete genomic complement of a species, and therefore the concept of pan-genome has been proposed. The latter is composed of a Core Genome (CG), common to all individuals of a species, and a Dispensable Genome (DG), absent in at least one individual. DG appears to be largely the youngest and most dynamic component of the pan-genome. Smaller deletions and insertions, due to recent movement of transposable elements and larger variants referred to as Copy Number Variants (CNVs) contribute to high levels of structural variation. In plants, the dispensable fraction of the genome may be widely influenced by the very active transposable elements.

More than 50 grapevine varieties were re-sequenced at high coverage and, based on a variety of approaches, used for the detection of Single Nucleotide Polymorphisms (SNPs) and Structural Variants (SVs). SNP markers were used to explore the grapevine population structure and to assess the genetic relationships between individuals. In order to gain knowledge about the composition of the dispensable fraction, structural variants of different size and origin were investigated based on paired-end mapping information. We will describe the dispensable fraction of the grapevine pan-genome, its extent and composition, and its epigenetic effects. In addition, we will explore the mechanisms that are at the origin of the dispensable portion. Lastly, by mean of transcriptomic data we will discuss the effects that DG has over the gene expression. Gaining insights into the composition and function of the DG will contribute to understand the mechanisms that create genetic diversity and phenotypic variation.

EXPLORING LINEAGE-SPECIFIC ENHANCERS BY INTEGRATING ENHANCER TRANSCRIPTION, EPIGENOMIC FEATURES, SEQUENCE MOTIFS, AND TRANSCRIPTION FACTOR EXPRESSION

Venkat S Malladi[1], Anusha Nagari[1], Hector L Franco[1,2], W. Lee Kraus[1]

[1]University of Texas Southwestern Medical Center, Cecil H. and Ida Green Center for Reproductive Biology Sciences and Division of Basic Reproductive Biology Research, Dallas, TX, [2]University of North Carolina School of Medicine, Lineberger Cancer Center, Department of Genetics, Chapel Hill, NC

The identification of transcription factors (TF) driving the formation of active enhancers that regulate the expression of target genes remains an open problem. We have developed a computational framework that identifies cell type-specific enhancers and their cognate TFs by integrating multiple genomic assays that probe the transcriptomes (GRO-seq and RNA-seq) and epigenomes (ChIP-seq) of various samples. Our method, called Total Functional Score of Enhancer Elements (TFSEE), integrates the magnitude of enhancer transcription (GRO-seq), enrichment of marks associated with enhancers (H3K4me1 and H3K27ac ChIP-seq), TF mRNA expression levels (RNA-seq), and TF motif p-values (MEME). This method has allowed us to explore the enhancer landscape in different cell types that share common origins or are biologically related, including distinct molecular subtypes of breast cancer, and embryonic stem cells (ESCs) and their derived lineages. Using TFSEE, we have identified key breast cancer subtype-specific transcription factors that are bound at active enhancers and dictate gene expression patterns determining growth outcomes. To demonstrate the broader utility of our approach, we have used this algorithm to identify transcription factors during the differentiation of embryonic stem cells into pancreatic cells. Taken together our results show that TFSEE can be used to perform multilayer genomic data integration to uncover novel cell type-specific transcription factors that control lineage-specific enhancers.

This work is supported by grants from the Cancer Prevention and Research Institute of Texas (CPRIT) (RP110471-P1, RP160318) and the NIH/NIDDK (R01 DK058110), and support from the Cecil H. and Ida Green Center for Reproductive Biology Sciences Endowment, to W.L.K.

TREEDEX, THE TREE DATA EXPLORER: INTERACTIVE FRAMEWORK FOR VISUALIZATION AND ANALYSIS OF COMPARATIVE OMICS

Marco Mariotti[1], Toni Gabaldón[2], Vadim N Gladyshev[1]

[1]Brigham and Women's Hospital, Harvard Medical School, Division of Genetics, Department of Medicine, Boston, MA, [2]Centre for Genomic Regulation, Bioinformatics and Genomics Programme, Barcelona, Spain

Rapid technological advances in sequencing and mass spectrometry have provided various high-throughput "omics" techniques, which can supply comprehensive molecular profiles across thousands of traits at once (genes, transcripts, metabolites) at a reasonable cost. Omics quickly became standard tools in research, and are routinely applied in large scale studies comprising multiple factors; for example, characterizing the gene expression response to a certain treatment in different tissues (with factors in this case being genes, treatment, tissues). The magnitude of omics data presents theoretical and practical challenges, particularly when encompassing many factors. When multiple species are investigated at once, there's also a *phylogenetic factor* to consider. Since organisms are related by their phylogeny, a meaningful interpretation of the data requires the consideration of their tree, and evolutionary theory must be applied to account for the effect of shared history on observed correlations.

We present here a novel computational tool: Treedex, the Tree Data Explorer, is designed as an integrative platform oriented to analyses of all kinds of comparative data. Thanks to its interactive environment, Treedex allows any scientist to intuitively navigate through complex observations across multiple species, and relate at all times the data points with the underlying species tree.

Treedex offers diverse built-in methods for comparative analysis (e.g. phylogenetic generalized least squares (PGLS) for tree-corrected regression), as well as coevolution-based techniques for evolutionary inference (e.g. phylogenetic profiling). The user can easily assemble data processing routines using such components, or customize them on the fly for specific purposes. This results in efficient and transparent analysis pipelines, as suited for complex omics data. By applying a geometrical abstraction to the co-evolution problem, Treedex allows to detect interactions between traits at any level, linking gene sequence variants, RNA expression, metabolite levels, and other quantitative and qualitative phenotypes. Treedex also aims to provide explicit representations for concepts relevant to modern evolutionary genomics, such as evolutionary trajectories (i.e. 'feature paths'), and the 'spaces' of all possible sequences and phenotypes.

Treedex is freely available on all operating systems for non-commercial uses: https://github.com/marco-mariotti/treedex

DEMOGRAPHIC HISTORY IMPACTS POLYGENIC RISK ACROSS DIVERSE POPULATIONS

Alicia R Martin[1,2,3], Christopher R Gignoux[3], Raymond K Walters[1,2], Genevieve L Wojcik[3], Duncan Palmer[1,2], Benjamin M Neale[1,2], Simon Gravel[4], Mark J Daly[1,2], Carlos D Bustamante[3], Eimear E Kenny[5]

[1]Massachusetts General Hospital, Analytic & Translational Genetics Unit, Boston, MA, [2]Broad Institute, Medical and Population Genetics Unit, Cambridge, MA, [3]Stanford University, Genetics Department, Stanford, CA, [4]McGill University, Department of Human Genetics, Montreal, Canada, [5]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences, New York, NY

Over the past decade, genome-wide association studies (GWAS) have successfully identified thousands of disease loci, but the vast majority of this research has prioritized populations of European ancestry. A key question remains regarding the generalizability of findings from complex trait studies to other populations. Population genetic theory indicates that effects of linkage disequilibrium, allele frequencies, and genetic architecture will impact transferability; however, there is a dearth of empirical data to demonstrate this. We have examined transferability by using published summary statistics for several well-studied traits and diseases and identified directional inconsistencies in polygenic risk scores across populations. These biases indicate that less heritability is explained in non-European cohorts.

To gain deeper quantitative insights into GWAS transferability, we developed a complex trait coalescent-based simulation framework recapitulating demographic parameters in an Out-of-Africa model. We consider the effects of polygenicity, causal allele frequency divergence, and heritability. As expected, correlations between true and inferred risk are typically highest in the population from which summary statistics were derived. We demonstrate that scores inferred from European GWAS are unpredictably biased in other populations, even when choosing the same causal variants. To address these issues, we have implemented novel methods that incorporate linkage disequilibrium into the correction of effect size estimates, thereby improving cross-population transferability. Our work cautions that summarizing findings from large-scale GWAS may have limited portability to other populations using standard approaches, and highlights the need for the adoption of improved polygenic risk methods, as well as the inclusion of more diverse individuals in medical genomics.

IDENTIFYING CANCER CODING AND NON-CODING DRIVERS INVOLVED IN SOMATIC STRUCTURAL VARIATIONS

Alexander Martinez Fundichely, Ekta Khurana

Weill Cornell Medical College, Institute for Computational Biomedicine, New York, NY

Somatic copy-number alterations are known to play an important role in cancer development. Recent availability of thousands of cancer whole-genome sequences allows an assessment of the impact of the full variety of large genomic rearrangements on this heterogeneous disease. Here we describe a new computational approach with enhanced power to identify genes targeted by somatic genomic structural rearrangements that could potentially drive cancer growth. The method combines the effects of the entire profile of somatic structural variations (SSVs, including deletions, duplications, translocations, inversions and other complex events) to identify the functional elements under positive selection in tumor cells. By analyzing the combined effects of all SSVs affecting every chromosome, we develop a probabilistic method that estimates background affected rates for each gene's coding sequence and regulatory elements, such as promoters and enhancers. The outcome is a p-value ranked list of significantly rearranged genomic elements. In particular, prostate cancer is a high-incidence cancer and the landscape of prostate cancer genomes is dominated by SSVs compared to point mutations. Application of our method on 123 prostate cancer whole genomes reveals 10 regions are significantly rearranged on 9 different chromosomes. These regions span known drivers and novel driver candidates. Thus, our new algorithm can be used for identifying coding and non-coding drivers involved in somatic structural variations across diverse tumor types.

USING SINGLE-CELL RNA-SEQ FOR ASSESSING THE EFFECT OF COMMON GENETIC VARIANTS ON DIFFERENTIATING HUMAN iPSCs

Davis J McCarthy[1], Jose Garcia-Bernardo[2], Mariya Chhatriwala[2], Shradha Amatya[2], Marc Jan Bonder[1], Yasin Memari[2], Ian Streeter[1], The HipSci Consortium[1,2], Ludovic Vallier[2], Oliver Stegle[1]

[1]EMBL-EBI, Systems Genomics, Hinxton, United Kingdom, [2]Wellcome Trust Sanger Institute, Cellular Genetics, Hinxton, United Kingdom

Human induced pluripotent stem cell lines (iPSCs) hold great potential for regenerative medicine and as model systems for studying disease by differentiating iPSCs into specific cell types of interest. However, the impact of genetic variation on the differentiation process and cell states is poorly understood.

Here, we leverage a large panel of genetically diverse human iPSC lines from the Human Induced Pluripotent Stem Cells Initiative (HipSci; www.hipsci.org) to assess the effects of common genetic variants on gene expression heterogeneity and cell states during cell differentiation. We study iPSC differentiation towards definitive endoderm at three distinct time points, assaying index-FACS sorted cells using single-cell RNA-seq. We describe an initial dataset of 10,000 cells with linked dense genotype data from a total of 30 human donors.

The availability of up to hundreds of cells from the same individual enables genetic analyses using a rich set of new molecular traits that can be derived from single-cell profiles. First, exploiting the continuous differentiation trajectories reconstructed using pseudo temporal orderings, we mapped genetic effects on gene expression at unprecedented temporal resolution, revealing between 554 (iPS) and 639 (endoderm) cis eQTLs (FDR<1e-04), the majority of which are specific to the stage of differentiation. Second, we estimated gene expression dispersion for individual genes and donors to map cis effects on gene expression variance (variants associated with the variance phenotype for a gene), revealing up to 586 variance-associated QTLs that are distinct from abundance QTLs. Finally, using FACS measurement of cell-surface markers, pseudotemporal ordering of cells, and pathway states derived using gene sets, we modelled cell-state specific genetic effects on gene expression. We identify 57 interaction QTLs (iQTLs) with the stage of differentiation and cell pluripotency. Several of these interaction effects overlap with variance QTLs, suggesting cell-state specific effects as a major explanation for variance QTLs.

Together, our study establishes single-cell RNA-seq as a phenotyping approach in genetically diverse populations, revealing effects on expression variance and cell-state specific effects in differentiating iPSCs.

STOCHASTICITY PROMOTES THE EVOLUTION OF COOPERATION IN A MULTILEVEL MODEL OF THE SNOWDRIFT GAME, WITH APPLICATIONS TO UNDERSTANDING COLLECTIVE DEFENSE IN MYCOBACTERIUM TUBERCULOSIS

Brian McLoone[1], Wai-Tong Fan[2], Adam Pham[3], Tracy M Doyle[4,5], Smead Rory[6], Caitlin S Pepperell[4,5], Laurence Loewe[1,7]

[1]Wisconsin Institute for Discovery, Systems Biology, Madison, WI, [2]University of Wisconsin–Madison, Mathematics, Madison, WI, [3]University of Wisconsin–Madison, Philosophy, Madison, WI, [4]University of Wisconsin–Madison, Department of Medical Microbiology and Immunology, Madison, WI, [5]University of Wisconsin–Madison, Medicine (Infectious Diseases), Madison, WI, [6]Northeastern University, Philosophy and Religion, Boston, MA, [7]University of Wisconsin–Madison, Genetics, Madison, WI

We explore the evolutionary dynamics of cooperation in a metapopulation composed of groups whose members play different instantiations of the Snowdrift Game, a well-known game in evolutionary game theory. In our model, within each group individual-level birth-death events are governed by a Moran process. Within the metapopulation, group-level fissioning and extinction events are also governed by a Moran process. We show that the combination of within-group stochasticity and this two-level selection process promotes the evolution of cooperation in the metapopulation, even its fixation, for a wide range of parameter values. We further characterize a phase transition for the fixation and extinction probabilities of cooperation in a finite population, based off of the cost to benefit ratio of cooperating in the Snowdrift Game. We relate our results to the literature on multilevel selection, studies of microbial social evolution—in particular, work on collective defense in Mycobacterium tuberculosis, the proximate cause of TB disease—as well as work that explores the importance of considering stochasticity when modeling evolutionary dynamics.

MACHINE LEARNING APPROACHES TO MODEL REGULATORY ARCHITECTURES

Maria K Mejia-Guerra[1], Tao Zuo[1], Edward S Buckler[1,2]

[1]Institute for Genomic Diversity, Cornell University, Ithaca, NY,
[2]Agricultural Research Service, USDA/ARS, Ithaca, NY

The rewiring of regulatory networks from which phenotypic variation may arises is the result of the paired turnover of *trans* specificity and *cis* regulatory elements within enhancers and promoters. An emerging view of the regulatory architecture suggests that recognition of the *cis*-information by *trans* regulators involves a combined readout of **DNA structural features**, **DNA sequence motifs**, **motif spatial organization**, and the **underlying sequence** in which motifs are embedded, and its interplay with the chromatin context.

We developed a computational approach to identify patterns of regulatory words (**lexicons**) and words relationships (**grammar**), present in functional non-coding genomic regions, aimed to score the putative effect of non-coding variation on regulation of gene expression. Our approach consist in applying a combination of machine learning approaches to identify patterns of lexicons and grammar rules from functional genomic datasets available in Maize. We found that our model discriminate holdout regulatory regions from random genomic background controlling for GC% and chromosomal distribution with great accuracy, as evaluated with the area under the receiver operating characteristic curve (AUROC), for which 1 is the highest possible score for a perfect classification. The model reach an average of **0.93 (AUROC) for transcription factors binding peaks** derived from ChIP-seq data, **0.91 (AUROC) for histone marks** peaks derived from ChIP-seq data, **0.95 (AUROC) for enhancer regions** derived from MNA-seq data and **0.94 (AUROC) for core promoters** derived from experimentally characterized transcription start sites.

NEANDERTAL AND DENISOVAN DNA SEQUENCES FROM PLEISTOCENE SEDIMENT

Matthias Meyer, Viviane Slon, Svante Pääbo, Ancient Sediment Analysis Group

Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany

Because skeletal remains of ancient hominins are rare, it is often difficult to determine which hominin groups have occupied archaeological sites where artefacts and other traces of human activity are found. Using automated laboratory procedures we have screened a large number of sediment samples from caves across Eurasia for the presence of DNA. We show that sediments from Late and Middle Pleistocene cave sites, some older than 200,000 years, contain large amounts of DNA from mammals such cave bears, mammoths, and hyenas. We also recovered eight Neandertal mtDNA sequences, some nearly complete, from sediments at four sites, including one at which no skeletal remains of Neandertals have been found. In the East Gallery of Denisova Cave, Russia, we find Denisovan mtDNA near the bottom of the stratigraphy, suggesting that Denisovans were present earlier in this region than previously known.

DNA from sediments is thus a much more abundant source of ancient hominin DNA than bones. It opens possibilities to detect the presence of archaic and ancient modern human groups in regions and at sites where no skeletal remains are found. Targeted enrichment of mitochondrial and possibly nuclear DNA from sediments is likely to shed light on the presence and the genetic diversity of known as well as hitherto unknown hominin groups across space and time.

EGALUX: AN ULTRA-FAST AND ACCURATE ALLELIC READS COUNTER FOR DNA SEQUENCING

Zong Miao[1,2], Marcus Alvarez[1], Päivi Pajukanta[1,2,3], Arthur Ko[1,3]

[1]UCLA, Dept. of Human Genetics, Los Angeles, CA, [2]UCLA, Bioinformatics Interdepartmental Program, Los Angeles, CA, [3]UCLA, Molecular Biology Institute, Los Angeles, CA

Allele-specific interrogation of epigenomic sites would help elucidate biological mechanisms underlying complex traits. However, mapping bias is a major obstacle in the allele-specific transcription factor (TF) binding, histone modification, and chromosomal interaction analysis. Although aligning reads both to the reference genome and alternative genome can efficiently reduce the bias caused by SNPs, alignment to the two genomes takes twice as much time as the alignment to the reference genome alone, slowing down the allele-specific epigenomic analysis. Some currently available aligners, such as SNP-o-matic and GSNAP, integrate the person's SNP information with the reference genome to perform a SNP-tolerant alignment. However, these existing tools are relatively slow, which makes them not ideal for allele-specific alignment. To perform accurate and fast alignment at the individual's epigenomic sites, we developed a new method EGAlux that focuses on SNP-tolerant alignment of allelic reads at the epigenomic sites. Since these reads are only ~10% of for example the ChIP-seq data, we can save a significant amount of time by ignoring non-allele specific reads during the alignment. EGAlux utilizes personal SNP information to extract the sequence around the SNPs and form a dynamic reference, reflecting thus the particular individual's genome around the SNP sites. Our new aligner EGAlux is significantly faster when compared to existing alignment tools because it identifies allelic reads with the dynamic reference before the genome-wide alignment. In more detail, EGAlux is ~38X faster than GSNAP (2600 queries/s), and ~11X faster than Bowtie (8600 queries/s). Our data on a simulated CTCF ChIP-seq data set show that EGAlux is not only fast but also accurate since EGAlux shows the least mapping bias of the three tested tools, with or without introducing sequencing errors. These test results demonstrate that EGAlux has a great accuracy and ultra-high speed for allelic reads alignment, making it a promising new tool for allele-specific analysis of TFs, histone modification and chromosomal interaction sites.

IOBIO DEV KIT: RESOURCES FOR MAKING GENOMIC, REAL-TIME WEB APPLICATIONS AND SERVICES

Chase A Miller[1,2], Yi Qiao[1,2], Tony DiSera[1,2], Alistair Ward[1,2], Gabor T Marth[1,2]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, USTAR Center for Genetic Discovery;, Salt Lake City, UT

IOBIO (http://iobio.io) is an open-source web-based genomic platform enabling real-time analysis of large, remotely-stored, distributed datasets, with analysis algorithms operating as web servers on data-streams in standardized data formats. Several IOBIO apps have been previously released and can be found at http://iobio.io/applications.html. Here we introduce the IOBIO Dev Kit (https://developer.iobio.io), which provides libraries, examples, and documentation to build your own IOBIO apps and services. The dev kit consists of three main libraries: iobio.js is a client-side javascript library that streamlines creating and executing iobio commands, while hiding complicated web-socket connection code; iobio.viz is a d3-based, streaming genomic visualization library that consumes iobio data streams; and minion is a server library that wraps (with only a few lines of code) standard command line tools and converts them into iobio web services that can be mixed and matched with all other iobio web services. As an instructive example, we have rebuilt our bam.iobio.io app (http://bam.iobio.io) using dev kit components with all code found here: https://github.com/chmille4/bam.iobio.io. Using the IOBIO Dev Kit, developers will be able to quickly create new genomic analysis apps on the web.

LOSS-OF-FUNCTION GENE VARIANTS IN NEUROMODULATORY PATHWAYS AMONG A POPULATION OF FREE-RANGING RHESUS MACAQUES

Michael J Montague[1], Noah Snyder-Mackler[2], Seth Madlon-Kay[1], Lauren Brent[3], J H Skene[4], Julie Horvath[5,6], Michael L Platt[1]

[1]University of Pennsylvania, Department of Neuroscience, Philadelphia, PA, [2]Duke University, Department of Evolutionary Anthropology, Durham, NC, [3]University of Exeter, Centre for Research in Animal Behaviour, Exeter, United Kingdom, [4]Duke University, Duke Institute for Brain Sciences, Durham, NC, [5]North Carolina Central University, Biological & Biomedical Sciences, Durham, NC, [6]NC Museum of Natural Sciences, Genomics & Microbiology Research Laboratory, Raleigh, NC

Evidence suggests that individual variation in social behavior arises from a combination of genetic predispositions and individual experience, yet the underlying biological mechanisms remain poorly understood. To address this gap, we have sought to understand the genetic, developmental, and neurobiological contributions to social behavior in a large, free-ranging population of rhesus macaques (*Macaca mulatta*) with a known pedigree and detailed behavioral phenotypes. We hypothesized that variants in genes related to neurotransmitters and neuromodulatory pathways may be associated with behavioral variation in this socially complex species. For example, glutamate receptor interacting protein 1 (*GRIP1*) is neuronal scaffolding protein involved in stabilization of glutamate receptors at excitatory synapses, and studies of neuronal-specific loss-of-function mice resulted in increased rates of prosocial behavior. To this end, we generated whole genome sequences for 217 individuals and identified over nineteen million population-wide single nucleotide variants. This included 254,777 exonic variants, of which 37% were predicted to alter transcript splicing sites or translated protein sequences. Predicted amino acid changes were found in key genes in neuromodulatory pathways, including dopamine receptors, oxytocin and vasopressin receptors, serotonin transporters, and the mu-1 opioid receptor. Of the 2,022 highest-impact variants, seventeen were predicted to affect the strongest candidate genes for autism spectrum disorders, per the online database, SFARI Gene. One such variant, with a population allele frequency of 0.16, was predicted to eliminate the start codon of *GRIP1*. Here, we describe the social behavioral differences found among the seventeen heterozygous and eleven homozygous macaques with this variant, suggesting approaches for integrating natural loss-of-function mutations with long-term behavioral data.

BRAIN-SEQ OR HOW TO SEQUENCE THE ENTIRE BRAIN AT SINGLE NEURON RESOLUTION: TOWARD DECIPHERING THE GENEALOGY OF NEURONS USING SCRNA-SEQ, NCRNAS AND RNA MODIFICATIONS

Leonid L Moroz[1,2], Andrea B Kohn[2]

[1]University of Florida, Neuroscience, Gainesville, FL, [2]University of Florida, Whitney Laboratory, St. Augustine, FL

One of the major challenges confronting the biomedical sciences is to understand how the genes of an organism regulate its behaviors from individual cells to complex networks and the brain. How and Why neuron-specific activity of thousands of genes leads to precise and persisting changes in synaptic efficiency mediating distinct memory traces are also unknown. Ideally, the understanding of genome-wide events requires an unbiased analysis of the entire scale of gene regulatory pathways simultaneously in each brain's neuron as they learn and remember. Although such an analysis is yet impossible in humans, it can be successfully achieved today using the simpler nervous system of the mollusc Aplysia, where many neurons and connections have been identified.

Here, using modified drop-seq/10x genomics strategies, we isolated and sequenced virtually all individual cells (neurons, glia, sensory and novel cell types) composing the entire brain of Aplysia, with a resolution currently impossible to achieve elsewhere. Second, we re-sequenced major components of the memory-forming circuit during two forms of learning. Next, we sequenced several peripheral organs including salivary glands and respiratory system at the single cell resolution. Finally, we developed a metric to quantitate each brain's region transcriptional relationships. Combined, this massive parallel scRNA-seq, with unsupervised clustering, allowed us to classify both neurons and non-neuronal cells. Surprisingly, we found that neurons, in spite of similar appearance, are more transcriptionally diverse than other phenotypes. We discover ~149 neuronal superclasses distributed across four levels of complexity. Hundreds of novel lineage-specific neuronal markers and secretory molecules were cross-validated using in situ hybridization and inherently linked to neuronal genealogy. Even more surprising was a discovery of neuron-specific ncRNAs and RNA modifying machinery leading to highly predictive neuronal individuality with nearly 90% accuracy! The unbiased cell-specific genomic portrait of the entire brain clearly illustrates that most neurons are epitranscriptomically unique. ScRNA-seq also allowed us to show that individual neurons both learn and age differently. In summary, we identified subsets of evolutionary conserved and novel genes associated to (a) neuronal identity and (b) plasticity as well as (c) glial phenotypes and consequently reconstructed their complex genealogies.

EXPANDING AND IMPROVING GENCODE GENE ANNOTATION TO AID HUMAN CLINICAL GENETICS

Jonathan M Mudge[1], James Wright[2], Jyoti Choudhary[2], Toby Hunt[1], Irwin Jungreis[3], Adam Frankish[1]

[1]Wellcome Trust Sanger Institute, Computational Genomics, Hinxton, United Kingdom, [2]Wellcome Trust Sanger Institute, Department of Proteomics, Hinxton, United Kingdom, [3]Massachusetts Institute of Technology, Computational Biology, Cambridge, MA

Most efforts to interpret the human genome sequence utilize gene annotation, and yet reference annotation catalogs are currently 'unfinished'. This causes particular problems in human genetics, most obviously where annotation deficiencies prevent the identification or correct interpretation of disease-associated variants. Indeed, the majority of GWAS variants currently fall outside genic regions. Here, we discuss a series of methodologies aiming to improve the content and usability of the GENCODE 'geneset'. Firstly, modern transcriptomics datasets indicate that thousand of exons, transcripts and even entire genes are currently missing from GENCODE. As well as integrating such data into our annotation pipeline, we are also targeting hundreds of specific loci for long-read sequencing based on Capture-seq and RACE assays. However, if annotation is to be truly useful it must not only describe the structure of transcripts, but also provide information into their functionality. An obvious question asks which sequences are protein-coding, and we have identified hundreds of additional coding sequences – including entirely novel genes – via a bespoke workflow combing mass spectrometry and comparative annotation based on phyloCSF. Finally, it is clear that many disease-associated variants fall within regulatory sequences as opposed to exons. Datasets such as Hi-C can allow for putative connections to be made between promoter and enhancer regions, while CHiP-seq assays can indicate specific sites of transcription-factor binding. We will describe our efforts to integrate such information to create 'extended' gene models, and thus to expand the space within which variants of interest can be tied to their target genes.

On behalf of the GENCODE consortium.

NUANCED EMPIRICAL MODELING OF OBSERVED TRANSCRIPTIONAL VARIANCE USING ULTRA-EFFICIENT 3’-BIASED RNASEQ DATA

Swagatam Mukhopadhyay, Sagar Damle, Steven Kuntz, Christopher E Hart

Ionis Pharmaceuticals, Functional Genomics, Carlsbad, CA

Genome-wide transcriptional profiling of biological samples remains one of the most informative methods for revealing underlying molecular and cellular states. For a fraction of the cost of full-coverage RNAseq, optimized 3'-biased sequencing methods provide quantitative resolution for nearly all well-expressed genes across 4-5 orders of magnitude. Unlike the established and robust quantification methods for RNAseq, estimating and quantifying changes in gene-specific variability requires novel methodology for low-coverage 3’-biased sequencing data. By dissecting biological variability's contributions to sequencing-readout noise, we set a prior expectation and quantify the significance of every gene's differential expression. Inspired by previous work on mathematical modelling of transcription, we assume that the variability in gene expression at a single cell level is distributed as a Negative Binomial (NB). Sequencing-readout noise and biological variability only add to this intrinsic noise (mixture, Poisson, and/or NB contribution), allowing us to compute -- using a maximum likelihood estimator and tens of biological replicates -- the dispersion parameter for every gene. We test this method using simulation. Our initial results strongly support the dispersion parameter's gene-specificity across multiple biological replicates, indicating that variability signatures of gene-expression can be learned from the variability in estimated transcripts-per-million (TPM) across biological replicates. Characterization of noise in TPM reduces false discovery in differential expression analyses. Using a physical model allows rigorous computation of measures of statistical confidence. These methods are implemented and distributed within an open-source Python package.

COMPARATIVE DYNAMICS OF MICRORNAS DURING MOUSE AND HUMAN PRENATAL DEVELOPMENT

Rabi Murad[1,2], Alessandra Breschi[3], Weihua Zeng[1,2], Brian Williams[4], Mark Mackiewicz[5], Carrie Davis[6], Thomas Gingeras[6], Barbara Wold[4], Richard M Myers[5], Roderic Guigó[3], Ali Mortazavi[1,2]

[1]Developmental and Cell Biology, University of California Irvine, Irvine, CA, [2]Center for Complex Biological Systems, University of California Irvine, Irvine, CA, [3]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain, [4]Division of Biology, California Institute of Technology, Pasadena, CA, [5]HudsonAlpha Institute for Biotechnology, Huntsville, AL, [6]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

microRNAs (miRNAs) are a class of small non-coding RNA that are critical post-transcriptional regulators of gene expression. The ENCODE project profiled the expression of miRNAs in various tissues during embryonic development of mouse and human using multiple complementary sequencing and hybridization techniques. We detected 643 miRNAs in mouse as well as 388 miRNAs in human. We find multiple tissue and developmental stage specific miRNA expression profiles dominated by small number of miRNAs. Comparative analysis of conserved miRNAs reveals clustering of expression patterns by tissue types rather than species. We used matching messenger RNA-seq (mRNA-seq) and histone modification ChIP-seq datasets to improve the annotation of microRNA primary transcripts. We show that the expression levels of a subset of primary miRNA transcripts predict the expression of their corresponding mature miRNAs. Our data provides the most comprehensive microRNA resource for mouse and human embryonic development as well as a comprehensive list of the mouse microRNAs that can be reliably measured by mRNA-seq of their primary transcripts.

PASSENGER PIGEON GENOMES REVEAL THE COST OF NATURAL SELECTION FOR A LARGE POPULATION

Gemma G Murray*[1], André E Soares*[1], Beth Shapiro[1], The Passenger Pigeon Genome Working Group[1,2]

[1]University of California Santa Cruz, Ecology & Evolutionary Biology, Santa Cruz, CA, [2]University of California Santa Cruz, Biomolecular Engineering, Santa Cruz, CA

*These authors contributed equally to this work.

The passenger pigeon was once the most abundant bird species in North America, but went extinct following a period of intensive commercial harvest in the late 19th century. The apparent absence of geographic structure in the passenger pigeon population prior to its decline suggests that in addition to its very large census population size this species had a large effective population size. Theory predicts that species with large effective population sizes experience a greater efficacy of natural selection. However, selection is also predicted to reduce the effective population size at linked sites, and when recombination rates are low, these linked regions are likely to be larger. Through comparative analyses of high-coverage whole genome assemblies of four passenger pigeons and two band-tailed pigeons, the passenger pigeon's closest living relative, we investigate the joint impact of a large census population size and the highly variable recombination landscape of the avian genome on the evolution of passenger pigeons. We find that passenger pigeons had a much lower effective population size than their census population size predicts, and that this was driven by natural selection rather than geographic structure or population size change. We find that passenger pigeons experienced a greater efficacy of natural selection than band-tailed pigeons, and that this and its impact on their effective population size were both modulated by differences in the recombination rate across their genome. This study provides clear evidence of an effect of the variable recombination landscape of the avian genome on effective population size and the efficacy of selection, and demonstrates how a larger population size can both increase the rate of adaptive evolution and the variation in the effective population size within a genome.

CHARACTERIZATION OF CHROMOSOME 21 rDNA REPEAT SEQUENCE AND VARIATION VIA TAR CLONING AND LONG-READ SEQUENCING

Jung-Hyun Kim*[1], Alexander Dilthey*[2], Ramaiah Nagaraja*[3], Hee-Sheung Lee[1], Sergey Koren[2], Dawood Dudekula[3], William Wood[3], Svetlana Shabalina[4], Koichi Utani[1], David Schlessinger[3], Adam Phillippy[2], Vladimir Larionov[1]

[1]National Cancer Institue, Developmental Therapeutics Branch, Bethesda, MD, [2]National Human Genome Research Institute, Computational and Statistical Genomics Branch, Bethesda, MD, [3]National Institute on Aging, Laboratory of Genetics and Genomics, Baltimore, MD, [4]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD

*Equal contribution

Human cells contain several hundred ribosomal DNA (rDNA) genes clustered in nucleolar organizer regions (NORs) on the short arms of five acrocentric chromosomes. Each NOR is a tandemly repeated unit of 43 kb transcribed into 13.3 kb 45S, encoding 18S, 5.8S and 28S rRNAs, and a 29.7 kb intergenic spacer (IGS). However, little is known about their precise structure and variation.

Prior attempts to characterize NOR sequences have failed due to their highly repetitive structure, and these regions are missing from the human reference genome. We targeted chromosome 21 NOR sequences using transformation-associated recombination in yeast, starting from a rodent/human hybrid cell line. After conversion to BACs they were sequenced on PacBio and Illumina platforms, enabling assembly. Structures were validated using an Oxford Nanopore protocol capable of ~90 kb of single-read sequence.

Thirteen clones, (0.4-fold coverage,~0.8 Mb) of the chr. 21 NOR, revealed variants of standard rDNA reference. We identified 216 variant alleles in the 45S region, 90 in mature rRNAs sequences, as well as 7 inversion breakpoints in the IGS resulting in palindromic structures.

The 90 sequence variations, compared to predicted 18S and 28S rRNA 2D structures, showed most located in species-specific expansion segments and rather than universal functional cores of 18S and 28S rRNA. Therefore, these variants are less likely to influence the catalytic core function of ribosome. However, variations in RNA 2D structure of the 5’ ETS would likely affect structural stability, consistent with a possible role in fine-tuning rRNA transcription.

In addition, a predicted candidate gene was found in the IGS for a subset of clones. The candidate gene, IGS breakpoints, and 61% of variant alleles were also seen in independent whole-genome PacBio and Illumina data. Thus, we have constructed a new ~45 kb rDNA reference sequence that has improved support from whole-genome data. This also provides an approach to full analysis of NORs and reagents for the study of human NOR function.

WHOLE GENOME SEQUENCING OF 175 MONGOLIANS UNCOVERS POPULATION-SPECIFIC GENETIC ARCHITECTURE AND PROVIDES AN INSIGHT INTO DISPERSAL OF EAST ASIANS

Narisu Narisu[1], Haihua Bai[2], Xiaosen Guo[3,6], Tianming Lan[3], Qizhu Wu[2], Yanping Xing[4], Yong Zhang[3], Stephen R Bond[1], Yanru Zhang[4], Chris R Gignoux[5], Huanming Yang[3], Lawrence C Brody[1], Jun Wang[3], Karsten Kristiansen[6,3], Francis S Collins[1], Burenbatu Burenbatu[2], Huanmin Zhou[4], Ye Yin[3]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [2]., Inner Mongolia University for the Nationalities, Inner Mongolia, China, [3]., BGI-Shenzhen, Shenzhen, China, [4]College of Life Science, Inner Mongolia Agricultural University, Hohhot, China, [5]Department of Genetics, Stanford University, Stanford, CA, [6]Department of Biology, University of Copenhagen, Copenhagen, Denmark

The genetic variation in Northern Asian populations is currently under-sampled. To address this, we have created a new genetic variation reference panel from whole-genome sequencing of 175 Mongolians representing six tribes. We identified 15.2 million single nucleotide polymorphisms (SNPs), of which 3.9 million (25.8%) are novel when compared against the 1000 Genomes project (1000G). Our new reference panel fills an important gap between the East Asians and Admixed Americans provided by the 1000G and significantly improves our ability to impute missing genotypes. This will aid future genome wide association studies (GWAS) for Mongolians and less studied populations of central and North Asia. We observe evidence of strong population stratification among tribes, and variance among Mongolians exceeds that of other East Asian populations. Incorporating our results with the 1000G reveals shared ancestral alleles between Finns and Mongolians, suggesting a path for gene flow between Europeans and East/North Asia. Furthermore, we report multiple lines of evidence supporting the hypothesis that the ancestors of modern East Asians colonized the region from North to South, likely via Western Siberia; this evidence takes into account our Mongolian samples, 1000G, and the Human Genome Diversity Panel (HGDP-CEPH), as well as additional public datasets covering Siberians, Indians, and Tibetans. From an adaptation and human evolution perspective, we have also identified positive selection on dozens of genes known to be involved in metabolism and diseases. In summary, the reference haplotype panel presented here will be a valuable resource for identifying population specific genetic causes of diseases in Central/North Asia, and exploring human evolutionary scenarios throughout Asia.

GENOME-WIDE CHARACTERISATION OF HIGH-TURNOVER REGULATORY REGIONS

Alexander J Nash, Boris Lenhard

Computational Regulatory Genomics Group, MRC London Institute of Medical Sciences, Imperial College London, London, United Kingdom

Metazoan comparative genomics has revealed dense clusters of highly conserved non-coding elements (CNEs), principally around developmental genes. It has been repeatedly demonstrated that CNEs function as enhancers that contribute to the establishment of complex spatio-temporal expression patterns during development. Clusters of CNEs, termed genomic regulatory blocks (GRBs), can span multiple genes, but generally act co-operatively to regulate a single target. While the majority of GRBs are identifiable using species comparisons spanning large evolutionary distances, there exists a subset in which the depth of conservation is greatly reduced. We hypothesize that for this subset of GRBs the observed reduction in depth of conservation is due to continuous CNE turnover, precluding GRB identification by sequence conservation alone. In this study, we define a novel measure of CNE turnover and use this measure to identify a set of high-turnover GRBs. Initial characterization of high-turnover GRBs suggests that they are under less negative selective pressure than low-turnover GRBs and are, on average, expressed later in embryonic development. Gene ontology enrichment analysis identifies distinct functional classes of genes as the targets of high- and low-turnover GRB regulation, highlighting the importance of high-turnover GRBs in the regulation of cell adhesion, and extracellular matrix remodeling during neural development.

LARGE DISEASE COHORT HARMONIZATION WITH WHOLE-GENOME REPLICATES REVEALS SEQUENCING PLATFORM HETEROGENEITY ACROSS VARIANT SPECTRUM

Waleed Nasser[1], Olga Krasheninina[1], Jesse Farek[1], Adam Mansfield[1], Adam English[1], Ziad Khan[1], Will Salerno[1], Eric Boerwinkle[1,2], Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]University of Texas, School of Public Health, Houston, TX

Large-scale human disease sequencing programs such as the Centers for Common Disease Genomics (CCDG), Trans-Omics for Precision Medicine (TOPMed), and Alzheimer's Disease Sequencing Project (ADSP) require analysis of tens of thousands of whole genome sequences with the goal of understanding the genomic basis of complex diseases including cardiovascular, lung, neuropsychiatric, autoimmune/inflammatory, and Alzheimer's diseases. These large cohorts also inform clinical sequencing standards variant interpretation for Mendelian disease and Precision Medicine programs. Quality Control (QC) harmonization of sequencing data from multiple sequencing centers is required prior to phenotype/genotype tests. Here we show harmonized variant sets generated from replicates across multiple sequencing and informatic protocols. These results indicate that indels and structural variants (SV) are strongly sensitive to protocol heterogeneity and require extensive cleaning prior to tertiary analysis.

We will describe the heterogeneity inherent to three sequencing protocols, multiple reference builds (hg19, h37d5, GRCh38) and two sequencing platforms (HiSeq 2000/2500 and HiSeq X) in 584 whole-genome samples. These data include a three-sample replicate set that was sequenced on both sequencing platforms, on each of the three protocols, and aligned to multiple references.

The SNV and small indel heterogeneity in these data is largely driven by sequence read coverage (coverage range 25x to 40x) and sequencing platform. Consistent reprocessing of BAMs on standard protocols resolved SNV heterogeneity across both platforms and all sequencing protocols, though HiSeq X data still generated more indel calls. In a merged deletion set from 12 SV callers, we identified 173,000 deletion events (49%) specific to a single sequencing protocol, although exclusion of these protocol-specific deletions had no effect on heterogeneity. However, by applying naive harmonization algorithm, we were able to exclude only 60,000 deletion loci (17%) to generate a harmonized data set with no apparent heterogeneity. These excluded loci were enriched for 100 bp to 1,000 bp deletions. Taken together, these results suggest how to generate high-specificity variant call sets when aggregating data across multiple projects.

AUTONOMOUS AND PERVASIVE TRANSCRIPTION DECOUPLING REVEAL TRANSCRIPTIONAL ACTIVITY OF LINE-1 ELEMENTS IN SOMATIC TISSUE AND THEIR IMPACT IN HUMAN TUMORS.

Fabio Navarro[1,2,3], Mark Gerstein[1,2,3]

[1]Yale University, Compututational Biology & Bioinformatics Program, New Haven, CT, [2]Yale University, Molecular Biology and Biochemistry, New Haven, CT, [3]Yale University, Computer Science, New Haven, CT

The broad distribution and high copy number of Long Interspaced Nuclear Element (LINE-1 or L1) elements across the human genome hinders the quantification of LINE-1 autonomous transcription. RNA sequencing analysis are specially confounded by RNA fragments emanating from pervasive transcription. We modeled and implemented a new approach that discerns pervasive transcription from autonomous transcription of L1 subfamilies and estimate their autonomous transcription level. We processed more than 9,000 RNA-Seq experiments from multiple datasets (GTEx,TCGA and ENCODE) to evaluate the autonomous activity of LINE-1 subfamilies in human cell lines, healthy organs and tumor tissue. We demonstrate that most of LINE-1 signal emanates from pervasive transcription, however recent and potentially active, LINE-1 subfamilies are autonomously transcribed in healthy tissues and tumors. Most of L1 subfamily autonomous transcription is found in cytoplasmic poly-adenylated transcripts. Moreover, we found that basal ganglia harbor higher activity of L1Hs than other adult brain regions, but, have relatively small autonomous transcription of L1 compared to tissues such as tibial nerve, testes, skin. Transcription of L1 is upregulated in most tumors when compared to their counterpart healthy tissue. When investigating the correlation between genome instability and L1 autonomous activity we find a directly correlation between the L1 transcriptional activity and number of indels in tumor samples. Further investigation suggests that there is an enrichment of L1 endonuclease recognition motif overlapping INDELs, suggesting that emergence INDELs could be facilitated by the activity of L1 endonuclease. We suggest a mechanism in which L1 endonuclease creates double strand breaks that are fixed by error prone DNA repair mediated by NHEJ, thus, creating INDELs.

MULTICLONAL INVASION IN DCIS IDENTIFIED BY TOPOGRAPHIC SINGLE-CELL DNA SEQUENCING

Anna Casasent[1,4], Annalyssa Long[1], Aislyn Schalck[1,4], Alexander Davis[1], Emi Sei[1], Ruli Gao[1], Funda Meric-Bernstam[5], Mary Edgerton[4], Nicholas Navin[1,2,4]

[1]MD Anderson Cancer Center, Genetics, houston, TX, [2]MD Anderson Cancer Center, Bioinformatics, houston, TX, [3]MD Anderson Cancer Center, Pathology, houston, TX, [4]University of Texas, GSBS Graduate School, houston, TX, [5]MD Anderson Cancer Center, Surgical Oncology, Houston, TX

Ductal Carcinoma in Situ (DCIS) is an early stage breast cancer and nonobligatory precursor to invasive breast cancer (IBC). Several models have been proposed to explain the genomic progression of DCIS to IBC, including independent lineages, population bottlenecks and multiclonal invasion. These models have been difficult to resolve in bulk tissues due to the limited number cells in the ducts and extensive intratumor heterogeneity. To address these challenges, we developed an approach that combines laser-capture-microdissection and laser-catapulting with single cell DNA sequencing to measure genomic copy number profiles of single tumor cells in tissue sections, while preserving their spatial geography. We applied this method to sequence 1012 single cells from 10 DCIS patients with high-grade matched in situ and invasive tumor cells to delineate genome evolution during invasion. In parallel we performed deep-exome sequencing of laser-microdissected in situ and invasive regions. Our data revealed early punctuated evolution in the ducts, followed by the migration of multiple subclones that established the invasive carcinomas. These data suggest that multiple subpopulations will need to be targeted to treat DCIS patients and prevent invasion.

INTEGRATED ANALYSIS OF RARE VARIATION IN SCHIZOPHRENIA AND OTHER NEURODEVELOPMENTAL DISORDERS

Hoang T Nguyen[1], Douglas M Ruderfer[2], Giulio Genovese[3], Menachem Fromer[1,4], Pamela Sklar[1], Shaun M Purcell[1], Xin He[5], Patrick F Sullivan[6], Eli Stahl[1]

[1]Icahn School of Medicine at Mount Sinai, Division of Psychiatric Genomics, Department of Genetics and Genomic Sciences, New York, NY, [2]Vanderbilt University, Division of Genetic Medicine, Nashville, TN, [3]Broad Institute, Stanley Center and Medical and Population Genetics Program, Boston, MA, [4]Verily Life Sciences, Palo Alto, CA, [5]University of Chicago, Department of Human Genetics, Chicago, IL, [6]University of North Carolina, Departments of Genetics and Psychiatry, Chape Hill, NC

Next-generation sequencing studies of rare variation from families have successfully implicated specific genes contributing to risk of the neurodevelopmental disorders intellectual disability (ID), severe developmental disorders (DD), epilepsy (EPI), and autism spectrum disorder (ASD). Family and case-control sequencing studies have implicated large gene sets in schizophrenia (SCZ), however, very few individual risk genes have been identified. Here, we develop hierarchical Bayesian models of rare variation in disease, and infer the proportion of risk genes and distribution of risk variant effect sizes across four variant annotation categories: gene-disruptive (nonsense, frameshift, essential splice site), damaging missense and silent-CFPK (silent mutations within frontal cortex-derived DHS) de novo mutations, and disruptive+missense damaging case-control singletons. We applied this method to the largest available schizophrenia collection of exome sequences (1,077 trios, 6,699 cases and 13,028 controls), and to 10,792 families and 4,058 cases/controls of ASD, ID, DD and EPI. We estimate that 8% of genes harbor SCZ risk variants (95% credible interval, CI, 4.6-12.9%), with mean relative risks (95% CI) of 12 (4.8- 22.2), 1.4 (1-3.2) and 1.2 (1.01-2.2) for disruptive, damaging missense and silent-CFPK de novos, respectively, and 2.1 (1.04-3.5), 2.4 (1.04, 5.7) and 1.04 (1-1.2) for disruptive+damaging missense singleton variants in each of three case-control samples. We identify two SCZ risk genes with FDR<0.05, SETD1A and TAF13, and two other genes with FDR<0.1, RB1CC1 and PRRC2A. Estimated proportions of risk genes in other neurodevelopmental disorders were smaller than in SCZ (<5% for disorders). We identify 164 and 58 genes (FDR<0.05) for DD and ID, respectively, including 101 novel DD genes and 15 novel ID genes. Overall, our results in schizophrenia replicate those of previous studies for known gene sets as well as for the single known gene SETD1A, confirming the robustness of the approach, and we achieve greater power in DD and ID by virtue of disease-specific genetic architecture inference.

GENOME-WIDE ASSOCIATION ANALYSIS OF ATRIAL FIBRILLATION IDENTIFIES TWO NEW RISK LOCI AND HIGHLIGHTS BIOLOGICAL PATHWAYS AND REGULATORY ELEMENTS INVOLVED IN CARDIAC DEVELOPMENT

Jonas B Nielsen[1,2], Lars Fritsche[3,4,5], Wei Zhou[2,4], Gonçalo R Abecasis[4,5], Kristian Hveem[3,5], Cristn J Willer[1,2,4]

[1]University of Michigan, Department of Internal Medicine, Division of Cardiovascular Medicine, Ann Arbor, MI, [2]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [3]Norwegian University of Science and Technology, HUNT Research Centre, Department of Public Health and General Practice, Levanger, Norway, [4]University of Michigan, Center for Statistical Genetics, Ann Arbor, MI, [5]Norwegian University of Science and Technology, K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Trondheim, Norway

Atrial fibrillation is a common cardiac arrhythmia and a major risk factor for stroke, heart failure, and death. The pathogenesis of atrial fibrillation remains poorly understood, which contributes to the current lack of highly effective treatment regimens. To understand the genetic variation underlying AF, we undertook a genome-wide association study of 6,337 individuals with AF and 61,607 controls from Norway, with replication in an additional 7,297 atrial fibrillation cases and 133,518 controls. Based on genotyping and dense imputation mapping from whole-genome sequencing, we tested almost 9 million genetic variants across the genome and identified 2 novel risk loci. One of the novel loci comprised several highly correlated missense variants situated in the I, A and M-band of titin, the largest protein in humans and responsible for the passive elasticity of heart and skeletal muscle. The other novel locus comprised a Norwegian-specific risk locus that has previously been associated with QRS-amplitude and QRS-duration, both electrocardiogram (ECG) derived measurements that reflect cardiac structure and function likely related to development of atrial fibrillation. We further applied the framework DEPICT to identify predicted gene function from multiple sources of evidence including protein-protein interactions, phenotypic data from gene knockout experiments in mice, and co-regulation of gene expression. We identified genes at atrial fibrillation-associated loci to be highly enriched in gene sets and pathways important for muscle cell differentiation and tissue formation. These findings were substantiated by enrichment analyses of functional and regulatory elements indicating that many single nucleotide variants at atrial fibrillation-associated loci fall within regions of open chromatin state during fetal heart development. Altogether, these results point to a mechanism of impaired muscle cell differentiation and tissue formation in the developing heart as an important risk factor for atrial fibrillation in adult life.

CHARACTERIZATION OF COMPLETE MOBILE GENETIC ELEMENTS IDENTIFIED BY LONG-READ SEQUENCING OF HUMAN GUT MICROBIOMES

Suguru Nishijima[1,2,3], Yoshihiko Suzuki[2], Wataru Suda[2,3,4,5], Shinichi Morishita[2], Masahira Hattori[2,3,5]

[1]National Institute of Advanced Industrial Science and Technology, Computational Bio-Big Data Open Innovation Lab., Tokyo, Japan, [2]The University of Tokyo, Department of Computational Biology and Medical Sciences, Chiba, Japan, [3]Waseda University, Faculty of Science and Engineering, Tokyo, Japan, [4]Keio University School of Medicine, Department of Microbiology and Immunology, Tokyo, Japan, [5]RIKEN, Center for Integrative Medical Sciences (IMS), Yokohama, Japan

Despite accumulation of large metagenomic datasets of human gut microbiomes, overall states of mobile genetic elements (MGEs) such as plasmids and phages in the community are largely unknown due to the difficulty of completion of their assembly and their separation from microbial chromosomes in the metagenomic data, demonstrating the typical limitations of short-read sequencing.
To enumerate dozens of complete MGEs in the human gut microbiome, we conducted an extremely long metagenomic sequencing of high-molecular weight DNA prepared from the feces of 12 healthy individuals by using the PacBio single-molecule real-time (SMRT) sequencing technology.
We identified 94 complete sequences of MGEs including 69 plasmids and 7 phages hitherto unknown by the assembly of a total of 135 Gb PacBio reads with >9 kb average length generated. The MGEs showed higher diversity between individuals than microbial species in the community. Furthermore, the mapping analysis using the publicly available gut metagenomic data from five nations (Japan, The United States, China, Denmark and Spain) to the MGEs found that 34 plasmids and 3 phages were prevalent in the five nations (>30% in frequency). On the other hand, 41 MGEs (34 plasmids and 7 phages) were significantly different proportion of individuals among the 5 countries. Our data suggest that a profile and distribution of MGEs is associated with host's life style, diet or geography. Metagenomics using long-read sequencing provided the first global picture of the MGEs in the human gut microbiome.

AN INTEGRATIVE ANALYSIS OF PHILADELPHIA-LIKE ACUTE LYMPHOBLASTIC LEUKEMIA

Conor Nodzak[1], Gabriel Centoducatte[2], Andrew Quitadamo[1], J. Andres Yunes[2], Xinghua Shi[1]

[1]University of North Carolina at Charlotte, Bioinformatics and Genomics, Charlotte, NC, [2]Centro Infantil Boldrini, Laboratorio de Biologia Molecular, Campinas, Brazil

The hematological cancer precursor B-cell acute lymphoblastic leukemia commonly afflicts children and adolescents and shows poor prognosis due to the release of immature antigen-presenting cells with unchecked replicative capacity into circulation. [1][2] Among patients suffering from the disease, there are many subtypes that can be defined based upon particular genomic rearrangements and the presence or absence of expressed fusion products. These subtypes may be experimentally determined from cytogenetics, molecular profiling or immuno-phenotyping protocols. [3] From this classification scheme, the Philadelphia-like subtype may be identified by a reciprocal translocation between the long arms of chromosomes 9 and 22 with a lack of an expressed fusion product of the BCR and ABL1 kinases. To assess the differences in gene activity for this rare class of leukemia, microarrays from the NCI TARGET initiative were used and transcript abundances for 223 Ph-like samples were quantified and tested for differential expression and gene-set enrichment against 1,096 samples representing eight distinct ALL subtypes. [4] Furthermore, the Ph-like group was divided in half by a disruption of the CRLF2 gene and reanalyzed for variation across the groups. As a result, it could be shown that pathways involved in cytoskeletal rearrangement, cell signaling and proliferation were consistently up regulated in Ph-like precursor B-cell ALL patients relative to subtypes harboring other genomic rearrangements. In addition, the Ph-like subtype of ALL showed similar enrichment in kinase-mediated signaling pathways to patients known to express the BCR-ABL1 fusion protein. A subset of 126 Ph-like patients from TARGET were also analyzed using RNA-seq data, from which heterozygous sites exhibiting allele specific expression were found using an FDR corrected binomial test on the allelic ratios of uniquely mapped reads. Using a combination of gene expression and SNP arrays, significant eQTLs were found and mapped back to the differentially expressed genes and regions exhibiting allele specific expression. The results of this analysis provide a set of examples whereby the genetic variation in patients with Ph-like leukemia may influence gene expression levels compared to other subtypes and where these differences can act in an allele specific manner.

PHYLOGENY-BASED NOMENCLATURE FOR OLFACTORY RECEPTORS IN DIVERSE VERTEBRATES

Tsviya Olender[1], Elspeth Bruford[2], Doron Lancet[1]

[1]Weizmann Institute of Science, Molecular Genetics, Rehovot, Israel, [2]European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom

Olfactory Receptors (ORs) are G protein-coupled receptors with a crucial role in odor detection. There are ~1000 OR genes and pseudogenes in a typical mammalian genome, however the number of functional ORs varies among species reflecting their adaptation to different environments, a process which involves gene duplication/deletion events. In human the widely accepted OR nomenclature is based on sequence similarity classification into 18 families, and further classification into subfamilies and members Thus, OR1A2 is gene number 2 from subfamily A in family 1, providing important phylogenetic, structural and functional insights. We have launched an effort to generate a similarly effective nomenclature for diverse vertebrates, a challenging task due to non-trivial orthology relationships among ORs. We describe the Mutual Maximum Similarity (MMS) algorithm for creating a human-based OR nomenclature for other species, via detecting inter-species hierarchical pairwise similarity relations. This was already applied to OR repertoires of mouse, rat, cow, dog, opossum, platypus, orangutan and chimpanzee. We are currently extending this work to non-mammalian species, including the maximally remote zebrafish whose nomenclature portrays 19 OR families, of which only 4 are shared with mammals. This suggested nomenclature is supported by both synteny and phylogenetic information.

The availability of a unified nomenclature system provides a powerful framework for diverse studies of vertebrate ORs. Thus, by using textual symbol comparison an immediate identification of potential ortholog groups is easily obtained. The symbols also portray species-specific OR repertoire expansions or diminutions, e.g. a relatively recent duplication of *OR52E5* in rat (*Or52e5*, *OR52e5b*). Another example is vertebrate OR subfamily 6Z that is entirely absent among apes OR symbols. In other mammals, members of this subfamily are disposed in one genomic cluster, suggesting a large deletion in the early ape lineage. While in dog and cow ~50% of the symbols are identical to human symbols, the number decreases to only ~30% in mouse and rat. This result is in line with the literature and reflects the adaptive changes of the OR gene superfamily across to diverse ecological niche.

The nomenclatures are available in the HORDE database (https://genome.weizmann.ac.il/horde/). Use of such proposed nomenclatures, under consideration among the relevant nomenclature committees, should help identify orthology relations by symbol-based scrutiny. Data for additional vertebrates will added to HORDE in due course.

CROSS-TISSUE PROTEIN EXPRESSION IN THE GENOTYPE-TISSUE EXPRESSION (GTEX) COLLECTION

Meritxell Oliva[1,2], Marian Fernando[1,2], Caroline Linke[1,2], Fan Wu[1,2], Andrew Skol[2,3], Barbara E Stranger[1,2,3]

[1]Section of Genetic Medicine, Department of Medicine, University of Chicago, IL, [2]The Institute for Genomics and Systems Biology, University of Chicago, IL, [3]Center for Data Intensive Science, University of Chicago, IL

The human transcriptome has been studied extensively but equivalent investigations of the proteome have been limited by the lack of robust technologies for large-scale protein quantification. Because of the dynamic nature of the proteome that responds in part to genetic variation, the proteome comprises a vital element in the causal network linking genetic variation to higher order organismal phenotypes. As part of the enhancing Genotype-Tissue Expression program (eGTEx), in which additional –omics assays are performed on human tissue samples that have been deeply characterized at the transcriptomic and genetic level, we investigate protein expression features across tissues of the GTEx collection. This initial phase of the project aims to characterize tissue-specific protein abundance profiles, and relate those to the transcriptome profiles derived from the same individuals and tissues.

We have adapted microwestern arrays to quantify protein expression of 353 proteins involved in transcription regulation and cell signaling in 203 GTEx samples comprising 33 tissues of 14 individuals (per tissue N=2-11). We characterized the protein expression landscape of these proteins across all samples in three technical replicates to characterize tissue specificity at the protein level, and to compare protein expression profiles to their corresponding transcript counterparts in order to assess the extent to which transcriptome-derived patterns are reflected at the protein level and vice versa. The GTEx tissue clustering profile derived from the protein expression data clearly recapitulates tissue similarity relationships derived from transcriptome data. Relative protein abundance levels reveal strong tissue specificity: proteins are expressed in a tissue-enriched (39%) or –enhanced (36%) manner. Brain, muscle and skin are enriched for tissue-specific expression relative to other tissues. For many proteins, cross-tissue ranks are not strongly correlated with ranks derived from transcript levels. Within tissues, we observe significant inter-individual protein variation and both positive and negative correlations of inter-individual mRNA and protein levels. We highlight examples of GTEx expression quantitative trait loci (eQTLs) that are consistent at the protein level, suggesting pQTLs.

Currently, we are quantifying the same 353 proteins within and between ten GTEx tissues (N=235 per tissue) to better characterize the protein expression landscape, assess the genetic basis of protein expression variation within and between tissues, and build protein-mRNA regulatory networks. These data will be analyzed in a multi-omics framework, utilizing other eGTEx datasets from the same samples and individuals.

FINE MAPPING GENOME-WIDE ASSOCIATION IN NARCOLEPSY DEFINES NOVEL DISEASE MECHANISMS.

Hanna M Ollila[1], Ryan Hillary[1], Ling Ling[1], Joachim Hallmayer[1], Jimmie Ye[2], Fang Han[3], Emmanuel Mignot[1]

[1]Stanford University School of Medicine, Psychiatry and Behavioral Sciences, Palo Alto, CA, [2]University of California San Francisco, Department of Epidemiology and Biostatistics, Institute of Human Genetics, San Francisco, CA, [3]Peking University People's Hospital, Department of Surgery, Beijing, China

Introduction

Type 1 Narcolepsy is characterized by sleepiness, REM sleep abnormalities and loss of muscle tone triggered by positive emotions (cataplexy). The cause of type 1 narcolepsy is a loss of neurons producing the hypocretin/orexin peptide of likely autoimmune origin. Our aim was to discover novel genetic variants in narcolepsy and fine map the potentially causative variants using a transethnic sample and cellular models.

Methods

We examined genetic variants using a transethnic GWAS in Asian, African American and Caucasian samples (N=5,500 cases and 21,500 controls). Function of the leading variants was examined using eQTL analysis in dendritic and T cell models, data from the ENCODE and GTEx consortiums and examining the effect of individual variants on flu vaccination and immune cell development using mass cytometry.

Results

We confirmed existing risk associations (TRA, TRB, IFNAR1, ZNF365, CTSH and P2RY11) and discovered novel loci that predisposed to narcolepsy in CD207, SIRPG, FLT3, ZFAND2A and PRF1. Fine mapping of association suggests a functional polymorphism in position A91V in PRF1, a variant that is directly affecting T and NK cell mediated cell killing. Furthermore, leading variant in IFNAR1 affected IFNAR1 expression after flu infection in dendritic cells suggesting causality for the development of narcolepsy.

Conclusions

The results further stress the effect of T cell-dendritic cell interactions in the development of narcolepsy and find causal pathways. The novel loci may explain how hypocretin cells are destroyed and support a T cell mediated autoimmune attack in narcolepsy susceptibility.

INFERRING THE EXTENT OF BACKGROUND SELECTION USING AN APPROXIMATE BAYESIAN APPROACH

Louise N Ormond[1], Susanne Pfeifer[1,2], Sebastian Matuszewski[1], Stefan Laurent[1], Jeffrey D Jensen[1,2]

[1]Ecole Polytechnique Federale de Lausanne, School of Life Sciences, Lausanne, Switzerland, [2]Arizona State University, Center for Evolution and Medicine, Tempe, AZ

Many studies have focused on identifying signatures of selective sweeps associated with beneficial mutations – both for better quantifying the adaptive process, as well as for characterizing the importance of linked positive selection in shaping genomic variation. Yet studies of the distribution of fitness effects (DFE) of newly arising mutations have demonstrated that a far greater proportion of mutations are deleterious. The resulting linked negative selection effects from this process are rarely accounted for in population genetic inference. This omission biases our understanding of the evolutionary process, and potentially also confounds signatures of selective sweeps and demographic processes. While strongly deleterious mutations will be rapidly purged from the population, and thus may be simply treated as a rescaling of effective population size, the effects of background selection (BGS) are most evident under a regime of weakly deleterious mutations. In this space a simple re-scaling is not sufficient, and the site frequency spectrum may become skewed towards low frequency variants. Here we use forward simulations to identify a range of potential parameters and summary statistics that capture the effects of BGS, and integrate these into an approximate Bayesian approach to estimate the impact of linked negative selection on genomic variation. Our model assumes a functional region experiencing a mix of weak and strong purifying selection flanked by linked neutral regions. We apply our method to estimate BGS levels from polymorphism data for humans and chimpanzees.

THE CANCER GENOME COLLABORATORY

Christina Y Yung[1], George L Mihaiescu[1], Bob Tiernay[1], Junjun Zhang[1], Francois Gerthoffert[1], Andy Yang[1], Jared Baker[1], Guillaume Bourque[2], Paul C Boutros[1,3], Bartha M Knoppers[2], B.F. Francis Ouellette[1,4], Cenk Sahinalp[5], Sohrab P Shah[6,7,8], Michelle D Brazas[1], Vincent Ferretti[1], Lincoln Stein[1,9]

[1]Ontario Institute for Cancer REsearch, Informatics and Biocomputing, Toronto, Canada, [2]McGill University, Medical Genetics, Montreal, Canada, [3]University of Toronto, Medical Biophysics, Toronto, Canada, [4]University of Toronto, Cell and Systems Biology, Toronto, Canada, [5]Simon Fraser University, Computer Science, Burnaby, Canada, [6]BC Cancer Agency, Molecular Oncology, Vancouver, Canada, [7]University of British Columbia, Molecular Pathology, Vancouver, Canada, [8]University of British Columbia, Computer Science, Vancouver, Canada, [9]University of Toronto, Molecular GeneticsToronto, Canada

The Cancer Genome Collaboratory (CGC) is an academic compute cloud designed to enable computational research on the world's largest & most comprehensive cancer genome dataset, ICGC. A subproject of ICGC, the PanCancer Analysis of Whole Genomes (PCAWG) alone has generated over 800TB of harmonized sequence alignments, variants & interpreted data from over 2,800 cancer patients. To facilitate the computational analysis on the ICGC data, the CGC has developed software solutions that are optimized for typical cancer genomics workloads, including well tested and accurate genome aligners and somatic variant calling pipelines. We have developed a simple to use, but fast & secure, data transfer tool that imports genomic data from cloud object storage into the user's compute instances. Because a growing number of cancer datasets have restrictions on their storage locations, it is important to have software solutions that are interoperable across multiple cloud environments. We have successfully demonstrated interoperability across the TCGA dataset hosted at University of Chicago's Bionimbus Protected Data Cloud, the ICGC dataset hosted at the CGC, & ICGC datasets stored in the Amazon Web Services (AWS) S3 storage. Lastly, we have developed a non-intrusive user authorization system that allows the CGC to authenticate against the ICGC DACO when researchers require access to controlled tier data. We anticipate that our software solutions will be implemented on additional commercial and academic clouds. The CGC is actively growing, with a target hardware infrastructure of over 3000 CPU cores and 15 petabytes of raw storage. As of November 2016, the CGC holds information on 2,000 ICGC PCAWG donors (500TB total). We anticipate expanding the CGC to host the entire ICGC dataset of 25,000 donors (approximately 5PB) and to extend its data management and analysis facilities across multiple clouds. During the current closed beta phase, the CGC has been successfully utilized by multiple research groups, most notably PCAWG project researchers who analyzed thousands of genomes at scale over a few weeks' time. The CGC will open to the public during the 2nd quarter of 2017. We invite cancer researchers to learn more about our cloud resources at cancercollaboratory.org, and apply for access to the CGC.

OPEN, COLLABORATIVE, SHAREABLE BIOINFORMATICS WORKSHOP TUTORIALS

Francis Ouellette[1,2], Michelle D Brazas[1], Ann Meyer[1]

[1]Ontario Institute for Cancer Research, Informatics and Biocomputing, Toronto, Canada, [2]University of Toronto, Cell and Systems Biology, Toronto, Canada

Bioinformatics.ca hosts workshops covering a wide range of topics in bioinformatics from introductory to advanced courses. We strongly support open science and believe that the educational resources for science should be open as well. A challenge in making workshop materials open is finding a platform that allows instructors to create content collaboratively and update that content on the fly, while allowing students in the workshop to be comfortable viewing and interacting with the content. GitHub, a powerful resource popular for sharing and collaborating on open source code projects, offers a solution to this challenge. Workshop content can be uploaded, stored, and edited by collaborating instructors. Using GitHub pages, the workshop content is displayed as a fully customizable website providing a familiar experience to students. All content in the workshop is openly available and anyone can view or download it. GitHub also makes it easy for workshop materials to be shared and modified for use by anyone. We will illustrate the use of GitHub and GitHub pages for workshop content creation and sharing.

AN ATTEMPT TO STUDY THE MODERN HUMAN PHENOTYPE

Kaja Moczulska[1], Felipe Mora-Bermúdez[2], Ben Vernot[1], J. Gray Camp[1], Michael Dannemann[1], Farhath Badsha[2], Sabina Kanton[1], Guido Vacano[3], Michael Boyle[1], Stephane Peyregne[1], Maria Schreiter[1], Kay Prüfer[1], David Patterson[3], Elena Taverna[1], Barbara Treutlein[1,2], Wieland B Huttner[2], Janet Kelso[1], Svante Pääbo[1]

[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany, [3]University of Denver, Department of Biological Sciences, Denver, CO

The determination of genome sequences of a Neandertal and a Denisovan has allowed 31,380 single nucleotide substitutions and 125 small insertions and deletions that are present in all or almost all humans today but not in Neandertals, Denisovans or other primates to be identified. We have embarked on a systematic study of these genomic changes, focusing on those that may affect the development and function of the central nervous system and that show evidence of having been positively selected on the human lineage.

We study these features by three approaches: *(i)* **Transgenic mice.** We have introduced primate and modern human-specific changes in the genes encoding the enzymes adenylosuccinate lyase and glycine decarboxylase and study metabolic effects in juvenile and adult mice. Similarly, we have introduced human-specific changes in the gene encoding the transcription factor Foxp2. Analysis of the transcriptomes of single cells reveal cell type-specific effects in the developing mouse brain. *(ii)* **Genome editing of human stem cells,** which are then differentiated into neurons and into brain organoids. As a starting point, we have analyzed brain organoids generated from ape and human iPS cells and detected a prolongation of the mitotic metaphase specific to human neural progenitors. We are now introducing six Neandertal-like amino acid changes in the spindle and kinetochore proteins CASC5, KIF18A, SPAG5 to explore how they affect progenitor mitoses. *(iii)* **Microinjection of mRNA into neural cells.** We study genes involved in neurite outgrowth that carry changes causing amino acid substitutions by injecting pools of in vitro transcribed, capped and polyadenylated mRNAs encoding the ancestral and derived versions of the proteins into neurons in culture and analyze the structure and complexity of dendritic trees, synaptogenesis and spine structure.

THE MAJOR DETERMINANTS OF GENOME-WIDE mRNA SPLICING EFFICIENCY IN FLIES

Athma A Pai[1], Telmo Henriques[2,3], Joseph Paggi[1], Adam Burkholder[4], Karen Adelman[2,3], Christopher B Burge[1,5]

[1]MIT, Department of Biology, Cambridge, MA, [2]NIEHS, Epigenetics and Stem Cell Biology Laboratory, Research Triangle Park, NC, [3]Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, Boston, MA, [4]NIEHS, Center for Integrative Bioinformatics, Research Triangle Park, NC, [5]MIT, Department of Biological Engineering, Cambridge, MA

The dynamics of gene expression may impact regulation, and the processing of nascent RNA molecules into mature RNA can be a rate-limiting step for establishing gene expression equilibrium. To assess the rates of pre-mRNA splicing, we used a short, progressive metabolic labeling strategy followed by RNA sequencing to capture nascent RNA molecules and estimate the intron half-lives of ~30,000 introns in *Drosophila melanogaster* S2 cells. We find that splicing rates are strongly correlated with several gene features. Splicing rates varied with intron length (independent of splice site strength) and were fastest for introns of length 60-70 nt, which is the most abundant intron length class in the *Drosophila* genome. Using our nascent sequencing data, we also identified hundreds of novel recursively spliced segments, where long introns are spliced in multiple segments rather than one unit. We expanded the catalog of known recursively spliced introns in flies by 4-fold, though sub-sampling and saturation analyses indicated that we are still underestimating the true number of recursive sites in the *Drosophila* genome. We find that recursive splicing is associated with much faster and also more accurate splicing of the ultra-long introns in which they occur. Together, intron length accounts for ~30% of variance in splicing rates and the presence of recursive sites is associated with a two-fold reduction in half-life. Building on these observations, we developed a model that accounts for greater than 50% of the variability in splicing rates across *Drosophila* introns. Surprisingly, introns within the same gene tend to have similar splicing half-lives and longer first introns are associated with faster splicing of subsequent introns. Our results indicate that genes have different intrinsic rates of splicing, and suggest that these rates are influenced by gene architecture and molecular events at gene 5' ends, likely tuning the dynamics of developmental gene expression.

GENE EXPRESSION AS A MEDIATOR OF TYPE 2 DIABETES SUSCEPTIBILITY VARIANTS

Anthony J Payne[1], Anne Ndungu[1], Jason M Torres[1], Cecilia M Lindgren[1,2], Martijn van de Bunt[1,3], Mark I McCarthy[1,3]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [2]Big Data Institute, University of Oxford, Oxford, United Kingdom, [3]Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, United Kingdom

To assess the potential of gene expression (GE) as a mediator of type 2 diabetes (T2D) susceptibility variants, we used two approaches that incorporated genotype data and human pancreatic islet RNA-sequencing data from 174 individuals.

We first identified single-variant islet eQTLs (FDR < 0.1) that were coincident with established T2D association signals. We focused on 10 independent eQTLs (involving 9 genes) for which the T2D association attained $p < 5 \times 10^{-5}$ in the DIAGRAM v3 T2D GWAS meta-analysis.

To obtain stronger genetic instruments for GE, we used LASSO regression to derive multi-variant (MV) predictors of islet GE. We found that these predictors often included SNPs that were correlated or had negligible impact on GE. Therefore, we filtered out SNPs in perfect LD with another included SNP or that contributed to the last 5% of a model's R^2. Remaining SNPs were then re-modelled with ridge regression. For each gene with prediction $R^2 > 0.05$, we applied MetaXcan with DIAGRAM results to identify genes that mediate genetic associations with T2D.

We found six significant genes (FDR < 0.1) using MetaXcan, three of which replicated our previous single-variant (SV) signals: *ADCY5*, *DGKB*, and *STARD10*. Six genes from the SV analysis did not replicate: the MV models of these genes incorporated additional SNPs that contributed to GE prediction but had no detectable association with T2D.

To explore this further, we decomposed the MetaXcan score of each gene into its constituent SNP effects, and identified the SNP making the strongest contribution (lead SNP). For 5/6 MetaXcan-significant genes, the lead SNP accounted for >50% of the gene's score. *ADCY5* and *STARD10* were in this category; the three others represented SV effects that were below thresholds in our initial SV analysis. *DGKB* was the only gene with two independent T2D-associated eQTLs in our SV analysis, and its lead SNP contributed only 27% of its MetaXcan score.

Our study has demonstrated the complexity of relationships and the potential discordance between the genetic bases of GE and disease. SV analysis and MetaXcan jointly identified 12 potential effector genes for T2D-associated variants, but only *ADCY5*, *STARD10*, and *DGKB* had consistent effects using both approaches. Thus, while the methods used can discover potential effector genes for susceptibility variants, the complex nature of significant effects should be considered when interpreting results.

INDEXCOV: FAST WHOLE-GENOME COVERAGE QUALITY-CONTROL USING BAM OR CRAM INDEXES

Brent S Pedersen[1], Ryan Collins[2], Michael Talkowski[2], Stephan Sanders[3], ASC/SSC WGS Consortium[4], Aaron Quinlan[1]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]Harvard, Center for Human Genetic Research, Boston, MA, [3]UCSF School of Medicine, Psychiatry, San Francisco, CA, [4]ASC/SSC, WGS Consortium, Various, UT

Whole genome sequencing (WGS) projects costing thousands of dollars produce datasets requiring potentially dozens of hours of analysis time per sample. Verifying the integrity of the data is crucial, but difficult, because of the size. An aligned whole-genome BAM at 30X coverage is 60GB on disk. Consequently, the simple task of iterating through the BAM records requires hours of processing time. Coverage-based quality control is a necessary step to precede variant-calling and especially copy-number and structural variant calling. In large cohorts, a problematic sample may go undetected until long after these steps are completed without a way to quickly assay coverage across all samples and see the results.

A BAM index contains a linear index that indicates the file position of the first read in each 16KB region of the genome. We show that the difference between file positions in this linear index and in the CRAM slice index serves as an accurate proxy for depth. We use this proxy in indexcov to quickly find coverage problems and chromosomal anomalies, and to infer sex **in as little as 2 seconds per sample**. The result of running indexcov is an HTML page with interactive plots showing a global indication of coverage, the inferred sex, PCA, and coverage of each chromosome. We demonstrate the utility of indexcov on data from the Simons Autism cohort. Indexcov is available at: https://github.com/brentp/goleft under the MIT license.

TOWARD THE GAPLESS ASSEMBLY OF VERTEBRATE GENOMES

Sergey Koren, Brian Walenz, Alexander Dilthey, Arang Rhie, Adam Phillippy

National Human Genome Research Institute, Computational and Statistical Genomics Branch, Bethesda, MD

A complete and accurate genome sequence forms the basis of all downstream genomic analyses. However, even the human reference genome remains imperfect, which affects the quality of experiments and can mask true genomic variations. For many other species, quality reference genomes do not exist. Long-read, single-molecule sequencing has begun to correct this deficiency and enabled the automated reconstruction of reference-quality genomes. Combination of this sequencing technology with chromatin conformation capture (Hi-C) scaffolding, may soon enable the gapless reconstruction of complete vertebrate genomes. We have illustrated the potential of this approach on several genomes, including hummingbird, domestic goat, and human. For the goat genome, this approach achieved a contig NG50 of 19 Mbp, a scaffold NG50 of 87 Mbp, and only 649 gaps using a combination of PacBio sequencing, BioNano optical maps, and Hi-C scaffolding. Thus, the de novo assembly of vertebrate genomes now approaches finished quality at a fraction of the cost previously required.

Unresolved assembly gaps include large segmental duplications and simple sequence repeat arrays characteristic of heterochromatin. The resolution of such regions will require either a drastic increase in read length, or an improvement in single-molequle sequence accuracy for the better differentiation of near-identical repeat copies. Oxford Nanopore sequencing offers a promising alternative to PacBio that is capable of >100 kbp read lengths. We have trialed this technology on BAC clones isolated from the complex human rDNA regions, and were able to generate single reads spanning entire BACs, revealing novel sequence structures. Other groups have recently sequenced whole human genomes to as much as 60X coverage using nanopore, which we have successfully assembled. In the best case, these assemblies resulted in a contig NG50 of 23 Mbp and included multi-kilobase assemblies of higher-order alpha satellite sequence. Thus, we conclude that Oxford Nanopore sequencing is now applicable to eukaryotic genome assembly, and will assist in closing the remaining human genome gaps.

THE EVOLUTION OF ALTERNATIVE SPLICING COMPLEXITY IN PRIMATES

Lenore Pipes[1,2,3], Adam Siepel[3], Christopher E Mason[2]

[1]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY, [2]Weill Cornell Medical College, Institute of Computational Biology, New York, NY, [3]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Alternative splicing produces a wide variety of distinct messenger RNA (mRNA) isoforms that may be translated into diverse protein products, which has been claimed to be the source of the key to mammalian phenotypic complexity. While the frequency and complexity of splicing increases along the phylogenetic lineage towards humans, the extent to which splicing is conserved in primates remains largely unknown. To study the variation in primate splicing patterns encompassing >70 million years of evolution, we assembled transcriptomes de novo by pooling >3 billion RNA-Seq reads across 14 matched tissues for human and 10 non-human primates from data generated from the Genome-Tissue Expression (GTEx) Project and the Non-Human Primate Reference Transcriptome Resource (NHPRTR), respectively. Using these transcript assemblies, we created comparative primate alternative splicing databases for the four main types of alternative splicing events that represent the majority of the type of splicing events prevalent in mammals: exon skipping, alternative acceptor site, alternative donor site, and intron retention. This allowed us to analyze one-to-one orthologous splicing events and quantify the percent spliced in index (PSI) per tissue for each primate species. We find that there are large variations in PSI that are often species-specific and/or taxonomically-restricted including evidence that there is an excess of changes from minor-form exons to major-form exons in skipped exon events. Conserved alternative exons also show increasing or decreasing PSI values which mimic the phylogenetic divergence from humans. Our study of taxonomically-restricted splicing events also revealed an association with splicing factor motifs and an enrichment of these events in a subset of tissues. We validated many of these taxonomically-restricted splicing events using RT-PCR in multiple individuals from the same species. We provide the first quantitative global assessment of PSI across different types of splicing in primates as well as an in-depth analysis to species-specific cases that could be the source of phenotypic innovation in primates. This not only represents one of the first attempts to understanding splicing complexity in primates but also represents a systematic approach for gaining insights into the evolution of splicing across species when annotation is not yet available.

IDENTIFYING FACTORS AND GENETIC VARIANTS UNDERLYING GXE INTERACTIONS WITH ATAC-SEQ

Roger Pique-Regi[1,3], Donovan Watza[1], Alexander Shanku[1], Xiaoquan Wen[4], Heejung Shim[2], Francesca Luca[1,3]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]Purdue University, Statistics, West Lafayette, IN, [3]Wayne State University, Obstetrics and Ginecology, Detroit, MI, [4]University of Michigan, Biostatistics, Ann Arbor, MI

A large fraction of loci important in determining human traits and disease conditions are located in non-coding regions of the genome. These regions likely contain specific regulatory sequences that control gene transcription and can also interact with changes in the cellular environment (e.g. drug treatment). Recent large scale efforts in functional genomics have facilitated the profiling of regulatory sequences across many cell-types and tissues, yet we are still very far from mapping the sequences that control the transcriptional response to many external stimuli. Importantly, much of the missing heritability in GWAS may be a consequence of small effect sizes dampened by unaccounted environmental interactions. Supporting this hypothesis, we recently demonstrated that genes with gene-environment (GxE) interactions at the molecular level are highly enriched in GWAS. More precisely, 50% of genes with condition-specific allele specific expression in a study screening 250 cellular environments were also found associated with GWAS traits. Here, we profiled five of the environmental conditions with very large gene expression changes for transcription factor binding activity. This was accomplished at a genome-wide scale by ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. We further modeled the Tn5 cleavage pattern "footprint" of transcription factors with known motifs with CENTIPEDE to identify bound binding sites. From our analyses we were able to resolve 383 actively bound motifs across all conditions. We were also able to characterize 5,236 regions that have significantly changed chromatin accessibility (FDR < 10%) in response to both copper and selenium. We have extended the CENTIPEDE model hierarchical prior to detect motifs that have differences in footprint activity in treatment vs. control experiments. For both metal ions we have detected a significant increase of binding for ETS and CRE motifs. Using a sequence model we previously developed for analyzing ENCODE data (now being updated with a deep learning strategy) we can annotate many genetic variants that are in binding sites specific to the treatment. We can validate several examples using allele specific Tn5 hypersensitivity (ASH) and conditional ASH (cASH). Our results demonstrate that ATAC-seq together with an improved footprint model are excellent tools for rapid profiling of transcription binding factor activity to study cellular regulatory response to the environment and molecular mechanisms underlying GxE.

UNDERSTANDING SUSCEPTIBILITY TO EPITHELIAL CANCERS USING COMPARATIVE GENOMICS

Jason Pizzollo[1], William J Nielsen[2], Yoichiro Shibata[3], Alexias Safi[3], Gregory E Crawford[3], Gregory A Wray[2,4], Courtney C Babbitt[1]

[1]University of Massachusetts Amherst, Department of Biology, Amherst, MA, [2]Duke University, Department of Biology, Durham, NC, [3]Duke University, Department of Pediatrics, Division of Medical Genetics, Durham, NC, [4]Duke University, Department of Evolutionary Anthropology, Durham, NC

Fibroblasts exposed to serum undergo a defined pattern of transcription activation that mimics the gene expression profile associated with wound healing. Tumors have been called wounds to do not heal, and exhibit similar patterns of gene expression to those found in actively healing wounds. The core serum response of fibroblasts has been well described in human fibroblasts and has prognostic value in evaluation of cancer progression with the ability to predict metastasis based on gene expression in tumors. In human populations, incidence of epithelial cancer is high, and causes up to 20% of deaths in modern populations. In our nearest living evolutionary ancestor, chimpanzees, however, rates of similar cancers are up to ten-fold lower. Disease risk is a function of both environmental and genetic factors, and though exposure to environmental agents contributes substantially to differential disease risk, we wanted to investigate if there is a genetic component as well between species. Here, we investigated the fibroblast serum response in a comparative manner using fibroblasts isolated from humans and chimpanzees. We explored phenotypic differences between species with RNA-Seq to look at global patterns of transcription, and DNase-Seq to look at changes in active chromatin during the serum response. Our data show that human fibroblasts have focused increases in expression of genes associated with wound healing and cancer pathways. Chimpanzee fibroblasts, on the other hand, do not engage in concentrated gene expression patterns. Although they have higher levels of gene expression before and during the assay, chimpanzee fibroblasts respond with a general decrease in expression. Similarly, chromatin accessibility increases in human but decreases in chimpanzee during the serum response. The relationship between level of open chromatin and gene expression has a weak but positive correlation, and in both species many changes in chromatin accessibility occur near genes that are associated with transcription and cell adhesion. Together, these data show distinct differences in gene expression and chromatin state between species with a focused response in gene expression and chromatin activation in human fibroblasts.

GENOME-WIDE PREDICTION OF REGULATORY TERRITORIES AND TARGET GENES UNDER LONG-RANGE TRANSCRIPTIONAL REGULATION

Ge Tan, Dimitris Polychronopoulos, Boris Lenhard

Computational Regulatory Genomics Group, MRC London Institute of Medical Sciences, Hammersmith Campus, Imperial College London, London, United Kingdom

Comparative genomics and high-throughput experimental methods like ChIP-Seq have led to the identification of regulatory elements in metazoan genomes. However, the assignment of those elements to their target genes still remains a challenging task. Conventional approaches based on assignment to the nearest gene, or manual and semi-intuitive processes are often unreliable, since regulatory regions can be located far away from their target genes, even within neighbouring genes. We previously showed that arrays of conserved noncoding elements, named genomic regulatory blocks (GRBs), span the loci of developmental regulatory genes ("GRB target genes") and several other genes ("bystander genes"). Moreover, target genes that respond to distal regulatory elements in those regions bear specific features that distinguish them from bystander genes in the vicinity, and in the genome. In this study, we propose a robust approach for the computational determination of GRB spans, and a method based on machine learning for genome-wide detection of target genes. We present a comprehensive catalogue of nearly one thousand human genes serving as hubs of long-range regulatory interactions. This catalogue consists of a large number of genes involved in development, transcription, axon guidance and other cell adhesion processes. In addition, most of those genes are involved in complex diseases, including diabetes and cancer. The GRB boundaries and target genes identified in this study provide a valuable resource for studying developmental regulation and disease-associated genomic variation.

MASSIVE A-TO-I RNA EDITING IS COMMON ACROSS ALL METAZOA

Hagit T Porath[1], Eli Eisenberg[2], Erez Y Levanon[1]

[1]Bar-Ilan University, The Mina and Everard Goodman Faculty of Life Sciences, Ramat-Gan, Israel, [2]Tel Aviv University, Sagol School of Neuroscience, Tel Aviv, Israel

RNA editing by adenosine deamination is probably the most frequent post-transcriptional modifications, where selected adenosines (A) are converted to inosine (I) within RNA molecules. The A-to-I deamination is catalyzed by the ADAR (Adenosine Deaminase that Acts on RNA) family of enzymes and in most cases is taking place in the primate specific *Alu* repeats. ADARs are ubiquitously expressed among multicellular metazoan, however genome-wide screenings of editing events have only been explored in a handful of organisms. Mainly, because matching DNA and RNA sequencing data from the same sample and SNP data, required for most RNA editing detection approaches, are available for only a limited number of animals. Here we apply a computational procedure to detect hyper-editing reads as a reliable method to identify RNA editing signals in RNA-seq data, without the need for corresponding DNA-seq or any prior knowledge about SNP data. 19 organisms from human to *Arabidopsis* were analyzed. As expected only the metazoan species that contain the *adar* gene showed editing signals. We detected numerous editing events in all the studied animals. Specifically, 665,790 hyper-edited reads were identified with 4,056,319 editing events. Of them 1,651,907 were unique editing sites. The normalized results demonstrated that human is actually not exceptional in its editing levels. Examining the evolution of ADAR sequence context across the tested animals showed that ADAR motifs were clustered into two different groups according to their phylogenetic position, suggesting that ADAR evolves along with the species itself. We found that although editing is very common in general, editing in coding regions are extremely rare, whereas hyper-editing has a remarkable tendency to be found in repeats regions and predicted dsRNA structures. Repeats can form dsRNA structures, which is the ADAR target, by hybridizing with nearby, oppositely oriented same sequence. The abundant editing in putative dsRNA regions compared with the almost non-existing levels in coding regions suggesting that the main activity of ADAR is to suppress the innate immune response to endogenous dsRNA structures. Accordingly, we found that hyper-editing signals significantly correlate with the potential to form dsRNA structures. In conclusion, we establish the high prevalence of RNA editing as a global phenomenon among metazoans.

HETEROCHROMATIN MODULATION BY VERTEBRATE-SPECIFIC HUSH PROTEIN COMPLEX

Daniil M Prigozhin, Christopher H Douse, Iva A Tchasovnikarova, Richard T Timms, Paul J Lehner, Yorgo Modis

University of Cambridge, Department of Medicine, Cambridge, United Kingdom

Majority of known factors responsible for heterochromatin establishment and maintenance were discovered using genetic screens in yeast and *Drosophila*. Recently, a screen performed in human cells identified a vertebrate-specific Human Silencing Hub (HUSH) complex as a central player in heterochromatin maintenance. HUSH complex comprises three proteins: TASOR, periphilin, and MPP8. It resides at genomic loci rich in H3K9me3 and can silence newly integrated DNA sequences. HUSH activity depends on SetDB1/ATF7IP H3K9 methyltransferase complex and on MORC2, a chromatin-compacting ATPase. The exact roles of HUSH components in gene silencing and what controls its localization to specific genomic locations remains unknown. The goal of my project is to determine the molecular mechanisms of interactions among HUSH components, its recognition of target loci, and recruitment of the key chromatin modifiers.

A HIGH-COVERAGE GENOME OF A NEANDERTAL FROM VINDIJA CAVE IN CROATIA

Kay Prüfer[1], Steffi Grote[1], Cesare de Filippo[1], Fabrizio Mafessoni[1], Mateja Hajdinjak[1], Petra Korlević[1], Benjamin Vernot[1], Michael Dannemann[1], Pavao Rudan[2], Željko Kucan[2], Ivan Gušic[2], Janet Kelso[1], Matthias Meyer[1], Svante Pääbo[1], The Vindija Genome Analysis Consortium[1,2]

[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Anthropology Center of the Croatian Academy of Sciences and Arts, Zagreb, Croatia

We present a 30x coverage genome from a Neandertal from the Vindija Cave in Croatia. This European Neandertal shared a most recent common ancestor with the previously published high-coverage Neandertal from Siberia around 130-140 thousand years ago (kya). Both Neandertals have a heterozygosity of ~1.5 per 10,000 bps, substantially lower than modern humans (~5-10/10,000bps), suggesting that Neandertal populations were comparatively small.

Previous studies have shown that Neandertals admixed with out-of-African populations around 50-60kya. We find that all present-day non-African populations share significantly more derived alleles with the European Neandertal than with the Asian Neandertal, indicating that the European Neandertal is more closely related to the source population(s) of the Neandertal introgression. Using the European Neandertal, we estimate that 2.0-2.9% of the genomes of present-day non-Africans trace their ancestry back to the Neandertals, somewhat higher than previous estimates.

Due to its closer relationship with the introgressing Neandertals, the high-coverage Vindija Neandertal genome provides a better proxy for studying the Neandertal gene flow into modern humans and allows us to refine our knowledge of past admixtures among Neandertals, Denisovans and modern human groups.

GENOME-WIDE CRISPR/CAS9 KNOCKOUT SCREENING AND POLY-GENOMIC INTERROGATION OF PRIMARY LEUKEMIA CELLS IDENTIFY NOVEL MECHANISMS OF GLUCOCORTICOID RESISTANCE IN PEDIATRIC B-LINEAGE ALL.

Robert J Autry[1,3], Steven W Paugh[1], Joseph R McCorkle[1], Calvin E Lau[1], Erik J Bonten[1], Jordan A Beard[1], Kristine R Crews[1], Wenjian Yang[1], Cheng Cheng[2], Deqing Pei[2], Seth E Karol[4], Kathryn G Roberts[5], Stanley Pounds[2], Charles G Mullighan[3,5], Sima Jeha[4], Ching-hon Pui[4], Mary V Relling[1,3], William E Evans[1,3]

[1]St. Jude Children's Research Hospital (SJCRH), Pharmaceutical Sciences, Memphis, TN, [2]SJCRH, Biostatistics, Memphis, TN, [3]The University of Tennessee Health Science Center, Integrated Biomedical Sciences Program, Memphis, TN, [4]SJCRH, Oncology, Memphis, TN, [5]SJCRH, Pathology, Memphis, TN

Glucocorticoids are essential components of combination chemotherapy to treat acute lymphoblastic leukemia (ALL), and resistance to glucocorticoids is associated with poor prognosis in patients with ALL. In order to understand the mechanisms underlying glucocorticoid resistance, we analyzed the prednisolone (PRED) sensitivity of primary leukemia cells from 370 pediatric patients with B-lineage ALL. To identify factors mediating PRED resistance, we interrogated multiple genomic and epigenomic features including mRNA expression, miRNA expression, DNA methylation, single nucleotide variants (SNVs) and copy number alterations (CNAs). Our analysis of these data resulted in individual signatures (56 mRNAs, 49 miRNAs, 203 CpG sites, 381 SNVs and 73 CNAs) for each of the features based on their ability to discriminate PRED resistant vs. sensitive leukemia. To determine which features were able to explain the most variability in PRED resistance, we used an integrative modeling approach which allows for integration of multiple feature types simultaneously with cross-validation and correction for bias. As an orthogonal method to validate our findings, we performed genome wide CRISPR/Cas9 knockout screening in the Nalm6 human Pre-B leukemia cell line. Resistant cells were selected by treatment with multiple concentrations of PRED (10, 100 and 500 μM). In addition to, validating previously reported mechanisms of resistance [e.g. decreased expression of GR (*NR3C1*), *SMARCA4* and *ARID1A*]; two top candidate genes (*CELSR2* and *E2F5*) were identified as highly significant in all analyses, warranting further interrogation of decreased expression in PRED resistant ALL. Confirmatory studies demonstrated that knockdown of these genes significantly increased PRED resistance in vitro. Our findings uncover new potential mediators of glucocorticoid resistance and reveal novel mechanisms for leukemic cells to escape the cytotoxic effects of glucocorticoids. Furthermore, these findings may translate to other diseases such as inflammatory/autoimmune disorders for which glucocorticoids are commonly used.

DISSECTING THE REGIONAL HETEROGENEITY AND MICROENVIRONMENT OF HUMAN GLIOBLASTOMA USING MASSIVELY PARALLEL SINGLE-CELL RNA-SEQ

Jinzhou Yuan[1], Hanna M Levitin[1], Veronique Frattini[2], Peter Canoll[3], Jeffrey N Bruce[4], Antonio Iavarone[2], Anna Lasorella[2], Peter A Sims[1]

[1]Columbia University Medical Center, Department of Systems Biology, New York, NY, [2]Columbia University Medical Center, Institute for Cancer Genetics, New York, NY, [3]Columbia University Medical Center, Department of Pathology & Cell Biology, New York, NY, [4]Columbia University Medical Center, Department of Neurological Surgery, New York, NY

Glioblastoma, the deadliest and most common form of malignant glioma, diffusely infiltrates the brain and exhibits distinct molecular signatures across different regions of the tumor. Because of the heterogeneous nature of tumor cells in glioblastoma and the complexity of the brain microenvironment, single cell analysis is essential to better characterizing progression and response to therapy in this disease. Using a massively parallel single-cell RNA-seq platform developed in our lab, we obtained thousands of high-quality single-cell transcriptomic profiles from specimens resected from both the cores and margins of primary tumors and from the cores of recurrent tumors. We assessed genetic alterations, lineage diversity, stemness, and proliferation state of individual glioma cells and detected non-neoplastic cells such as endothelial cells, pericytes, oligodendrocytes, T cells, and myeloid lineage cells. As expected, we found that both astrocytic and oligodendroglial glioma cells often co-exist within the same tumor. However, proliferating cells were highly restricted to the oligodendroglial lineage of the tumor, whereas astrocytic glioma cells were largely quiescent. In certain tumors, we observe significantly less lineage diversity among glioma cells in the diffuse margins of the tumor than in the highly cellular core, suggesting that a specific subset of glioma cells are capable of infiltrating the brain. In addition, we detected distinct subpopulations of myeloid lineage cells. Taken together, our study highlights regional heterogeneity and myeloid cell diversity in glioblastoma which have important implications for post-surgical therapy.

ADAPTIVE RESISTANCE TO CHEMOTHERAPY IN TRIPLE-NEGATIVE BREAST CANCER REVEALED BY SINGLE CELL DNA AND RNA SEQUENCING

Charissa Kim*[1,2], Ruli Gao*[1], Emi Sei[1], Rachel Brandt[1], Nicola Crosetto[3], Theodoros Foukakis[3], Nicholas Navin[1]

[1]The University of Texas MD Anderson Cancer Center, Department of Genetics, Houston, TX, [2]The University of Texas MD Anderson Cancer Center, Graduate School of Biomedical Sciences, Houston, TX, [3]Karolinska University Hospital, Department of Oncology-Pathology, Stockholm, Sweden

Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer that displays extensive intratumor heterogeneity and frequently (46%) develops resistance to neoadjuvant chemotherapy (NAC). Currently, the genomic and phenotypic evolution of chemoresistance is poorly understood in human patients. A major question is whether resistance to chemotherapy is driven by the selection of rare pre-existing subclones that confer resistance (adaptive resistance) or by the spontaneous induction of new mutations and expression changes that confer a resistant phenotype (acquired resistance). To investigate this question, we applied single cell DNA and RNA sequencing methods and deep-exome sequencing to longitudinal time-point samples collected from a cohort of 20 TNBC patients. Deep-exome sequencing of the cohort at three time-points revealed that most point mutations in clones decrease irrespective of genotype. In contrast, single cell copy number profiling of 892 cells and single cell RNA sequencing of 7137 cells in 8 TNBC patients showed that minor subclones from the pre-treatment tumors were selected and expanded in response to NAC. Our data support an adaptive resistance model in TNBC for copy number aberrations and expression programs, suggesting that it may be clinically feasible to identify chemoresistant clones in patients prior to the administration of NAC.

MOLECULAR DISSECTION OF COMPLEX GERMLINE STRUCTURAL VARIATION USING A MULTI-OMICS APPROACH

Sjors Middelkamp, Judith Vlaar, Wigard Kloosterman, Ewart Kuijk, Edwin Cuppen

UMC Utrecht, Division Biomedical Genetics, Utrecht, the Netherlands

Whole exome sequencing has greatly improved the genetic diagnosis of patients with congenital neurodevelopmental disorders such as intellectual disability in recent years. Whole genome sequencing (WGS) has the potential to increase the diagnostic yield even further, especially because it can also detect structural variants (SVs) that often cause such disorders. However, WGS also identifies many non-coding variants and SVs with unknown significance. To further improve the diagnosis of patients with potential pathogenic *de novo* SVs and to gain a better understanding of the molecular consequences of SVs we studied not only the genomes, but also the transcriptomes, epigenomes and genome organizations of patients with *de novo* germline SVs. In addition to direct effects on coding sequences of genes, SVs can have indirect, positional effects on gene regulation and expression which can not be easily inferred from WGS data. By using this multi-omics approach we can link variants of unknown significance to genes and enhancers that may have caused the phenotypes of the patients. For example, in a patient with germline chromothripsis we identified deregulation of *TWIST1*, which was not directly affected in this patient, but was located close to two breakpoint junctions, as one of the main contributors the phenotype by performing RNA-seq, Hi-C and 4C-seq on differentiated induced pluripotent stem cells derived from the patient. We demonstrate that such a multi-omics approach is a powerful method to improve the interpretation of pathogenic (complex) genomic rearrangements.

SPATIAL MAPS OF PROSTATE CANCER TRANSCRIPTOMES REVEALS AN UNEXPLORED LANDSCAPE OF HETEROGENEITY

Emelie Berglund[1], Jonas Maaskola[1], Niklas Schultz[2], Maja Marklund[1], Joseph Bergenstråhle[1], Stefanie Friedrich[3], Firas Tarish[2], Anna Tanoglidi[2], Christoph Ogris[3], Erik Sonnhammer[3], Thomas Helleday[2], Patrik Ståhl [1], Joakim Lundeberg[1]

[1]Science for Life Laboratory, Division of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden, [2]Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden, [3]Science for Life Laboratory, DBB, Stockholm University, Stockholm, Sweden

Tumor heterogeneity remains as one of the largest challenges in cancer treatments today. Prostate cancer (PCa) causes over 250,000 deaths annually worldwide and advances via clonal evolution wherein genetic and epigenetic changes cause subclonal selection. Subclonal diversity can be analyzed with next-generation sequencing (NGS) methods, including RNA-seq, single-cell RNA sequencing (scRNA-Seq) or spatial in situ methods. RNA-Seq, being based on bulk samples, do not resolve intratumour heterogeneity. ScRNA-Seq protocols faces other challenges such as dissociation of cells, limited throughput, and lack of spatial information. Thus the spatial dimension of prostate cancer transcriptomes remains under-explored.

We investigate tissue-scale gene expression heterogeneity throughout a whole multifocal prostate cancer by using the spatial transcriptomics (ST) method (Ståhl et al, Science 2016), which quantifies the transcriptome with spatial resolution in individual tissue sections. By a novel approach for deconvolution we are able to stratify the different tissue components (stroma, immune cells, PIN, cancer) into clear expression signatures that identifies subtle differences between normal as compared to stroma adjacent to tumor. We are also able to clearly distinguish areas with different Gleason scores and thereby provide a gene expression phylogeny of a prostate tumor. By comparison of transcriptome patterns with annotation by a pathologist, we note that a transcriptome-based spatial analysis confirms manual morphological stratification but also add novel information not identified by microscopy.

ULTRAFAST METAGENOMICS AT THE BENCH AND IN THE CLINIC

Aurélie Kapusta[1,2], Steven Flygare[3], Chase Miller[1,2], Yi Qiao[1,2], Gabor Marth[1,2], Edgar J Hernandez[1,2], Qing Li[1], Helena Safavi-Hemami[4], Samuel D Robinson[4], Aiping Lu[4,5], Baldomero Olivera[4], Guochun Liao[3], Martin G Reese[3], Robert Schlaberg[3,6,7], Mark Yandell[1,2,3]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]USTAR, Center for Genetic Discovery, Salt Lake City, UT, [3]IDbyDNA Inc, San Francisco, CA, [4]University of Utah, Department of Biology, Salt Lake City, UT, [5]Tongji University, Institute of Protein Research, Shanghai, China, [6]University of Utah, Department of Pathology, Salt Lake City, UT, [7]ARUP, Institute for Clinical and Experimental Pathology, Salt Lake City, UT

NGS-based metagenomics has the potential to revolutionize both basic science and clinical medicine. Long computation times, however, have hindered adoption. In response, researchers at the University of Utah, the Centers for Disease Control and Prevention (CDC), ARUP Laboratories, and IDbyDNA Inc. have developed Taxonomer [1] – an ultrafast engine for comprehensive metagenomics data analysis and interactive results visualization (www.taxonomer.com). Taxonomer is unique in providing integrated nucleotide and protein-based classification, and simultaneous host RNA transcript profiling. And it is blindingly fast: the latest development-version of Taxonomer can search and classify a typical Illumina RNA-seq dataset (~20 million reads) against the entirety of the Greengenes [2] 16s database (~203,000 16s rRNA sequences), UNIREF90 (47 million protein sequences), and profile expression of every coding and non-coding transcript of any organism in less than 90 seconds using 100 threads. Taxonomer thus opens new avenues for basic research applications. We will highlight this aspect with our discoveries of novel hormones, enzymes, and new superfamilies of conotoxins – small peptides which are a focus of worldwide pharma discovery efforts and source for new FDA approved drugs. Medical applications for ultra-fast metagenomics also abound. We recently published benchmarks demonstrating that Taxonomer's ability to identify pathogens in patient samples rivals that of an FDA-approved PCR panel [3]. Using actual (and often unusual) clinical cases, we will show how Taxonomer is already playing a major role in hospitals across the country. We will also provide details on the upcoming nationwide role out by ARUP and IDbyDNA of the Taxonomer-driven ARIA test, the first-of-its-kind commercial NGS test for respiratory infections. NGS has already revolutionized genetics; now ultrafast metagenomics is about to revolutionize comparative genomics and treatment of infectious diseases. Our results provide a preview.

[1] Flygare, Simmon et al. (2016). Genome Biology
[2] http://greengenes.secondgenome.com/
[3] Graf et al. (2016). J Clin Microbiol

NATURAL SELECTION AND LOCAL RECOMBINATION RATES SHAPE THE GENOME EVOLUTION OF SWORDTAIL HYBRIDS

Molly Schumer

Harvard University / Columbia University, New York, NY

How distinct species persist in the face of gene flow is a long-standing and central question in evolutionary biology, reinvigorated by the recent realization that hybridization is surprisingly common. Here, we address this question by generating a fine-scale genetic map for the swordtail fish, Xiphophorus birchmanni, and by analyzing genome sequences from three, independent naturally-occurring hybrid populations of X. birchmanni and X. malinche. Analyzing variation in ancestry proportions along the genome, we see clear evidence that local recombination rates are a key determinant of ancestry proportions. Moreover, we uncover genomic patterns consistent with the joint action of selection against "hybrid incompatibilities" but also of selection against mildly deleterious mutations introduced by hybridization. Thus, we provide genomic evidence for theoretical predictions about the roles that recombination and selection play in constraining the evolution of hybrid genomes.

A WOLF IN SHEEP’S CLOTHING: A SELFISH ELEMENT DISGUISED AS A LINKED PAIR OF DEVELOPMENTAL GENES UNDERLIES A GENETIC INCOMPATIBILITY IN *C. ELEGANS*

Eyal Ben-David, Alejandro Burga, Leonid Kruglyak

Department of Human Genetics, Department of Biological Chemistry and Howard Hughes Medical Institute, UCLA, Los Angeles, CA

Selfish genetic elements promote their own transmission while being neutral or detrimental to the fitness of the organism. In extreme cases, selfish elements promote their transmission by killing individuals that do not inherit them, leading to a genetic incompatibility between carriers and non-carriers. Genetic incompatibilities are found across the tree of life, but in only a few cases have the underlying genetic mechanisms been resolved. We discovered a novel genetic incompatibility between strains of the nematode *Caenorhabditis elegans*. The incompatibility is caused by a selfish element composed of two genes: a maternal-effect toxin and a zygotically expressed antidote. In crosses between strains that carry the element and ones that do not, the element acts via maternal effect to kill the progeny that do not inherit it. We identified the genes underlying the incompatibility, and were surprised to find that the toxin and the antidote were encoded by *sup-35* and *pha-1*, respectively. *pha-1* was originally proposed to be an essential developmental gene due to specific pharyngeal defects observed in mutants, and its defects were known to be fully suppressed by mutations in *sup-35*. Our results indicate that the phenotypes previously associated with mutations in *pha-1* are in fact a consequence of *sup-35* toxicity, which is rescued by *pha-1*. The lethality associated with *pha-1* mutations is known to require additional genes, suggesting that the *sup-35/pha-1* selfish element exerts its toxicity by hijacking a conserved developmental pathway. We *de novo* assembled the haplotypes of strains not carrying an active *sup-35/pha-1* element using a combination of Illumina short reads and Oxford Nanopore long reads, and found structural variation and remarkable nucleotide divergence, illustrating the ability of selfish elements to reshape the genome. We also identified strains carrying loss of function variants in *sup-35* that abolish the incompatibility, further supporting its role as a toxin rather than a developmental gene. Our results illustrate how the study of natural genetic variation can illuminate our understanding of gene function. Furthermore, our results suggest that selfish elements conferring genetic incompatibilities may be more common than previously thought, and that some of them may be hiding in plain sight, disguised as developmental genes.

DNA SEQUENCING OF SINGLE SPERM USING A NOVEL APPROACH FOR WHOLE GENOME AMPLIFICATION PROVIDES CRITICAL INSIGHTS INTO MEIOSIS AND RECOMBINATION

Anjali Hinch[1], Gang Zhang[1], Philipp Becker[1], Ben Davies[1], Daniela Moralli[1], Cath Green[1], Rory Bowden[1], Peter Donnelly[1,2]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]University of Oxford, Department of Statistics, Oxford, United Kingdom

We have developed a novel and reliable approach to whole-genome amplification from ultra-low DNA inputs which improves on currently available methods. This approach was used to allow whole-genome DNA sequencing from 217 individual mouse sperm, providing a new and powerful high-throughput approach to directly study meiotic recombination events. We detected more than 2500 crossovers, with the majority of these localised to within 1 kb. Nearly 400 crossovers were resolved to within 250bp, providing an unprecedented fine-scale view of crossover resolution in a mammal. Over 96% of all crossover events overlap known recombination hotspots. Comparison of stage-specific recombination markers H3K4Me3 and DMC1 with crossovers in the same inbred mouse strain provides a powerful tool with which to unpick various aspects of meiosis. While it has long been known that only a minority (~10%) of the double-strand breaks (DSBs) which initiate recombination are resolved as crossovers (with the remainder as non-crossovers), the factors involved in this decision are not understood. Our data suggest that DSBs which are first to pair with their homologous chromosome are more likely to be resolved as crossovers, that distance from the centromere tends to decrease homologue pairing times, and that hotspots where PRDM9 is likely to have bound to both homologous chromosomes tend to have lower repair times. We show that repositioning of nucleosomes by PRDM9 on the non-initiating, template, chromosome affects the lengths of repair tracts. Our results also shed novel light on the nature and functional definition of the pseudo-autosomal region (PAR) and the constraints on recombination within it.

BACKGROUND SELECTION IS THE DOMINANT MODE OF LINKED SELECTION IN HUMANS

David Murphy, Guy Sella

Columbia University, Biological Sciences, New York, NY

Analysis of genetic variation data across a range of taxa have conclusively demonstrated the importance of selection at linked sites in shaping levels of neutral diversity in the genome. However, much about the process remains unknown, including which genomic features underlie the selective effects. With the recent availability of detailed genomic annotations and fine-scale genetic maps, we can now address these questions by fitting models of linked selection to genome-wide variation data. We do so here, by fitting a joint model of positive selection ("selective sweep") and purifying selection against linked deleterious mutations ("background selection") to polymorphism data from the human 1000 genomes project. We show that the model of background selection alone can accurately predict diversity levels across the human genome, explaining more than 40% of the variance on the 1Mb scale, and that adding effects of selective sweeps does not increase the fit. In turn, the parameter estimates obtained from fitting the model suggest that more than 80% of strongly deleterious mutations occur in non-coding regions and that the best proxy for these regions is phylogenetic conservation in primates. Our findings further point to the existence of a class of functional sites that are human-specific (in that they are not conserved across primates), contribute to background selection effects, but are not captured by broad genomic annotations. More generally, our results indicate that background selection is the dominant mode of linked selection in humans and that it has a marked effect on diversity levels in most of the genome.

EVIDENCE FROM THE BOVINE OF MAJOR DIFFERENCES BETWEEN INDIVIDUALS IN THE RATE OF *DE NOVO* SINGLE NUCLEOTIDE MUTATION AND TRANSPOSON MOBILIZATION IN THE GERM-LINE

Chad Harland[1,2], Keith Durkin[1], Maria Artesi[1], Latifa Karim[1,3], Nadine Cambisano[1,3], Manon Deckers[1,3], Nico Tamma[1,3], Erik Mullart[4], Wouter Coppieters[1,3], Michel Georges[1], Carole Charlier[1]

[1]Unit of Animal Genomics, GIGA Research, University of Liège, Liège, Belgium, [2]Livestock Improvement Corporation, Research & Development, Hamilton, New Zealand, [3]GIGA-Genomics Platform, GIGA Research, University of Liège, Liège, Belgium, [4]CRV, Research & Development, Arnhem, the Netherlands

To study the process of *de novo* mutations in the bovine germ line, we have sequenced the whole genome of >750 individuals constituting 130 sire-dam-offspring trios with at least five grand-offspring each. A first study using four pedigrees revealed the common occurrence of somatic and germ-line mosaicism for *de novo* mutations pointing towards mutation-prone early cleavage cell divisions (http://biorxiv.org/content/early/2016/10/09/079863). We herein characterize *de novo* mutations in the remaining 126 pedigrees. Two observations point towards major inter-individual differences in the rate of *de novo* mutations.

We first identify one sire characterized by a mutation rate that is ~3-fold larger than the population average. We show that this remarkable increase is due to a ~7-fold excess of mutations occurring at the very early stages of development (on the basis of observed mosaicism). The corresponding mutations are characterized by a ~8-fold excess in C to T transitions outside the CpG context. The corresponding animal was shown to be the only individual of the pedigree to be homozygous for a rare disruptive mutation in components of the DNA repair or replication machinery: a P>L substitution in the REV1 DNA Directed Polymerase. The causality of this mutation is presently being examined.

We further developed a pipeline to detect *de novo* transposition and pseudogene mobilization events. We identified a family of LTR elements that are still active in the bovine genome. We detected five corresponding *de novo* transposition events, of which three occurred in the same individual including two in the same gamete.

Latest results of both studies will be presented.

RAPID EVOLUTION OF THE HUMAN MUTATION SPECTRUM

Kelley Harris[1], Jonathan K Pritchard[1,2,3]

[1]Stanford University, Genetics, Stanford, CA, [2]Stanford University, Biology, Stanford, CA, [3]Howard Hughes Medical Institute, Chevy Chase, MD

DNA is a remarkably precise medium for copying and storing biological information, with a mutation rate in humans of about 1e-8 per base pair per generation. This extraordinary fidelity results from the combined action of hundreds of genes involved in DNA replication and proofreading, and repair of spontaneous damage. Recent studies of cancer have shown that mutation of specific genes often leads to characteristic mutational "signatures"--i.e., increased mutation rates within particular sequence contexts. We therefore hypothesized that more subtle variation in replication or repair genes within natural populations might also lead to differences in mutational signatures. As a proxy for mutational input, we examined SNV variation across human and other great ape populations. Remarkably we found that mutational spectra differ substantially among species, human continental groups and even, in some cases, between closely-related populations. Closer examination of one such signal, an increased rate of TCC>TTC mutations reported previously in Europeans, indicates a burst of mutations from about 15,000 to 2,000 years ago, perhaps due to the appearance, drift, and ultimate elimination of a genetic modifier of mutation rate. Our results suggest the possibility of mapping modifiers of mutation rates within human populations and across species.

RESISTANCE TO MALARIA THROUGH STRUCTURAL VARIATION OF RED BLOOD CELL INVASION RECEPTORS

Ellen M Leffler[1,2], Gavin Band[1,2], George B Busby[1], Katja Kivinen[2], Quang S Le[1], Geraldine M Clarke[1], Christina Hubbart[1], Anna E Jeffreys[1], Kate Rowlands[1], Kirk A Rockett[1,2], Chris C Spencer[1], Dominic P Kwiatkowski[1,2], Malaria Genomic Epidemiology Network[1,2]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Plasmodium falciparum invades human red blood cells by a series of interactions between host and parasite surface proteins. Here we analyse whole genome sequence data from worldwide human populations, including 765 new genomes from across sub-Saharan Africa, and identify a diverse array of large copy number variants affecting the host invasion receptor genes *GYPA* and *GYPB*. We find that a nearby reported association with severe malaria is explained by a complex structural variant that involves the loss of *GYPB* and gain of two hybrid genes, each with a GYPB extracellular domain and GYPA intracellular domain. We show that this structural variant encodes the Dantu blood group antigen, reduces the risk of severe malaria by 40% and has recently risen in frequency in parts of Kenya. Finally, we analyse 119 non-human primate genomes at this locus, finding widespread copy number variation across African great apes and rapid evolution of extracellular sequences. These findings demonstrate that structural variation of these red blood cell invasion receptors is associated with natural resistance to *P. falciparum* malaria in humans and suggest they may be involved in repeated host-pathogen interactions across apes.

SEEDERSEEKER: A COMPUTATIONAL ALGORITHM FOR RECONSTRUCTING METASTATIC EXPANSION AT A SUBCLONAL LEVEL

Yi Qiao[1,2], Xiaomeng Huang[1,2], Gabor Marth[1,2]

[1]University of Utah, Eccles Institute of Human Genetics, Salt Lake City, UT, [2]USTAR Center for Genetic Discovery, Salt Lake City, UT

In most cancers, metastasis is the major cause of treatment failure and patient death. Understanding metastatic tumor evolution at a subclonal level is likely to offer vital insight into mechanism. The identification of aggressive subclones responsible for metastatic colonization into distal organs offers the possibility to preferentially target these subclones, rather than more benign groups of cells within the tumor.

The reconstruction of metastatic tumor evolution from bulk sequencing data is complicated by at two main factors: First, as in the analysis of longitudinal samples collected during cancer recurrence at a single tumor site, the identity and cellular frequency of the subclones (i.e. groups of cell with a specific complement of somatic mutations) present in each consecutive biopsy must be reconstructed from the bulk allele frequency measurements. Second, whereas longitudinal tumor samples have inherent time-ordering according to when each biopsy is collected, metastatic sites are typically removed at a single timepoint, as part of a rapid autopsy procedure hours after the patient’s death, and therefore the order of metastatic seeding and colonization across the sites has to be inferred directly from the data.

Here we present the SeederSeeker algorithm, a major extension of our published and popular SubcloneSeeker program. SeederSeeker (1) reconstructs subclonal composition at each metastasis; and (2) establishes the most likely order of colonization across all such sites, as well as the primary tumor, if available. SeederSeeker uses bulk somatic SNV allele frequency measurements, copy number variation (CNV) profiles, and loss of heterozygosity (LOH) estimates as input. We first extended the SubcloneSeeker algorithm currently capable of reconstructing subclones in one or two tumor sites, to reconstruct subclones at an arbitrarily large number of sites. Second, the algorithm establishes partial time ordering across pairs of subclones using the natural evolution of chromosomal events; and pairwise seeder - seeded tumor site relationships from inferred clonal allele frequencies of subclones shared between the pair. Third, seeding order across all tumor sites, facilitated by specific subclonal colonization events, are inferred from the pairwise relationships. We demonstrate the application of our algorithm for reconstructing subclonal metastatic expansion in data from multiple breast and ovarian cancer patients.

DETECTING POLYGENIC ADAPTATION IN AN ADMIXTURE GRAPH

Fernando Racimo[1], Jeremy J Berg[2], Joseph K Pickrell[1,2]

[1]New York Genome Center, New York, NY, [2]Columbia University, Department of Biological Sciences, New York, NY

It is now apparent that much of recent human phenotypic evolution may have occurred via polygenic adaptation (PA): small and concerted shifts in allele frequencies at several loci affecting a complex trait. In recent years, several methods have been developed to detect PA using SNP effect size estimates from GWAS data. Though powerful, these methods suffer from limited interpretability: they can detect which sets of populations have evidence for PA, but are unable to reveal where in the history of multiple populations these processes occurred. To address this, we created a method to detect PA in an admixture graph, which is a representation of the historical divergences and admixture events relating different populations through time. We have developed a MCMC algorithm to obtain posterior distributions of branch-specific parameters reflecting the strength of selection in each branch of a graph. We also developed a set of summary statistics that are fast to compute and can indicate which branches are most likely to have experienced PA. This, in turn, helps us reduce the possible space of candidate branches in our MCMC. We show via simulations that we have good power to detect PA in complex graphs with trait-affecting SNPs from published GWAS data. We also applied our method to human population genomic data from around the world, to determine when and where PA for a variety of anthropometric, metabolic and neurological traits occurred during recent human evolution.

A GENOME-WIDE INTERACTOME OF DNA-ASSOCIATED PROTEINS IN THE HUMAN LIVER

Ryne C Ramaker[1,2], Daniel Savic[1], Andrew A Hardigan[1,2], Gregory M Cooper[1], Richard M Myers[1], Sara J Cooper[1]

[1]HudsonAlpha Institute for Biotechnology, Genetics, Huntsville, AL, [2]University of Alabama at Birmingham, Genetics, Birmingham, AL

Large-scale efforts like the Encyclopedia of DNA Elements (ENCODE) Project have made tremendous progress in cataloging the genomic binding patterns of DNA-associated proteins (DAPs), such as transcription factors. However most chromatin immunoprecipitation-sequencing (ChIP-seq) analyses have focused on a few immortalized cell lines whose activities and physiology deviate in important ways from endogenous cells and tissues. Consequently, binding data from primary human tissue are essential to improving our understanding of in vivo gene regulation. Here we analyze ChIP-seq data for 20 DAPs assayed in two healthy human liver tissue samples, identifying more than 300,000 binding sites. We integrated binding data with transcriptome and phased whole genome data to investigate allelic DAP interactions and the impact of heterozygous sequence variation on the expression of neighboring genes.

We compared our genome-wide DAP occupancy data to that of a widely studied cancer cell line, HepG2. After integration with data from the Genotype-Tissue Expression (GTEx) project we found our primary tissue data showed higher enrichment for liver-specific expression quantitative trait loci (eQTL) and proximal binding near transcripts expressed primarily in liver tissue. Using tumor expression and whole genome mutation data from The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) projects, we have demonstrated how our tissue-derived DAP occupancy data can be complementary to existing cancer cell line-derived data in identifying regulatory somatic mutations and disruptions of normal liver expression programs.

Lastly, we used a machine learning-based approach for using our DAP occupancy data to prioritize impactful non-coding variation. With this approach, we ranked all liver eQTL SNPs based on their likelihood to disrupt assayed DAP binding or induce cryptic binding. This approach was validated with allele bias data and in vitro reporter assays in HepG2 cells. Overall, our work provides a valuable resource for investigation of genomic regulatory regions in liver and highlights the value of performing ChIP-seq in primary tissue as a complement to cell culture systems.

BENCHMARKING RNA-SEQ ANALYSIS IN PLANT SPECIES

Srividya Ramakrishnan[1], Fritz Sedlazeck[1], Michael Schatz[1,2]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY

High-throughput RNA sequencing is well-established as a versatile platform with the potential to study the complex transcriptomics of many organisms across the tree of life. Most RNA-Seq experiments aim to build a catalogue of species transcripts, assemble them and quantify the changing expression levels of transcripts during different stages of development or under different conditions. Despite the fact that a large number of mapping algorithms have been developed for RNA-seq read mapping, transcriptome assembly and differential analysis in recent years, accurate alignment of RNA-seq reads is still challenging and yet unsolved problem because of exon-exon spanning junction reads, relatively short read lengths and the ambiguity of multiple-mapping reads. In addition to these inherent complexities, RNA-seq within plant species is especially problematic for a variety of reasons. One of the biggest challenges is plant gene annotations, which are used during the aligning step and the assembly step, are far from complete and the annotations of many genes that are particularly unique to plants, are still of poor quality. Furthermore, the genomes of many plant species are lower quality draft sequences with gene families collapsed or entirely missing. These factors significantly impact the downstream gene expression estimates in plants. In this study, we evaluate the performance of several of the most popular RNA-Seq aligners, including Tophat2, HiSat2 and STAR, and the assembly and quantification of alignments from these aligners using Cufflinks and StringTie on both real and simulated RNA-Seq data from plants. Additionally we also assess these tool’s performance on gene abundance estimation and differential gene expression based on the presence/absence of complete gene annotations.

TRANSCRIPTIONAL PROFILING OF AGING EFFECTS IN HUMAN TRABECULAR MESHWORK

Shweta Ramdas[1], Li Guan[1], Qianyi Ma[2], Frank Rozsa[4], Julia Richards[4], Jun Z Li[2,3]

[1]University of Michigan, Bioinformatics Graduate Program, Ann Arbor, MI, [2]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [3]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [4]University of Michigan, Department of Ophthalmology and Visual Sciences, Ann Arbor, MI

The trabecular meshwork (TM) is an area of tissue in the outer eye, consisting of a network of fibers involved in the outflow of aqueous humor from the cornea. Increasing stiffness of this tissue with age is a risk factor for glaucoma, one of the leading causes of blindness in the world. In this study, we use microarray-based gene expression profiling of post-mortem samples from 93 subjects (age: 13-88 yrs) to analyze age-related gene expression changes in the TM. We identify a transcriptional signature of age in this tissue, involving genes such as SPOPL, CASC15 and TGM2. The TGM2 protein is responsible for cross-linking of extracellular proteins, making them resistant to degradation. Previous studies have reported higher levels of TGM2 (transglutaminase 2) in glaucoma samples than in controls. The age-related increase of TGM2 therefore could be related to the increasing stiffness of the TM with age and a risk factor for glaucoma. Comparisons with published aging signatures in the rat reveal convergence at the level of biological pathways, including those related to the immune system, the extracellular space and the ribosome. Genes in these pathways tend to have higher expression in older subjects. While aging studies of other tissues show a decreased expression of genes for ribosomal function, TM appears to be one of the few tissues showing an increase with age. Further, more genes show a higher variance among older samples than those showing higher variance in the young. This greater inter-individual variation in older samples is consistent with a decreased stringency of gene regulation with advancing age, as proposed by previous studies. Lastly, we were able to develop gene expression predictors of age using subsets of our samples as training sets, achieving a performance of correlation coefficient r of 0.68 between the predicted and actual age in the test set.

POOLED CRISPR SCREENING WITH SINGLE-CELL TRANSCRIPTOME READOUT

Paul Datlinger[1], Andre F Rendeiro[1], Christian Schmidl[1], Thomas Krausgruber[1], Peter Traxler[1], Johanna Klughammer[1], Linda C Schuster[1], Amelie Kuchler[1], Donat Alpar[1], Christoph Bock[1,2,3]

[1]CeMM Research Center for Molecular Medicine, Bock lab, Vienna, Austria, [2]Medical University of Vienna, Department of Laboratory Medicine, Vienna, Austria, [3]Max Planck Institute, Informatics, Saarbrücken, Germany

CRISPR-based genetic screens are accelerating biological discovery, but current methods have inherent limitations. Widely used pooled screens are restricted to simple readouts including cell proliferation and sortable marker proteins. Arrayed screens allow for comprehensive molecular readouts such as transcriptome profiling, but at much lower throughput. We have combined pooled CRISPR screening with single-cell RNA sequencing into a broadly applicable workflow, directly linking guide RNA expression to transcriptome responses in thousands of individual cells. Our method for CRISPR droplet sequencing (CROP-seq) enables pooled CRISPR screens with single-cell transcriptome resolution, which will facilitate high-throughput functional dissection of complex regulatory mechanisms and heterogeneous cell populations.

TRANSCRIPTION IS TIGHTLY COUPLED TO CHROMATIN ORGANISATION AND IS PREDICTIVE OF CELL SPECIFIC INTERACTIONS

Sarah Rennie, Robin Andersson

University of Copenhagen, Section for Computational and RNA Biology, Copenhagen, Denmark

The three dimensional organisation of a genome within a nucleus appears to be crucial for correct transcriptional activity. However, it is unclear to which extent transcriptional activity is attributable to chromosomal position, chromatin organisation, or gene-specific regulatory programs. Here, we show that information within transcription data alone is highly descriptive of chromatin organisations. We demonstrate this by constructing a model that computationally decomposes RNA expression data into two main components; one reflecting the positional relationship of neighbouring transcription units across chromosomes, and a further independent component reflecting the transcription levels not attributable to their relative positioning.

We demonstrate that this underlying positional component not only reflects a significant proportion of transcription, but also is highly informative of topological domain organisation; predicting boundaries and chromatin compartments. Furthermore, when integrated with enhancer and promoter locations, transcriptional components can accurately predict individual proximity interactions as defined by chromatin conformation capture data. This demonstrates a close coupling between transcriptional output and proximity. Intriguingly, our models are accurate at predicting chromatin interactions across a range of distances, both within and between chromatin domains. Different transcriptional attributes appear to drive long- and short-range interactions, with enhancer transcription being a strong predictor of distal interaction capabilities.

We apply our method to several cell types and tissues with CAGE data, predicting transcriptionally defined domain boundaries and proximity interactions between pairs of active regions, as well as chromatin compartment shifts at key identity genes. In all, we present a resource that demonstrates a close relationship between active transcription and higher order regulatory organisations.

CHROMATIN REORGANIZATION EVENTS IN ENDOTHELIAL CELLS ASSOCIATED WITH PULMONARY ARTERIAL HYPERTENSION

Armando Reyes-Palomares[1], Fabian Grubert[2], Mingxia Gu[3], Ivan Berest[1], Michael P Snyder[2], Marlene Rabinovitch[3], Judith B Zaugg[1]

[1]European Molecular Biology Laboratory, SCB unit, Heidelberg, Germany, [2]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [3]Stanford University School of Medicine, Department of Pediatrics, Stanford, CA

Pulmonary arterial hypertension is a rare condition characterized by high-blood pressure in the arteries that supply blood to the lungs. Here we aim to study the epigenetic and gene expression changes in primary pulmonary arterial endothelial cells from patients and controls. This study was performed taking into account the stratification of patients according to their disease causes, idiopathic or familial (BMPR2 mutation). Our multi-omics' approach consisted on the integration of diverse molecular phenotypes such as histone modifications (H3K27ac), gene expression (RNA-seq) and long-range chromatin interactions (RAD21 and CTCF). The multi-level integration allowed us to uncover disease-specific chromatin regulatory domains that can be entirely up-/down-regulated by a few transcription factors. The multi-level integration allowed us to uncover disease-specific chromatin regulatory domains that can be entirely up-/down-regulated by a few transcription factors. We used these chromatin modules as the basis for building a comprehensive regulatory network to understand the molecular mechanism underlying hypertension injury and vascular remodelling in endothelium. Our results suggest specific-epigenetic changes related to transcription factor activities and chromatin reorganization events that are sensitive for hypertension injury. These findings will be useful for getting a deeper insight of the genetic causes and the development of intervention therapies.

REGULATION OF NODULATION BY A LARGE FAMILY OF NODULE-SPECIFIC CYSTEINE RICH (NCR) PEPTIDES IN *MEDICAGO TRUNCATULA*

Mingkee Achom[1], Charlotte Rich[1], Sascha Ott[2], Miriam Gifford[1]

[1]University of Warwick, School of Life Sciences, Coventry, United Kingdom, [2]University of Warwick, Department of Computer Sciences, Coventry, United Kingdom

Legume plants form symbiotic relationships with nitrogen-fixing rhizobia in specialised polyploid cells forming nodules. Within these nodules, the rhizobia 'fixes' atmospheric nitrogen into useable ammonia which is required for plant growth. The key genes involved in the nodulation process are well characterized, but much less well understood are the regulatory functions that balance nodule development with other root development pathways such as lateral root development.

Medicago truncatula, a model legume with an indeterminate nodule-type has been found to express a large family of nodule cysteine rich peptides (NCRs) (Nallu et al., 2013) during different stages of nodulation. The size and sequence diversity of NCRs, coupled with their distinct spatial and temporal expression profiles has led to the idea that NCRs are signaling molecules with multiple functions in the control of nodulation and root development. Initial studies of NCRs have suggested some potential functions for individual NCRs, but the prevalence of such as large family of NCR is.

We performed gene expression analysis comparing wild-type with a hyper-nodulating mutant under different nitrogen conditions which revealed striking patterns of NCR expression during nodulation and nitrogen influx. Since these genes are strongly rhizobia and nitrogen regulated, we suggest that NCRs may act to regulate nodule numbers based on plant nitrogen status. To further understand the regulation of NCRs, we used *de novo* motif discovery and multiple alignment methods to discover highly conserved regions in the promoters of NCR genes which we will use to identify potential key regulators of NCRs and thus nodulation.

COMBINATORIAL RECOGNITION OF DNA BY BZIP TRANSCRIPTION FACTORS

Jose A Rodriguez-Martinez

University of Puerto Rico - Rio Piedras, Biology, San Juan, PR

Transcription factors rarely function by binding DNA as monomers. However, how transcription factor dimerization impacts their DNA binding specificity is poorly understood. bZIP transcription factors are obligate dimers that regulate gene expression in a myriad of cellular process. Guided by bZIP dimerization properties, we examined DNA binding specificities of 270 human bZIP pairs. DNA interactomes of for 102 bZIP dimers revealed that 72% of heterodimer motifs correspond to conjoined half-sites preferred by partnering monomers. Remarkably, the remaining motifs are composed of variably-spaced half-sites (12%) or “emergent” sites (16%) that cannot be readily inferred from half-site preferences of partnering monomers. Identified cognate sites were confirmed by binding assays. Furthermore, bZIP heterodimer-preferred sites were overrepresented in genomic regions co-occupied by two bZIP proteins as determined by ChIP-seq peaks. Focusing on ATF3, we observed distinct cognate site preferences conferred by different bZIP partners, and demonstrated that genome-wide binding of ATF3 is best explained by considering many dimers in which it participates. Notably, our compendium of bZIP-DNA interactomes were applied to predict the impact of non-coding mutations on bZIP-DNA binding.

FULL-LENGTH SEQUENCING OF CANCER GENE FUSIONS USING RNA-CAPTURE AND KILOBASE –LENGTH READS

Jeffrey A Rosenfeld[1], Sara Goodwin[2], Shridar Ganesan[1]

[1]Rutgers University, Cancer Institute of New Jersey, New Brunswick, NJ,
[2]Cold Spring Harbor Laboratory, Genome Center, Cold Spring Harbor, NY

Cancer genomes are full of novel proteins that are created from the fusion of different genes. These non-native proteins are often extremely harmful to the patient and are a driver of oncogenesis and tumor growth. There are current techniques using PCR coupled with short read sequencing that attempt to identify these proteins, but they are limited by short length of the reads. Junction locations can be detected, but it is difficult to obtain the full context of the fusion.

We designed capture probes for all kinases and a set of other proteins that are known to be involved in cancer fusions. These enrichment probes were used to selectively enrich RNA for sequencing. After enrichment, we used Oxford Nanopore long read sequencing to obtain reads of 5kb to 10kb which captured the entire fusion gene sequence. Each flow cell gave us several gigabytes of data. We validated our approach by using cell lines with known fusions including. TMPRSS2-ERG in prostate cancer and BCR-ABL in Acute myeloid leukemia (AML). For these validation samples, we have found that we capture the junction location along with the full surrounding protein. We have also started to look at primary patient samples to determine the sequence context of the fusions.

DISSECTING THE GENETIC AND EVOLUTIONARY BASES OF SPLICE SITE USAGE IN THE HUMAN IMMUNE RESPONSE

Maxime Rotival, Hélène Quach, Etienne Patin, Guillaume Laval, Christine Harmant, Nora Zidane, Lluis Quintana-Murci

Human Evolutionary Genetics Unit, CNRS URA3012, Institut Pasteur, Paris, France

The immune system is one of the most adaptive systems that exist. Recent studies have shown that inter-individual and between-population variation in transcriptional responses to immune stimulation are, to a large extent, accounted for by regulatory variants (response eQTL), which have been, in turn, preferential targets of positive selection during recent human evolution. Yet, the extent to which genetic variants affect splicing in response to immune stimulation, and the adaptive potential of these variants in the human population remain largely unexplored. Here, we took advantage of 970 RNA-Seq profiles generated in primary monocytes from 200 individuals of African- and European-descent, which were coupled to genome wide SNP and exome data, at the resting state and following stimulation of major innate immunity pathways (TLR4, TLR1/2, TLR7/8) and Influenza A virus infection. First, we identified 1,354 genes (26% of all tested genes) whose splicing patterns are altered upon immune stimulation, and showed that these patterns strongly differ depending on population ancestry, i.e., 799 genes presented differential splicing between populations upon stimulation, including essential immune regulators such as NOD1, IRF1 or NFKB1. Focusing on the genetic determinants of splicing, we found that 14% of genes have a splice QTL, with events of exon skipping and last exon switching being enriched in genetic control (17%, $P<0.002$). We also report cases of splice sites whose usage has been the target of positive selection, in genes such as NADSYND1, associated to vitamin D deficiency (Fst=0.65, |iHS|=2.7) or CTSH, associated to type I diabetes (Fst=0.59, |iHS|=3.4), or has been altered by Neanderthal introgression, in genes such as OAS1, CAST, RAB5A, or FCGR2A. Looking at the conservation of splice sites active in immune cells, we further showed that while 98% of splice sites are conserved across species (GerpRS>2), 17% of immune response genes constitutively use at least one splice site that is not conserved, indicating that splicing variation of immune genes participates to their strong adaptability. Lastly, our analysis revealed that the usage of non-conserved, unannotated splice sites is increased upon stimulation at the genome wide level ($P<10^{-16}$), independently of gene expression, but decreased in cytokines (e.g., IFNG, IL12A/B or CCL20) and genes involved in response to bacterium (e.g., MYD88, IRAK3, or LYZ), suggesting mechanisms for increased quality control, and constraint, of splicing at immune genes upon stimulation. Together, this study increases our understanding of how regulatory variants influence immune responses at the splicing level, and how these variants have participated to population adaptation at different time scales, highlighting the important, adaptive role of splicing in reshaping host defense mechanisms against pathogens.

CTCF-MEDIATED INTRAGENIC CHROMATIN LOOPING REGULATES ALTERNATIVE EXON USAGE

Mariana Ruiz-Velasco[1], Manjeet Kumar[1], Mang C Lai[1], Pooja Bhat[2], Ana B Solis-Pinson[1], Alejandro Reyes[3], Kyung M Noh[4], Toby Gibson[1], Judith B Zaugg[1]

[1]European Molecular Biology Laboratory, Structural and computational biology, Heidelberg, Germany, [2]Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna, Austria, [3]Dana-Farber Cancer Institute, Biostatistics and Computational Biology, Boston, MA, [4]European Molecular Biology Laboratory, Genome Biology, Heidelberg, Germany

The three-dimensional structure of the DNA has been proposed to play a major role in gene regulation. However, despite our current knowledge about the formation of long range DNA-contacts, which are mainly established by the CCCTC-binding factor (CTCF), our understanding of the functional impact of chromatin structure on downstream processes, such as gene expression, remains largely descriptive with little mechanistic insight. Here we propose a novel mechanism of how CTCF-mediated intragenic chromatin-loops can regulate alternative exon usage. We validate our model for alternative exon usage across 18 individuals in lymphoblastoid cell lines by integrating RNA-Seq data with ChIP-Seq for CTCF and PolII, cell-line-specific chromatin contact maps, and matching genotype information. Overall, our results provide strong evidence that CTCF-mediated chromatin looping is regulating alternative exon usage, which seems driven by genetic variation in CTCF binding sites across individuals. Furthermore, we find that the exons with the potential of being CTCF-regulated are more likely to disrupt annotated protein domains and particularly enriched for being involved in signalling and cellular stress-response pathways. By integrating multiple levels of molecular phenotype data, our study provides strong evidence for alternative exon usage being regulated by chromatin structure, and thus increases our understanding of functional consequences underlying variation in chromatin architecture.

DE NOVO GENE EVOLUTION: STUDYING THE GENE TRANSITION FROM NON-CODING TO CODING.

Jorge Ruiz-Orera[1], José Luis Villanueva-Cañas[1], William R Blevins[1], M. Mar Albà[1,2,3]

[1]Evolutionary Genomics Group, Research Programme in Biomedical Informatics, Hospital del Mar Research Institute (IMIM), Barcelona, Spain, [2]Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3] Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain

Recent years have witnessed the discovery of many genes which appear to have evolved *de novo* from previously non-coding sequences. This has changed the long-standing view that coding sequences can only evolve from other coding sequences. However, there are still many open questions regarding how new protein-coding sequences can arise from non-genic DNA.

Two prerequisites for the birth of a new functional protein-coding gene are that the corresponding DNA fragment is transcribed and that it is also translated. Transcription is known to be pervasive in the genome, producing a large number of transcripts that do not correspond to conserved protein-coding genes, and which are usually annotated as long non-coding RNAs (lncRNA). Recently, sequencing of ribosome protected fragments (Ribo-Seq) has provided evidence that many of these transcripts actually translate small proteins, therefore suggesting that not only transcription, but also translation, is a pervasive process. We have used mouse non-synonymous and synonymous variation data to estimate the strength of purifying selection acting on the translated open reading frames (ORFs) previously identified as translated in different public ribosome profiling experiments. A subset of the lncRNAs are likely to be functional and conserved protein-coding genes translating small proteins, since many studies have proposed that a fraction of the small proteome is yet not annotated. However, the bulk of lncRNAs code for peptides which show variation patterns consistent with neutral protein-coding gene evolution. We also show that the ORFs that have a more favorable, coding-like, sequence composition are more likely to be translated than other ORFs in lncRNAs. This study provides the first evidence to date that there is a large and ever-changing reservoir of lowly abundant proteins; some of these peptides may become useful and act as seeds for *de novo* gene evolution.

ANALYSIS OF LONG NONCODING RNA AND CHROMATIN INTERACTIONS BY CHROMATIN ISOLATION BY RNA PURIFICATION (CHIRP)

Kan Saito, Vi Chu

MilliporeSigma, Cellular Assays, Biological Reagents & Kits, R&D, Temecula, CA

Gene regulation plays a critical role in complex cellular processes such as development, differentiation, and cellular response to environmental changes. While the regulation of gene expression by transcription factors and epigenetic influences has been well studied over time, pervasive genomic transcription and the role of non-coding RNAs in this process is a rapidly evolving field that remains to be thoroughly explored.

Chromatin Isolation by RNA Purification (ChIRP) is one of the methods, which allows analysis of DNA, RNA and protein in the RNA complex by using probe-based hybridization to target RNA molecules in chromatin. DNA can be isolated from recovered chromatin and analyzed by quantitative PCR or next generation sequencing (ChIRP-seq). Alternatively, proteins can be isolated and analyzed by western blotting or mass spectrometry (ChIRP-MS).
To enable the exploration of these RNA interactions in chromatin regulation, we have optimized the methods and developed ChIRP reagents. Using these reagents ChIRP experiments can be performed with reliable recovery of chromatin using lncRNA or other chromatin associated RNA as targets. We have performed ChIRP experiments with various cell lysates and capture oligos for several lncRNA targets (TERC, HOTAIR, Xist, NEAT1, U1 and U2snRNA).

In summary the methods optimized allow discovery of RNA-associated DNA and RNA sequences and also proteins.

EVOLUTION OF DELETION POLYMORPHISM OF THE GENE FOR A CELLULAR METABOLIZING ENZYME GLUTATHIONE-S-TRANSFERASE M1 IN HUMANS AND CHIMPANZEES

Marie Saitou[1], Takafumi Ishida[2], Yoko Satta[3], Omer Gokcumen[1]

[1]SUNY at Buffalo, Dept. of Biological Sciences, Buffalo, NY, [2]The University of Tokyo, Dept. of Biological Sciences, Tokyo, Japan, [3]SOKENDAI, School of Advanced Sciences, Kanagawa, Japan

More than half of the contemporary human chromosomes carry a deletion of a gene for a cellular metabolizing enzyme, glutathione-s-transferase μ1 (GSTM1). The GSTM1 null allele is thought to have been generated by homologous recombination of two segmental duplications (SDs). However, the evolutionary mechanisms through which deletion has been maintained in humans remain unknown.
To address this, using available reference genome sequences, we first reconstructed the evolutionary history of GSTM gene family, of which GSTM1 is a member. We revealed a dynamic history of this gene family where the chimpanzee GSTM5, gorilla GSTM4 and rhesus macaque GSTM1 were pseudogenized in a lineage-specific manner. It is likely that some GSTM genes may be under only weak functional constraints because GSTM genes compensate each other when one gene was lost. We also found that the primate GSTM1 genes were also flanked by the two SDs, which may cause the gene deletion. We then sequenced the breakpoints of this deletion by long-range PCR in humans and also in chimpanzees for comparative purposes. To our surprise, we found that there is an polymorphic deletion at the orthologous site of GSTM1 in approximately 40% of chimpanzee chromosomes as well. We then used a combination of phylogenetic tools to show clear separation of haplotypes carrying human deletion alleles and chimpanzee deletion alleles, suggesting that the human GSTM1 deletion allele and chimpanzee GSTM1 deletion allele were generated independetnly recurrently in human and chimpanzee lineages. Based on our results, the most likely explanation is that a combination of high mutation rate for deletion formation due to segmental duplications adjacent to this locus and the relatively small fitness effect of loss-of-function of GSTM1 may have allowed recurrent evolution of deletion alleles in humans and chimpanzees. It is of note that we did not find any deletion affecting GSTM1 among 69 macaques; however, the GSTM1 gene is pseudogenized in macaques due to premature stop-codon inducing single nucleotide variations. Overall, such different mechanisms of loss-of-function of the GSTM1 gene in several primate lineages suggest that the GSTM1 has been under weak functional constraints at least since the divergence between humans and macaques.

NANOPORE SEQUENCING REVEALS CONCERTED EVOLUTION IN VACCINIA VIRUS DRIVEN BY HOST-PATHOGEN CONFLICT

Thomas A Sasani, Kelsey Rogers-Cone, Ryan M Layer, Nels Elde, Aaron Quinlan

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

Viruses are locked in conflict with host organisms and rapidly adapt to combat antiviral host defenses. For example, the vaccinia virus (VACV) genome encodes two proteins, E3L and K3L, that each disrupt key elements of the human response to viral replication and propagation. In prior work, populations of VACV that lacked the E3L gene were shown to rapidly adapt under selective pressure by duplicating K3L in tandem gene arrays. Additionally, these populations began to accumulate K3L^{H47R} single-nucleotide mutations, which appeared to confer a fitness benefit comparable to duplication of wild-type K3L. The interplay between these two genomic adaptations has remained mysterious, largely because short-read sequencing technologies are unable to sequence through tandem arrays of K3L duplications.

In this study, we utilized the Oxford Nanopore (ONT) long-read platform to characterize K3L copy number and K3L^{H47R} accumulation in VACV populations that survived selective pressure during successive passages in HeLa cell lines. By sequencing the genomes of experimentally evolved VACV following 10, 15, and 20 passages (P10, P15, and P20), we gained insight into two key mechanisms of vaccinia adaptation. Using long reads, we directly characterized viral genomes harboring up to 19 tandem copies of K3L, likely a result of recombination-driven gene expansion. Interestingly, population distributions of K3L copy number are nearly identical at P10, P15, and P20; this may suggest that recombination can generate stable copy number increases at the population level. We also discovered that the K3L^{H47R} allele spreads rapidly over the course of vaccinia evolution, jumping from a population frequency of 12% at P10 to nearly 90% at P20. Using long ONT sequencing reads, we were able to move beyond a population-level view of K3L^{H47R} accumulation, and tracked the spread of the variant *within* tandem K3L arrays in individual viral genomes. We determined that K3L^{H47R} rapidly "homogenizes" within these arrays during virus evolution. After 10 passages through HeLa cells, nearly all tandem arrays are composed of wild-type alleles, but by P20 these arrays comprise entirely K3L^{H47R} alleles. This pattern of allele accumulation is consistent with a model of concerted evolution, which describes the homogenization of alleles within repetitive genome sequences. Though concerted evolution has been observed in numerous eukaryotic and bacterial species, our observations of a similar, conflict-driven process in vaccinia reveal a new and exciting facet of virus evolution.

FAST, HEURISTIC ARG INFERENCE FOR LARGE DATA SETS

Nathan K Schaefer, Richard E Green

University of California, Santa Cruz, Biomolecular Engineering, Santa Cruz, CA

Many phylogenomic studies seek to quantify and map patterns of ancestry and natural selection across panels of genomes. Various scans and statistics have been developed for this purpose; such techniques summarize or approximate part of the ancestral recombination graph (ARG). The ARG is a data structure that captures all information in a phylogenomic data set, describing the evolutionary relationships between all haplotypes at every variable site, along with historical recombination events that stitched together segments of haplotypes with different histories. Despite its power and utility, however, ARG inference remains computationally resource intensive and slow. We present a new, heuristic ARG inference algorithm built on the four haplotype test that can efficiently run on genomes from a large panel of individuals. We also use it to map Neanderthal ancestry in modern human genomes and detect higher average Neanderthal allele frequencies in Peruvians than other 1000 Genomes Project populations.

LESSONS LEARNED FROM BUILDING PERSONALIZED PHASED DIPLOID GENOMES OF THE EN-TEX SAMPLES

Michael C Schatz[1,2], Fritz J Sedlazeck[2], Han Fang[1,3], Alex Dobin[1], Anna Vlasova[4], Yunjiang Qiu[5,6], David Gorkin[5,6], Sora Chee[5], Laurent Luo[2], Maria Nattestad[2], Srividya Ramakrishnan[2], Charlotte Darby[2], Carrie Davis[1], Alessandra Breschi[4], Julien Lagarde[4], Roderic Guigo[4], Bing Ren[5,6], Thomas R Gingeras[1]

[1]Cold Spring Harbor Laboratory, Functional Genomics, Cold Spring Harbor, NY, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [3]Stony Brook University, Department of Computer Science, Stony Brook, NY, [4]Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain, [5]Ludwig Institute for Cancer Research, Genomics, La Jolla, CA, [6]UCSD School of Medicine, Genomics, La Jolly, CA

Most current transcriptome and other functional genomics studies begin by mapping sequencing data to a standard haploid reference genome. While this approach has been very effective for measuring major trends to gene expression and regulation across individuals and tissue types, it has been largely blind to the underlying genomic differences between individuals, and especially to the phase of those variants. Consequently, the community suffers from limited power to study the role of genomic variation on such effects as allele specific binding or allele specific expression modulated by distant cis- and trans-regulatory elements. To address this critical need, as part of the ENCODE project we have sequenced the genomes of 4 human samples obtained from an ENCODE-GTEX collaboration (EN-TEx) using a large collection of genomic technologies to construct a set of personalized genomes. This includes deep coverage of Illumina short read sequencing for high quality variant identification (60X), PacBio long read sequencing for phased structural variant analysis (55x), 10X Genomics Chromium (35x) sequencing for long range variant phasing, and Hi-C short reads for chromosome-span phasing and SV analysis (100x). Using this combination of data for both alignment-based and de novo assembly techniques, we have identified millions of single nucleotide and short indel variants per genome as well as thousands of larger structural variations in each individual. We have further processed the variant and read data to establish high quality phased personalized diploid genomes, with most genes and variants fully phased at chromosome-span in each individual genome. Based on the catalog of detected variations we have further prepared individualized genome gene annotations. When combined with over 500 RNA-seq, ChIP-seq and other genome wide functional datasets in progress or collected from approximately 25 tissues from each donor, these new phased personal genomes provide a foundation for unprecedented exploration into the interplay between variation, tissue specific expression, and regulation with allele-specific resolution. By integrating these different functional data types in the context of an annotated personalized phased diploid genome, we identify notable effects of genetic variations on gene expression profiles that cannot observed by using the consensus human genome sequence.

LOCAL ADAPTATION IN CHIMPANZEES

Joshua M Schmidt[1], Marc de Manuel[2], Tomas Marques-Bonet[2,3], Sergi Castellano[1], Aida Andrés[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]Universidad Pompeu Fabra, Institut de Biologia Evolutiva, Barcelona, Spain, [3]Institucio Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

Local adaptation has been extensively studied in modern human populations. While there are notable examples of positive selection, a consensus view has emerged suggesting a limited role for positive selection in shaping the patterns of pairwise differentiation between modern human populations, which are mainly determined by the combined effects of demographic history and drift (including the effects of background selection). How specific this is to humans in comparison to other primates is not known, as equivalent processes in other species, including our closest living relatives the chimpanzees, have not been studied.

Chimpanzees are divided in four subspecies that differ in geographical range, genetic diversity, and estimated effective population size (Ne). Their pairwise levels of genetic differentiation range from those among modern human populations to those between modern humans and Neandertals. Studying local adaptation in chimpanzee sub-species not only allows us to put local adaptation of human populations in its broader evolutionary context, but it can also inform about how these endangered populations have adapted to their environments, as well as about the influence of relevant parameters (such as Ne and environmental pressure) that influence the likelihood and mode of adaptation.

Here we present analyses on the genome-wide patterns of differentiation in chimpanzees utilizing recently published high-coverage genome-wide polymorphism data of 58 chimpanzees from the four sub-species. Genome wide we find that, just as in humans, patterns of population differentiation can be explained with plausible levels of background selection, except in the particular case of Central and Eastern chimpanzees. In contrast to equivalent patterns in humans, this cannot be realized in coalescent simulations using realistic demographic models, either with identical or asymmetric strengths of background selection. This provides strong evidence of the action of recent, local positive selection. Further, both haplotype based statistics and a simple, new extension of the population branch statistic (PBS) – that summaries the joint site frequency spectrum of all four chimpanzee sub-species – identify candidate loci responsible for local adaptation in each of the sub-species. The human paralogs of some of these genes are known to play key roles in variation in pathogen resistance, suggesting a predominate role for immunity related functions in chimpanzee local adaptation.

GENOME-WIDE SURVEY OF RARE MUTATIONS INFLUENCING PROTEIN ABUNDANCES IN YEAST

Olga T Schubert[1], Joshua S Bloom[1,2], Leonid Kruglyak[1,2,3]
[1]University of California, Department of Human Genetics, Los Angeles, CA,
[2]University of California, Howard Hughes Medical Institute, Los Angeles, CA,
[3]University of California, Department of Biological Chemistry, Los Angeles, CA

Despite significant advances in identifying genetic loci underlying phenotypic variation, our understanding of how specific genetic variants cause trait differences remains sparse. The phenotype of a cell is mostly determined by proteins, serving as essential structural components, enzymes, and constituents of signaling networks, from receptors to transcription factors. Therefore, variation in protein abundance as a consequence of DNA sequence polymorphisms is a key contributor to the connection between genotypes and cellular and organismal phenotypes. Here we aim to characterize general principles of the genetic architecture underlying protein abundance variation, including the number of cis- and trans-acting loci that can influence the abundance of a given protein, their effect sizes, and whether they are enriched in non-coding DNA or coding regions of specific classes of genes. To achieve this, we are using random chemical mutagenesis, high-throughput single-cell protein quantification based on fluorescent tags, and pooled sequencing. In contrast to earlier studies focused on natural variation segregating in a cross, this approach is independent of existing natural polymorphisms, and can therefore query a broader mutational space for effects on protein abundance.

Our experimental workflow is based on the Yeast GFP Collection, which contains over 4000 yeast strains, each of which carries a translational fusion of an endogenous protein with GFP. The first step is chemical mutagenesis of an individual GFP-tagged yeast strain. A small fraction of the resulting mutant cells are expected to contain mutations that either increase or decrease the abundance of the protein of interest. In the next step, we isolate these mutants with extreme protein levels by fluorescence-activated cell sorting. We can then identify causal loci directly by whole-genome sequencing of pools of these selected cells. Pilot studies and simulations indicate that this approach can identify causal loci even if the mutational target size includes several dozen genes, provided that the pooled samples are sequenced to a sufficient depth, and that sequencing errors are kept as low as possible, which can be accomplished by using overlapping paired-end sequencing and stringent filtering on base call quality.

To unravel general principles of the genetic architecture underlying protein abundance variation, we will apply the approach described above to dozens of proteins representing different functional classes. Furthermore, we will interrogate yeast orthologs of human disease-relevant proteins to gain insight into genetic networks and biological processes underlying human diseases, potentially revealing new candidates for medical intervention. Overall, this work will enhance our knowledge on how genome variation modulates the molecular phenotype of a cell, thereby providing a key missing functional link for genotype-phenotype associations.

LINKED LONG READ SEQUENCING AND OPTICAL MAPPING FOR *DE NOVO* GENOME ASSEMBLY OF AN ENDANGERED SPECIES

Alan F Scott[1], David W Mohr[1], Ahmed Naguib[2], Deanna Church[3]

[1]Johns Hopkins University, Genetics, Baltimore, MD, [2]BioNano Genomics, San Diego, CA, [3]10X Genomics, Pleasanton, CA

The widespread sequencing of mammalian genomes has been hindered for a variety of reasons including cost, sample requirements and analysis. Current short-read methods have come to dominate genome sequencing because they are cost-effective, rapid, and accurate. However, short reads are most applicable when data can be aligned to a known reference and assembled with more traditional methods. We have explored de novo genome assembly of the endangered Hawaiian monk seal using the10X Genomics Chromium linked-reads approach combined with BioNano Genomics (BNG) optical mapping. We show that linked reads, assembled with Supernova v1.1 alignment software produced scaffolds with an N50 of 22.23 Mbp with the longest individual scaffold of 84.06 Mbp. When combined with BNG optical maps the scaffold N50 increased to 29.65 Mbp and the longest individual scaffold increased to 84.78 Mbp. A total of 170 hybrid scaffolds were created that represented approximately 98% of the expected genome. Along with improving scaffold length, the optical maps served as a useful standard to which the sequence could be aligned and regions of compression or insertion identified. BUSCO analysis and manual identification of translated scaffolds showed that most expected protein coding genes were present and that the gene order was often conserved with human or other mammals over tens of Mbps. Combining the orthogonal single-molecule based methods of 10X linked reads with BNG optical maps is likely to make the assembly of high quality genomes more routine and significantly improve our understanding of comparative genome biology.

PATERNALLY INHERITED NONCODING STRUCTURAL VARIANTS CONTRIBUTE TO AUTISM

Jonathan Sebat[1], William M Brandler[1], Danny Antaki[1], Madhusudan Gujral[1], Venter C J[2], Christina Corsello[3], Keith Vaux[1], Lilia Iakoucheva[1], Amaia Hervás [4], Maria Arranz[4], Boyko Kakaradov[2], Amalio Telenti[2]

[1]UC San Diego, Department of Psychiatry, La Jolla, CA, [2]Human Longevity Inc., HLI, La Jolla, CA, [3]Rady Children's Hospital, Autism Discovery Institute, San Diego, CA, [4]Mutua Terrassa Hospital, Department of Child Psychiatry, Barcelona, Spain

The genetic architecture of autism spectrum disorder (ASD) is known to consist of contributions from gene-disrupting *de novo* mutations and common variants of modest effect. We hypothesize that the unexplained heritability of ASD lies between these two extremes, and also includes rare inherited variants with intermediate effects. We investigated the genome-wide distribution and functional impact of structural variants (SVs) through whole genome sequence analysis (>30X coverage, Illumina HiSeq X10) of 3,169 subjects from 829 families affected by ASD. Genes that are intolerant to inactivating variants in the exome aggregation consortium (ExAC) were depleted for SVs in parents, specifically within promoters, UTRs and exons. Rare paternally-inherited SVs that disrupt promoters or UTRs were over-transmitted to probands (P = 0.0006) and not to their typically-developing siblings. Protein-coding SVs were also associated with ASD (P = 0.001) and displayed a maternal bias. Recurrent functional noncoding deletions implicate the gene LEO1 in ASD. Our results establish that rare inherited SVs predispose children to ASD, with differing contributions from each parent. Potential mechanisms to explain the paternal origin effect of *cis*-regulatory SVs will be discussed.

ACCURATE AND FAST DETECTION OF COMPLEX AND NESTED STRUCTURAL VARIATIONS USING LONG READ TECHNOLOGIES.

Fritz J Sedlazeck[1,2], Philipp Rescheneder[3], Moritz Smolka[3], Han Fang[4], Maria Nattestad [4], Arnd von Haeseler[3], Michael C Schatz[1,4]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Baylor College of Medicine, Baylor College of Medicine, Houston, TX, [3]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna, Vienna, Austria, [4]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

The impact of structural variations (SVs) is becoming more prominent within a variety of organisms and diseases, especially human cancers. Short-read sequencing has proved invaluable for recognizing copy number variations and other simple SVs, although has been highly limited for detecting most other SVs because of repetitive elements and other limitations of short reads. The advent of long-read technologies, such as PacBio or Nanopore sequencing that now routine produce reads over 10,000bp, offer a more powerful way to detect SVs. However, currently available methods often lack precision and sensitivity when working with highly erroneous reads, especially for more complex or nested SVs.

Here we present Sniffles, a method for detecting all types of SVs from long-read sequencing data. A unique feature of Sniffles is detecting nested SVs (e.g. chromothripsis), such as inversions flanked by deletions, which we now commonly detect in several samples. Sniffles finds SVs from split-read alignments as well as the analysis of “noisy regions”, where sequencing errors mask the presence of biological differences. A self-balancing interval tree enables a fast runtime so that whole human genome datasets can be analyzed in minutes. Furthermore, Sniffles offers read-level phasing to study complex breakpoints, as in chromoplexy. Using real and simulated data, we demonstrate the enhanced ability of Sniffles to detect SVs over existing methods like PBHoney or short read methods such as Lumpy, Delly, or Manta. We further introduce a new long-read mapping method called NGM-LR to enhance the accuracy of Sniffles and reduce the false discovery of SVs even further. NGM-LR is uniquely capable to detect and react if a read overlaps with a disturbed region (e.g. SVs) during the mapping phase and thus provides a more accurate alignment.

Working with genuine PacBio and Oxford Nanopore reads with human cancer samples (SKBR3), healthy human samples (GIAB), and other species, we show how Sniffles combined with NGM-LR reduces the coverage, and therefore cost, required per sample for highly sensitive and specific SV detection. Sniffles and NGM-LR are available open-source at Github, and are already being used by multiple institutes around the world.

INTEGRATING eQTLS ACROSS 44 HUMAN TISSUES WITH GWAS UNCOVERS NEW ASSOCIATIONS AND CAUSAL GENES FOR COMPLEX DISEASES

Ayellet V Segre[1], Francois Aguet[1], GTEx Consortium[2], Gad Getz[1,3], Kristin G Ardlie[2]

[1]The Broad Institute, Cancer Program, Cambridge, MA, [2]The Broad Institute, Medical and Population Genetics Program, Cambridge, MA, [3]Massachusetts General Hospital, Dept. of Pathology, Boston, MA

Thousands of common variants associated with complex diseases and traits have been detected through genome-wide association studies (GWAS), but many associations of modest effect remain to be found. The majority of detected associations lie in noncoding regions, suggesting that variation in gene regulation plays a primary role in disease etiology. Identifying the regulatory mechanisms, causal genes and biological processes through which genomic loci exert their effect on disease in relevant tissue contexts remains a major challenge. The Genotype-Tissue Expression (GTEx) project, designed to identify DNA variants associated with gene expression changes (eQTLs) across healthy human tissues, is uniquely suited to address these challenges. We developed a two-step statistical method that integrates GWAS summary statistics with eQTL and pathway data to propose new causal genes and tissues of action for known and new genetic associations. The method tests whether eQTLs significant in a given tissue are enriched for modest to genome-wide significant trait associations (e.g. GWAS $p<0.05$), correcting for potential confounding effects. If enrichment is found, target genes of eQTLs (eGenes) with GWAS $p<0.05$ are further tested for enrichment in signaling and metabolic pathways or phenotype ontologies. We applied our method to GWAS meta-analyses of 18 traits with available summary statistics (metabolic, cardiovascular, anthropometric, autoimmune and neurodegenerative), using eQTLs from 44 GTEx tissues (~800-10,000 eGenes/tissue, FDR<5%), and tissue-specific subsets of these eQTLs. Significant enrichment of trait associations among eQTLs was found for all traits in a range of tissues ($p<6E-05$, fold-enrich.=1.1-2.2), including expected pathogenic tissues, such as whole blood and colon for Crohn's disease and hippocampus for Alzheimer’s disease. Enrichment of trait associations among tissue-specific eQTLs was found for a limited number of tissue-trait pairs, with the strongest enrichment found for systolic blood pressure (SBP) and aorta-specific eQTLs (P=2E-05, fold-enrich.=2.2). Among novel SBP associations and genes enriched in cardiovascular-related gene sets, one was recently validated in a 5-fold larger SBP GWAS meta-analysis (rs4691707, aorta-specific eQTL for GUCY1B3), demonstrating the power of our integrative approach. In addition to tissue-specific eQTL associations, we characterize the extent of tissue-specificity of eQTLs underlying each complex trait.

ESTIMATION OF GENE EXPRESSION AND MUTATION DETECTION BY RNA-SEQ USING NANOPORE SEQUENCING

Masahide Seki, Eri Katsumata, Yutaka Suzuki

The University of Tokyo, Department of Computational Biology and Medical Science, Kashiwashi, Chiba, Japan

Recently-launched portable USB sequencer, MinION has the potential for omics analyses to be conducted even in the laboratories outside the sequencing core labs or in developing countries, where initial investment for the next generation sequencing is not available. In this study, we optimized RNA-Seq by MinION sequencer. We started from lung cancer cell line LC2ad and prepared the double stranded full-length cDNA library using Smart-Seq cDNA synthesis kit. The obtained template was directly introduced to library preparation and sequencing by the latest MinION cell, R9.4 (FLO_MIN106). Four MinION flow cells yielded 801,449 reads with high quality in total. Its improved sequencing fidelity at >92% allowed the precise mapping of the sequences using LAST. On average, each of single reads covered sixty percent of the entire transcript regions. Overall correlation of the expression levels comparing that measured by MinION sequencing and that by the Illumina True RNA Seq was 0.89, suggesting the expression estimations by MinION and Illumina sequencing are comparable. Each of the exons was located precisely, representing the full-length splicing pattern of the transcript. Notably, the fusion gene transcripts, which are important to characterize cancers though sometimes difficult to identify solely using Illumina reads, was represented as a single sequence read spanning both of the fusion gene partners. So-called phasing information, that is, allelic background information of cancer-causing driver mutations and secondary acquired mutations related to drug resistance, was obtained by inspecting the obtained long sequence reads. Taking advantages of not only its portable use but also its long read, MinION sequencing have the potential to further accelerate the various types of omics studies by providing convenient sequencing platform to wider research communities.

GENOMIC RECONSTRUCTION FROM LONGITUDINAL METAGENOMICS DATASETS DURING PREGNANCY

Myrna G Serrano, Vaginal Microbiome Consortium at VCU

Virginia Commonwealth University, Center for the Study of Biological Complexity, Richmond, VA

The vaginal microbiome in pregnancy plays an important role in both maternal and neonatal health outcomes. Despite the critical role of the human microbiota in health, our understanding of microbiota compositional dynamics during pregnancy is incomplete. In our Multi Omic Microbiome Study-Pregnancy Initiative (MOMS PI), we generated multi-omic data from ~90 women who delivered at term (>37 weeks). Samples were obtained during the prenatal visit of women from all trimesters through delivery/discharge of healthy pregnant MOMS-PI participants. A very comprehensive microbiome profiling by 16S rRNA survey includes: longitudinal maternal vaginal, rectal, buccal samples, neonatal meconium, 1st/2nd stool and buccal swabs within an hour of birth and at discharge. Our 16S rRNA taxonomic analysis using the STIRRUPS database generates species-level information of the bacteria. In addition, we selected 575 samples for whole metagenome sequencing, including maternal longitudinal vaginal samples (427) and infant meconium/1st stool (148). Metagenomic sequence contigs were binned into clusters using k-mer strategy in conjunction with read abundance data. Using this approach a total of 452 bacterial genomes were identified. Over 70% of the reconstructed genomes were high quality assemblies. The metagenomic sequence provides significant additional information relevant to the biology and potential pathogenesis of dozens of strains of bacteria. This analysis provides information about genomic variations in strains of known bacterial species. We are able to generate and analyze the sequences of newly identified bacteria that have not been previously identified or characterized. Assembly of metagenomics sequence data into genomes of individual species has a fundamental value to expand our knowledge of the composition of the microbial and will allow us to explore in depth the microbial community and the associated host response during pregnancy.

This work was supported by grants from the NIH Common Fund Human Microbiome Project program (1UH2/UH3AI08326 and 8U54HD080784 and to G. Buck, K. Jefferson and J. Strauss). We thank the members of the Vaginal Microbiome Consortium at VCU (vmc.vcu.edu) for their hard work on this project, and the Global Alliance for the Prevention of Prematurity and Stillbirth for their participation in launching the MOMS-PI project.

IDENTIFICATION OF NOVEL DISEASE LOCI BY BAYESIAN LATENT VARIABLE RE-CODING OF PHENOTYPES IN EXISTING GWAS

Afrah Shafquat[1], Jason Mezey[1,2]

[1]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY, [2]Weill Cornell Medicine, Department of Genetic Medicine, New York, NY

Poorly or inconsistently defined disease phenotypes are among the numerous factors that can negatively impact discovery of disease loci in Genome Wide Association Studies (GWAS). Here we describe a novel association approach that addresses this issue by assuming measured GWAS phenotypes can be used to learn a latent, correlated phenotype that has better properties for identifying disease loci associations.

Our method makes use of a hierarchical Bayesian model that allows probabilistic re-coding of a measured phenotype given the information present in genotype associations. Parameters of the model, including those describing genotype associations with the latent phenotype and modification probabilities associating the latent and observed phenotypes, are inferred using a parallelized MCMC algorithm. Modification probabilities are used to define alternative latent phenotypes for association analysis, where the complete methodology involves multiple checks on false positive artifacts, including assessment of genome-wide inflation factors and phenotype correlation analysis.

We evaluated the performance of the method on simulated GWAS data where the measured disease phenotype was correlated with a latent phenotype that had stronger associations with disease loci. In these simulations, our method was able to correctly recover the latent phenotype coding and disease associations that were ambiguous or invisible when analyzing the original phenotype. This was the case when assuming realistic heritability and disease loci effect sizes, as well as moderate to high correlations between the measured and latent phenotype. We also applied our method to several existing GWAS data sets, including those available from the UK Biobank, the Wellcome Trust Case Control Consortium, and the NIDDK IBD Genetics Consortium. From the re-analysis of these data, we found several novel disease loci for IBD and for psychiatric disorders, many of which had highly suggestive functional annotations, where these loci had moderate to low non-significant associations with the original measured phenotypes. Overall, our methodology and analyses indicate there are many additional disease-associated loci that could be identified from existing GWAS when assuming a measured phenotype contains information about, but is not necessarily itself, the best measurement for disease locus mapping.

A COMPUTATIONAL APPROACH TO THE MUTATIVE EFFECTS OF AID ON EBV LATENT REPLICATION

Maxwell Shapiro*[1], Teresa Martinez*[1], Thomas MacCarthy[1,2]

[1]Stony Brook University, Applied Mathematics & Statistics, Stony Brook, NY, [2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY

Epstein-Barr Virus (EBV) is a gammaherpesvirus that establishes different latency stages in human B-cells characterized by distinct patterns of virus gene expression. The enzyme Activation Induced Deaminase (AID) is expressed in germinal center B-cells, and is required for the processes of somatic hypermutation and class-switch recombination that are part of normal antibody affinity maturation. Specifically, AID induces cytosine to uracil mutations on ssDNA within the antibody gene (Immunoglobulin, or Ig) loci. Mistargeting of AID to other (non-Ig) loci can lead to malignancies, including Burkitt Lymphoma, which is associated with EBV. Using computational methods, we investigate the mutative properties of AID on EBV, focusing particularly on the origin of plasmid replication (oriP) and latency genes. AID has been shown to preferentially deaminate at the "hotspot" motif WRC (W=A/T, R=A/G) and after analyzing 33 EBV sequences for evidence of AID mutation, we identified statistically significant deamination that is preferentially targeted to the lagging strand in the oriP, consistent with targeting by other cytosine deaminases such as APOBEC3. Of those EBV sequences found to have higher mutation frequencies on the lagging strand, a majority come from EBV-associated Burkitt Lymphoma cells. We also find that certain subregions of the oriP tend to have higher mutation frequencies, particularly at motifs such as AGCA and TGCT, where AID hotspots overlap on both strands, and which mainly occur in the repeat region of oriP.

*Authors contributed equally

IDENTIFICATION OF MIR-31 AS A MOLECULAR STRATIFIER OF CROHN'S DISEASE PHENOTYPES

Ben Keith[1], Nevzat Kazgan[2], Jasmine Barrow[2], Neil Shah[2], Terrence S Furey[1], Praveen Sethupathy[1], Shehzad Z Sheikh[1,2]

[1]UNC, Genetics, Chapel Hill, NC, [2]UNC, Center of Gastrointestinal Biology and Disease, Chapel Hill, NC

Background: There is currently no way to determine a personalized approach for therapy in Crohn's Disease (CD). Our group recently showed that protein-coding gene expression patterns define two clinically distinct molecular signatures in non-inflamed colon tissue from a cohort of adult CD patients. One signature largely resembled the expression profile of a normal colon (colon-like), while the other displayed expression patterns characteristic of the ileum (ileum-like). We found that these two CD subtypes associated with clinical phenotypes, such as presence of rectal disease and need for post-surgical anti-TNF therapy. We hypothesized that micro(mi)RNA expression profiles would stratify these same molecular phenotypes, and that specific miRNAs are fundamental in the distinction of these major CD subtypes.

Methods: We generated and analyzed colonic miRNA sequencing (seq) data from the same adult cohort as our previous study (20 CD, 11 non-CD), along with colonic and ileal miRNA-seq from formalin-fixed paraffin-embedded tissue in a treatment-naive cohort of pediatric patients (94 CD, 54 non-CD). Principal component analysis was used to cluster samples with similar miRNA expression profiles. Specific differentially expressed miRNAs in disease vs control or between colon-like and ileum-like samples were identified using the DESeq2 package. Potential master miRNA regulators disrupted in CD pathogenesis that contribute the most to shaping changes in gene expression were determined using miRhub, a bioinformatic strategy that integrates miRNA and mRNA information to identify miRNAs that contribute the most to shaping changes in gene expression.

Results: We found that CD patients formed identical clusters based on miRNA expression patterns as those previously observed using mRNA expression profiles. Notably miR-31-5p contributed significantly to the segregation of patients with CD into two main molecular phenotypes, with significantly increased expression in ileum-like samples with respect to colon-like samples ($p = 9.8e-09$). It was a major driver in differentiating ileum-like samples from colon-like samples in adult and pediatric patients. Using RNA-seq data from the adult patient samples, we identified miR-31-5p as a candidate master regulator of gene expression pathways associated with ileum-like CD, a CD phenotype that is associated with the use of post-operative anti-TNF drugs.

Conclusion(s): Our results show for the first time that miRNA levels in colon tissue segregate adult and pediatric patients into two clinically distinct forms of CD. Specifically, our findings suggest that miR-31-5p contributes most to the discrimination between ileum-like and colon-like phenotypes and could serve as an effective biomarker of CD subtypes.

DISCOVERY OF LONG NOVEL INSERTIONS IN AFRICAN DESCENT POPULATIONS

Rachel M Sherman[1,2], Rasika Mathias[3], Margaret A Taub[4], Terri H Beaty[5], Ingo Ruczinski[4], Kathleen C Barnes[6], Steven L Salzberg[1,2,4,7]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]JHU, McKusick-Nathans Institute of Genetic Medicine, Center for Computational Biology, Baltimore, MD, [3]JHU, Medicine, Baltimore, MD, [4]JHU, Biostatistics, Baltimore, MD, [5]JHU, Epidemiology, Baltimore, MD, [6]University of Colorado, Colorado Center for Personalized Medicine, Aurora, CO, [7]JHU, Biomedical Engineering, Baltimore, MD

Current human genomics work relies heavily upon the human reference genome, which despite periodic updates and the recent inclusion of alternative loci, remains primarily representative of European ancestry populations. To date, most whole genome population variation studies in non-Europeans have been on a small scale, and often miss large variants due to current methods. Among these understudied populations are African populations. African populations are of particular interest both evolutionarily, and due to extensive evidence that their genomes are highly divergent, relative both to one another and to non-African populations. Though available whole genome data from individuals of African ancestry is increasing, existing variation studies in African populations primarily utilize alignment to the human reference (currently GRCh38) to discover variants. Although the simplest approach to variant discovery, reference based comparisons are inherently limiting. While this approach allows for SNP, deletion, and small insertion discovery, large insertions are impossible to find since reads within insertions do not align. On the other hand, *de novo* assembling high quality genomes from hundreds or thousands of individuals then performing comparisons is computationally infeasible. Here, we examine 910 deeply sequenced genomes of individuals of African ancestry from North and South America, the Caribbean, and West Africa, searching for long (> 1kb) novel insertions. We find the pan-genome from these individuals includes over 600 Mbp of previously uncatalogued DNA in the form of roughly 150,000 long insertions, over 450 Mbp of which is shared by more than one individual. To circumvent the limitations of reference based detection and eliminate the need for intensive whole genome *de novo* assembly, we combine the two methods. By aligning to the reference and assembling only unaligned reads, we are able to assemble long novel insertions. We then recover insertion locations using mate pair information. In addition to finding over 450 Mbp of shared, novel sequence, we discover several insertions within known exons, which could be linked to phenotypic differences in the populations we examined. The new pan-genome will provide a greatly improved basis for future genetic studies in this population.

UNIVERSAL 3-NUCLEOTIDE PERIODICITY DRIVES EFFICIENT PROTEIN TRANSLATION THROUGH STEPPED MRNA-RRNA BASE PAIRING

Ruchi B Sheth[1], William Barr[1], Jacob W Glickman[1], Abigail E Cram[1], Om Chatterji[1], Kelly M Thayer[2], Danny Krizanc[3], Michael P Weir[1]

[1]Wesleyan University, Biology, Middletown, CT, [2]Wesleyan University, Chemistry, Middletown, CT, [3]Wesleyan University, Mathematics and Computer Science, Middletown, CT

The idea that base pairing between mRNAs and structural rRNAs of ribosomes might contribute to protein translation has long been an intriguing possibility. The 530 loop of 16S rRNA has been implicated in translation initiation, elongation and termination. This loop and the corresponding highly-conserved loop in 18S rRNA are located in the mRNA entrance tunnel of ribosomes. As noted before high-resolution ribosome structures were described, the 530 loop contains a 3-nucleotide-repeating pattern complementary to the 3-nucleotide periodicity of protein open reading frames characterized by overrepresentation of (GCN)n. We find that the 3-nucleotide periodicity is significantly enhanced downstream of translation start codons of highly expressed Saccharomyces cerevisiae genes with significant depression of G at nucleotides 2 and 3 of these codons. We propose that during translocation steps, exposed rRNA nucleotides G530 and G529 can transiently base pair with the second and third mRNA nucleotides of the incoming A-site codon before engagement of the A-site tRNA and that this promotes efficient launching of the ribosome through the start region codons permitting high protein expression. This cooperation between the A-site and 530 loop requires a precise positioning of the mRNA reading frame, which if compromised in several adjacent codons leads to low translation levels. We suggest that rRNA nucleotides C528 and G527 may also base pair to the mRNA and help pull the mRNA into an S-shaped conformation as it advances through the ribosome entrance tunnel.

iMETHYL: AN INTEGRATIVE HUMAN DNA METHYLATION VARIATION DATABASE - DEVELOPMENT FOR MULTI OMICS DATA IN 3 TYPES OF HUMAN BLOOD CELLS

Yuh Shiwa[1], Ryohei Furukawa[1], Tsuyoshi Hachiya[1], Hideki Ohmomo[1], Shohei Komaki[1], Ryo Otomo[1], Mamoru Satoh[1], Jiro Hitomi[2], Kenji Sobue[2], Makoto Sasaki[2], Atsushi Shimizu[1]

[1]Iwate Medical University, Division of Biomedical Information Analysis, Iwate Tohoku Medical Megabank Organization, Iwate, Japan, [2]Iwate Medical University, Iwate Tohoku Medical Megabank Organization, Iwate, Japan

PURPOSE: Both hereditary and environmental factors can influence disease onset and phenotype. Omics analysis is a method to comprehend the influence of the environment on the amount of change in biomolecules. In recent years, DNA methylation has gained attention owing to its ability to evaluate long-term environmental exposure and transgenerational stress effects. However, only limited DNA methylation analyses are published. Here, we report a database that integrates whole DNA methylation analysis and whole transcriptome analysis in 3 principal blood cells (monocytes, T-lymphocytes, and neutrophils) of 100 cohort study subjects.

METHODS: We collected blood samples (8 ml) from the participants at the Tohoku Medical Megabank Project (TMM) in the Iwate prefecture, and several blood cells were isolated using FACS with high purities. DNA and RNA were extracted from the sorted cells and refrigerated at -80°C. Taking sex and age into consideration, we performed a whole DNA methylation analysis and whole transcriptome analysis using DNA and RNA collected from the 3 types of blood cells. Simultaneously, a whole genome analysis was also performed by extracting DNA from whole blood stored at the TMM biobank. We constructed the database on our website after integrating analytical data obtained from each study subject.

RESULTS: We performed whole-genome bisulfite sequencing of purified monocytes, CD4+ T-lymphocytes, and neutrophils collected from apparently healthy Japanese subjects and obtained comprehensive DNAm profiles covering ~90% of the CpG sites. Based upon the DNAm profiles, we estimated the average and standard deviation (SD) of DNAm levels for ~24 million CpG sites, and then calculated histograms for the DNAm levels of each CpG site. We successfully implemented the obtained data pertaining to DNAm, genetic variants, and gene expressions as an open database named “iMethyl” (http://imethyl.iwate-megabank.org). In the iMethyl browser, regions of interest (ROIs) can be specified using gene symbols, dbSNP rsIDs, and genomic positions. On the ROIs, the average and SD of DNAm levels as well as SNVs and gene expression levels are shown regarding the 3 types of blood cells. By clicking on the bar in the CpG tracks, histograms of the DNAm levels for each CpG site appear in the pop-up window.

CONCLUSION: We published the world’s first database of DNA methylation variation in 100 individuals. It is expected that these data will be useful not only to researchers specialized in epigenomics but also to those interested in the interactive analysis of DNA methylation, gene expression, and complex omics.

AN EPIGENETIC SWITCH CONFERS PLEIOTROPIC RISK FOR BONE MINERAL DENSITY AND HYPERGLYCAEMIA

Nicholas A Sinnott-Armstrong*[1,2], Isabel S Sousa*[1,3], Elizabeth R Ruedy[4], Richard C Sallari[1], Xing Chen[5,6], Simon E Nitter Dankel[7], Gunnar Mellgren[7], Anyonya Guntur[4], David Karasik[5,8], Hans Hauner[3], Clifford J Rosen[4], Yi-Hsiang Hsu[5,6], Douglas P Kiel#[5], Melina Claussnitzer#[1,3,5,6]

[1]Broad Institute of MIT & Harvard, Cambridge, MA, [2]Stanford University, Department of Genetics, Stanford, CA, [3]Technical University Munich, Else Kröner-Fresenius-Center for Nutritional Medicine, Munich, Germany, [4]Maine Medical Center Research Institute, Center for Molecular Medicine, Scarborough, ME, [5]University of Bergen, Department of Clinical Science, Bergen, Norway, [6]Hebrew SeniorLife, Insitute for Aging Research, Boston, MA, [7]Beth Israel Deaconess Medical Center, Boston, MA, [8]Bar-Ilan University, Faculty of the Medicine in the Galilee, Ramat Gan, Israel

Recent studies suggest shared etiologies of skeletal and glycemic traits, but the underlying genetic factors remain unknown. Here, we apply our complex trait dissection framework to the most significant bivariate genome-wide association signal for bone mineral density (BMD) and fasting glucose levels, at the 3q21.1 locus (bivariate association test on GWAS meta-analyses p=1.9 x 10-9). We show that the variant rs56371916 is causal at 3q21.1, i.e. the T allele at rs56371916 results in hyperglycemia and higher BMD, whereas the C allele (14% frequency in Europeans) causes normal blood glucose but lower BMD. The 3q21.1 locus, which contains an enhancer that allows for activation during mesenchymal differentiation processes, is repressed in human progenitor cells of adipocytes and osteoblasts. De-repression depends on a conserved SREBP1 activator motif in the hyperglycaemia major allele. However, the low BMD allele fails to fully de-repress by disrupting the SREBP1 motif and continuing activity of the Polycomb-repressive complex 2 subunit EZH2. This failure to de-repress results in downregulation of adenylate cyclase 5 (ADCY5) in both osteoblasts and adipocytes. In osteoblasts, ADCY5 downregulation leads to a cell-autonomous perturbation of fatty acid oxidation and failure to differentiate, and is recapitulated in murine osteoblast expression profiling. In adipocytes, we observed a decreased adrenergic lipolysis rate and improved insulin sensitivity for the minor allele, consistent with the epidemiology. CRISPR/Cas9 C-to-T editing of rs56371916 in patient derived osteoblasts and adipocytes restored ADCY5 activation by SREBP1, restored osteoblasts differentiation programs, and accelerated lipolysis rate and glycerol release in adipocytes, which is known to confer hyperinsulinemia and hyperglycemia. Together, our results uncover a pleiotropic risk locus, acting through the ADCY5 gene on bone and fat. Specifically, we establish rs56371916, by altering lineage-specific Polycomb derepression, to lie at the root of the strongest genetic correlation between BMD and glucose levels.

FUNCTION AND SPECIFICITY SHAPE THE PSEUDOGENE LANDSCAPE IN 17 MOUSE STRAINS

Cristina Sisu, Paul Muir, Mark Gerstein

Yale University, Molecular Biophysics and Biochemistry, New Haven, CT

The advances of the Mouse Genome Project toward completing the de-novo assembly for a variety of mouse strains, provide a unique opportunity to get an in-depth picture of the evolution and variation of these organisms. Here we present the analysis of the first draft of the mouse pseudogene annotation in 17 mouse strains and the reference genome. We annotated the pseudogenes in the 18 genomes using a combination of automatic and manual curation pipelines. While the majority of protein coding genes seemed to be conserved over the 6MY of evolution, the pseudogene complements tell a story of lineage specific genome remodeling processes. Looking at L1 transposable elements (TE) activity, we noticed that by contrast with human, where the TEs became silent following the last retrotransposition event, the mouse TE families exhibit multiple successive retrotransposition bursts resulting in a continuous renewal of the pseudogene pool. Moreover, pseudogene parents are enriched in essential genes with higher expression levels at multiple time points during mouse embryonic development, suggesting that these genes might lead to additional retrotransposition events resulting in new pseudogenes.

Next, by comparing the phylogenetic trees of various pseudogenes classes (ribosomal proteins, CDK, olfactory receptors) we noticed a strain specific evolutionary pattern that is reflected in strain specific phenotypes. The same specificity is recorded in the gene ontology and pseudogene family characterization of the strains. For example, Mus Spretus specific pseudogenes are enriched in apoptosis related genes and are characterized by the “Death” superfamily. This result is in concordance with the previous reports describing the strain specific tumor resistant phenotype and multiple active apoptotic pathways. Analysis of the pseudogene repertoire also points towards multiple genomic rearrangements. By examining the pseudogene loci we found that the proportion of un-conserved loci follows a logarithmic curve that matches closely the divergent evolutionary time scale of the mouse strains suggesting a uniform rate of genome remodeling across the murine lineage.

Despite the evolutionary specificity we found that all the strains share a spectrum of pseudogene biochemical activity. In particular, we found a uniform proportion of pseudogene transcription in brain tissue across all strains.

Finally, we compared the human and mouse lineage to gain insights into the evolution of loss and gain of function events. As such we were able to identify almost 200 new unitary pseudogenes in human and a comparable number in mouse that pin-point to interesting LOF and GOF events.

NONCANONICAL UORFS FORM A CONSERVED CLASS OF *CIS*-REGULATORY ELEMENTS IN YEAST WITH DISTINCT REGULATORY FEATURES.

Pieter J Spealman[1], Armaghan W Naik[2], Gemma E May[1], Scott Kuersten[3], Lindsay Freeberg[3], Robert F Murphy[2], Joel C McManus[1]

[1]Carnegie Mellon University, Department of Biological Sciences, Pittsburgh, PA, [2]Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA, [3]Illumina, Inc., Madison, WI

Upstream Open Reading Frames (uORFs) are a major class of cis-acting elements involved in regulating mRNA turnover and translation (Wethmar, 2014; Ingolia 2014). Recent genome-wide ribosome profiling studies suggest that many uORFs initiate with non-AUG start codons and may play a distinct role in translational regulation in response to stress (Ingolia et al. 2009, Zhang et al., 2011, Brar et al., 2011). While suggestive, these yeast non-AUG uORF predictions have been made without statistical control or validation, thus the importance and identity of these elements remain to be demonstrated.

We used a comparative genomics approach to study AUG and non-AUG uORFs. We first mapped the Transcript Leaders (TLs) in three yeast species, *Saccharomyces cerevisiae*, *S. paradoxus*, and *S. uvarum* using targeted RNA-sequencing. We then applied a novel machine learning algorithm (uORF-seqr) to ribosome profiling data from each species. This generated hundreds of statistically significant predictions, of which we have manually tested and validated over a dozen. Our analysis has revealed both conserved and species-specific AUG- and NCC-uORFs in hundreds of orthologous TLs. Furthermore, AUG and NCC-uORFs have several distinguishing trends that separate the two types. We found that AUG-uORFs tend to act as translational repressors while NCC-uORFs act as enhancers, consistent with the two types having distinct regulatory roles. Similarly, the two types have different rates of inclusion within transcript isoforms with NCC-uORFs having significantly higher rates of inclusion, suggesting transcription may play a role in uORF selection. Finally, the two types have distinct patterns of RNA binding protein occupancy around start codons. NCC uORFs exhibit increased protein occupancy downstream of their start codons, suggesting that RNA binding proteins may regulate NCC-uORF translation. By demonstrating the conservation of significant NCC uORFs, our results support important biological functions for this emerging class of cis-regulatory sequences.

GENOME-WIDE RECONSTRUCTION OF COMPLEX STRUCTURAL VARIANTS USING READ CLOUDS

Noah Spies[1,2,3], Ziming Weng[2,3], Justin M Zook[1], Robert B West[3], Serafim Batzoglou[4], Marc Salit[1], Arend Sidow[2,3]

[1]National Institute of Standards and Technology, Joint Initiative for Metrology in Biology, Stanford, CA, [2]Stanford University, Dept of Genetics, Stanford, CA, [3]Stanford University, Dept of Pathology, Stanford, CA, [4]Stanford University, Dept of Computer Science, Stanford, CA

Recently developed methods that utilize partitioning of long genomic DNA fragments, and barcoding of shorter fragments derived from them, have succeeded in retaining long-range information in short sequencing reads. These so-called read cloud approaches represent a powerful, accurate, and cost-effective alternative to single-molecule long-read sequencing. We developed software, GROC-SVs, that takes advantage of read clouds for structural variant detection and assembly. We apply the method to two 10x Genomics data sets, one chromothriptic sarcoma with several spatially separated samples, and one breast cancer cell line, all Illumina-sequenced to high coverage. Comparison to short-fragment data from the same samples, and validation by mate-pair data from a subset of the sarcoma samples, demonstrate substantial improvement in specificity of breakpoint detection compared to short-fragment sequencing, at comparable sensitivity, and vice versa. The embedded long-range information also facilitates sequence assembly of a large fraction of the breakpoints; importantly, consecutive breakpoints that are closer than the average length of the input DNA molecules can be assembled together and their order and arrangement reconstructed, with some events exhibiting remarkable complexity. These features facilitated an analysis of the structural evolution of the sarcoma. In the chromothripsis, rearrangements occurred before copy number amplifications, and using the phylogenetic tree built from point mutation data we show that single nucleotide variants and structural variants are not correlated. We predict significant future advances in structural variant science using 10x data analyzed with GROC-SVs and other read cloud-specific methods.

HIGH PREVALENCE OF COUPLED R-GENE DISEASE RESISTANCE SYSTEMS IN 11 SPECIES OF THE *ORYZEAE* TRIBE.

Joshua C Stein[1], Kapeel Chougule[1], Sharon Wei[1], Rod A Wing[2], Doreen Ware[1,3]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]University of Arizona, Arizona Genomics Institute, Tucson, AZ, [3]USDA-ARS-NAA, Robert W. Holley Center for Agriculture and Health, Ithaca, NY

Disease pathogens, such as rice blast (*Magnaporthe oryzae*), severely impact rice production and may pose an increasing threat as climate change alters the geographical range of pests in the future. Breeding for natural host resistance is a proven strategy but limited by sources of variation. Wild relatives of rice, collected from around the world, provide an untapped reservoir of resistance genes. Taking advantage of an 11-species set of reference genomes in Gramene (www.gramene.org) that spans the *Oryzeae* tribe (including wild species in the *Oryza* and *Leersia* genera), we discovered over 4500 NLR (nucleotide-binding domain and leucine-rich repeat containing) genes in 28 families. Rapid diversification of complex haplotypes by gene expansion and loss is typical of NLR genes, contributing to disease adaptation. Applying phylogenetic reconciliation methods to gene trees in these 28 NLR families, we found a 10-fold increase in duplication rates in lineages leading to both Asian and African cultivated rice, consistent with selection for resistance traits prior to domestication. Most NLR were positionally clustered, often forming complex arrangements of distantly related genes. Yet, clear orthologous relationships and evidence of conserved underlying haplotype structures could be drawn, even in the most distantly related (~17 MY) species of *Leersia*. Adjacent heterogeneous pairs, with head-to-head arrangement, showed a disproportionate prevalence, conservation, and association with putative integrated decoy domains, suggesting function as coupled NLR gene pairs. Striking variation in domain structure suggests that swapping of various decoy domains contributes to the evolution of haplotype diversity and resistance specificity. This study has opened a treasure trove of potentially novel resistance functions that may help in the future development and sustainability of rice. Funded by NSF awards #1026200 and #1127112.

RECONSTRUCTING GENOME BIOLOGY

Thomas Stoeger[1,2], Martin Gerlach[3], Richard I Morimoto[4], Luís A Amaral[2,3]

[1]Northwestern University, Center for Genetic Medicine, Chicago, IL, [2]Northwestern University, Institute on Complex Systems, Evanston, IL, [3]Northwestern University, Chemical & Biological Engineering, Evanston, IL, [4]Northwestern University, Molecular Biosciences, Evanston, IL

Why do genome biologists know what they know? Is it just curiosity, serendipity, and societal relevance – or is there something else? Without a complete understanding of the formation of knowledge we cannot distinguish missed research opportunities from biology that can be overlooked. I attempt to close this gap by reconstructing the formation of genomic knowledge during the last three decades and by building models that predict the amount of publications and annotations for individual genes. We find that following the publication of the first mammalian genomes, genome biology transitioned towards subcellular proteins, and miniscule biological processes, and a continued decrease of impact and growth. Along the same lines, during the last ten years genome biology stopped to diversify, and scientists continue to accumulate explicit knowledge solely about a small set of genes. Irrespective of the year, the majority of publications can be predicted by a series of largely redundant chemical and biological attributes of genes and their products. By superimposing knowledge onto maps of those chemical and biological properties we identify vast territories of uncharted biology. Finally, we will present strategies how individual researchers may be able to reduce the risk of studying something arbitrary upon jumping into the unknown. We believe that such alternative strategies could complement approaches relying on guilt-by-association. We hope that our work can help to counteract the decreasing impact of genome biology by equipping individual researchers with the first predictive framework of any scientific discipline.

DE NOVO IDENTIFICATION OF DNA MODIFICATIONS ENABLED BY GENOME-GUIDED NANOPORE SIGNAL PROCESSING

Marcus Stoiber[1], Joshua Quick[2], Rob Egan[3], Ji Eun Lee[3], Susan Celniker[1], Robert K Neely[4], Nicholas Loman[2], Len A Pennacchio[3], James Brown[1]

[1] Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology, Berkeley, CA, [2]University of Birmingham, Institute of Microbiology and Infection, Birmingham, United Kingdom, [3]Lawrence Berkeley National Laboratory, Joint Genome Institute, Berkeley, CA, [4]University of Birmingham, School of Chemistry, Birmingham, United Kingdom

Advances in single molecule sequencing technology have enabled the investigation of the full catalogue of covalent DNA modifications. We present an assay, Modified DNA sequencing (MoD-seq), which leverages information-rich raw nanopore data to directly survey DNA modifications without the need for any prior training dataset. This assay is facilitated by a new, open source software package, nanoraw, for precisely assigning raw nanopore signal to genomic positions, enabling novel data visualization strategies, and the discovery of covalently modified bases in native DNA. Case studies applying MoD-seq identify three distinct marks, 4mC, 5mC, and 6mA, and demonstrate quantitative reproducibility across biological replicates processed in different labs. In a ground-truth dataset created via in vitro treatment of synthetic DNA with selected methylases, modifications are detected in a variety of distinct sequence contexts. We recapitulated known methylation patterns in native human and E. coli samples, and propose a pipeline for the comprehensive discovery of DNA modifications in any genome without a priori knowledge of their chemical identities.

RAREVARIANTVIS 2: A SUITE FOR ANALYSIS OF RARE GENOMIC VARIANTS IN WHOLE GENOME SEQUENCING DATA

Adam Gudys[1], Torunn Fiskerstrand[2,3], Rita Holdhus[2], Gunnar Houge[3], Mark Gerstein[4], Inge Jonassen[5], Vidar Steen[2], Tomasz Stokowy[2,4,5]

[1]Silesian University of Technology, Department of Informatics, Gliwice, Poland, [2]University of Bergen, Department of Clinical Science, Bergen, Norway, [3]Haukeland University Hospital, Center for Medical Genetics and Molecular Medicine, Bergen, Norway, [4]Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT, [5]University of Bergen, Department of Informatics, Bergen, Norway

The search for causative genetic variants in rare diseases of presumed monogenic inheritance has been boosted by the implementation of whole genome sequencing (WGS). Analysis and visualisation of WGS data is demanding due to its size and complexity. To aid this challenge, we have further developed and significantly extended the RareVariantVis tool (Stokowy et al., Bioinformatics 2016). RareVariantVis 2 annotates, filters and visualises a whole human genome in less than 30 minutes. Additionally, it allows calling of homozygous regions from sequencing data. The method accepts and integrates vcf files for single nucleotide, structural and copy number variants produced by various callers (GATK, speedseq) and provides annotated rare variant lists and relevant chromosome visualization. RareVariantVis 2 was successfully used to disclose causes of three rare monogenic disorders, including one non-coding variant.

A SYSTEMATIC ASSESSMENT OF THE POPULATION GENETIC EVIDENCE FOR SELECTION ACROSS BRAIN RELATED PHENOTYPES

Evan R Beiter[1], Ekaterina A Khramtsova[2,3], Celia Van Der Merwe[4], Emile Chimusa[5], Corrine N Simonti[6], Dan Stein[4], John A Capra[6], James Knowles[7], Lea K Davis*[6], Barbara E Stranger*[2,3,8]

[1]Washington University, Department of Biology, St. Louis, MO, [2]University of Chicago, Section of Genetic Medicine, Chicago, IL, [3]University of Chicago, Institute of Genomics and Systems Biology, Chicago, IL, [4]University of Cape Town, Department of Psychiatry, Cape Town, South Africa, [5]University of Cape Town, Department of Human Genetics, Cape Town, South Africa, [6]Vanderbilt University, Vanderbilt Genetics Institute, Nashville, TN, [7]University of Southern California, Department of Psychiatry and Behavioral Sciences, Los Angeles, CA, [8]University of Chicago, Center for Data Intensive Science, Chicago, IL
*Contributed equally

Inter-individual variation in neuropsychiatric traits is present across diverse human populations, has persisted through history, and has a genetic basis primarily accounted for by common single nucleotide polymorphisms (SNPs). Motivated by recent observations that SNPs with high minor allele frequency (MAF) contribute disproportionately to some neuropsychiatric phenotypes, we tested the hypothesis that common susceptibility variants for neuropsychiatric phenotypes have experienced weak positive selection. We performed multiple analyses using genome-wide association study summary statistics from studies of neuropsychiatric traits, personality measures, MRI subcortical brain structure volumes, and autoimmune phenotypes to assess evidence of selection. Consistent with expectations for polygenic phenotypes, no trait displayed significant enrichment of population differentiated SNPs or strong recent positive selection. However, congruent with recent reports, we identified enrichment of risk alleles for schizophrenia (p=0.004) and neuroticism (p<0.002) within regions of the genome under selection since divergence from Neanderthal. We assessed each phenotype for evidence of polygenic selection using an approach that detects coordinated shifts in the MAF of many trait-associated SNPs after accounting for genetic drift (Berg and Coop 2014). Significant evidence of polygenic adaptation was found for extraversion (p<0.001), schizophrenia (p<0.001), hippocampus volume (p<0.001), and putamen volume (p<0.001). We found evidence of very recent derived-allele selection in schizophrenia (p<0.001), putamen volume (p<0.001), and hippocampus volume (p<0.001) through Singleton Density Score analysis. Our results suggest that associated alleles for multiple neuropsychiatric and brain volume phenotypes have experienced weak selective pressures. Importantly, however, the results of these analyses do not indicate the targets of selection. We conducted expression Quantitative Trait Locus (eQTL) and gene set enrichment analyses to shed additional light on biological processes that may underlie the observed selection. Among SNPs associated with schizophrenia and putamen we found significant evidence of eQTL enrichment in brain (p<0.001) and immune tissues (p<0.001), respectively.

THE ENCODE ANNOTATION PIPELINE: CLOUD TO GROUND FOR CHIP-SEQ, RNA-SEQ, DNASE-SEQ, AND WHOLE-GENOME BISULFITE EXPERIMENTS

J Seth Strattan[1], Timothy R Dreszer[1], Ben C Hitz[1], Esther T Chan[1], Jean M Davidson[1], Idan Gabdank[1], Jason A Hilton[1], Cricket A Sloan[1], Zhiping Weng[2], Anshul Kundaje[1], J Michael Cherry[1]

[1]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [2]University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA

The ground-level annotations ENCODE applies to the epigenome and the transcriptome are produced by defined cloud-based computational pipelines. These pipelines are run and shared for the primary analysis of ChIP-seq, RNA-seq, DNase-seq, and whole-genome bisulfite experiments. By standardizing the computational methodologies for analysis and quality control, results from multiple labs can be directly compared and integrated into higher-level annotations, such as ENCODE Candidate Regulatory Elements (CREs). The ENCODE Data Coordinating Center (DCC) have deployed the pipelines to a cloud-based computing environment so that the computational load can be distributed across scalable resources. The particular platform we have chosen (DNAnexus) features a web-based graphical user interface (GUI) to the pipelines that is accessible to anyone with an account on the platform. Users pay for their own compute and can produce analysis results and quality-control metrics from their own data. In this way, their results are directly comparable to ENCODE's annotations.

ENCODE analyses are distributed through the ENCODE Portal at https://www.encodeproject.org/ The pipelines are available as “ENCODE Uniform Processing Pipelines” at https://platform.dnanexus.com/projects/featured The ENCODE DCC codebase is at https://github.com/ENCODE-DCC

WIDESPREAD ACCUMULATION OF 3' UTR MRNA FRAGMENTS IN SPECIFIC NEURONAL CELL POPULATIONS OF THE AGING BRAIN

Peter Sudmant, Myriam Heiman, Christopher B Burge

Massachusetts Institute of Technology, Biology, Cambridge, MA

Aging differentially impacts the various regions and cell types of the brain. Here we observe accumulation of 3' UTR mRNA fragments without corresponding coding regions for hundreds of genes in aging D1 medium spiny neurons (MSNs) of the mouse striatum. Increased levels of oxidative stress and dramatic changes in expression of translation and oxidative phosphorylation genes accompany fragment accumulation in aged D1 MSNs. We provide evidence that these mRNA fragment species also accumulate in the aging human brain, particularly in more metabolically active nuclei. Fragment accumulation can be stimulated by treatment with compounds that induce oxidative stress and is associated with increased expression of mitochondrial genes. Our findings support a model in which mRNA 3' UTR mRNA fragments accumulate as a result of oxidative stress-induced failure of the ribosome recycling protein ABCE1, eliciting mRNA cleavage via the No-Go decay pathway. We conclude that fragment accumulation is a hallmark of the aging brain

DE NOVO METAGENOME ASSEMBLY AND METHYLOME OF THE HUMAN GUT MICROBIOME USING SMRT SEQUENCING

Yoshihiko Suzuki[1], Suguru Nishijima[1,2,3], Yoshikazu Furuta[4], Wataru Suda[1,3,5,6], Kenshiro Oshima[1], Masahira Hattori[1,3,6], Shinichi Morishita[1]

[1]The University of Tokyo, Department of Computational Biology and Medical Sciences, Chiba, Japan, [2]National Institute of Advanced Industrial Science and Technology, Computational Bio Big-Data Open Innovation Laboratory, Tokyo, Japan, [3]Waseda University, Faculty of Science and Engineering, Tokyo, Japan, [4]Hokkaido University, Division of Infection and Immunity, Sapporo, Japan, [5]Keio University School of Medicine, Department of Microbiology and Immunology, Tokyo, Japan, [6]RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan

Assembly of complex metagenomic samples such as the human gut microbiome using short-read sequencers results in extremely fragmented genomic contigs due to repeat elements, and raw state of DNA methylation in microbiomes has been insufficiently characterized. PacBio sequencers based on single-molecule, real-time (SMRT) sequencing technology can produce reads long enough to span the most common repeat in microbes. SMRT sequencing also has a unique ability to determine positions and motif sequences of DNA methylation such as 6-methyladenine and 4-methylcytosine directly from uncultured microbes. We sequenced 13 fecal samples including one pair of biological replicates from healthy Japanese individuals using a PacBio RS II sequencer. To assemble PacBio long reads, we exploited the FALCON genome assembler. Although FALCON could manage to find structural variations without dividing contigs, it sometimes tended to merge contigs from distinct microbes; therefore, we used unitigs, which are unambiguous blocks in contigs that are shorter but more reliable contiguous sequences than contigs. To reconstruct circular sequences, we utilized results of unitig binning based on differential abundance among existing 106 short-read data and sequence composition. The N50 length of the PacBio contigs reached hundreds of Kbp, and we obtained 7 complete circular microbial chromosomes and 94 putative circular mobile genetic elements such as phages and plasmids. We also found several horizontally transferred genes that were related to antimicrobial resistance and prevalent among microbes. We identified a total of 503 manually curated DNA methylation motifs from all of the samples, including 431 motifs absent in the REBASE PacBio database. Moreover, we revealed a considerable inter-indivisual diversity of the methylation motifs. Assuming that methylation patterns are common in plasmids and their host bacteria, we assigned 46 plasmids to its host microbe, which has been difficult.

UNEVEN CONTRACTIONS AND EXPANSIONS IN THE HUMAN GENOME

Lifei Li, Leila Taher

University of Erlangen-Nuremberg, Department of Biology, Erlangen, Germany

The regulatory role of repetitive elements such as transposable elements (TEs) was already recognized by Barbara McClintock in the 1940s and 1950s. Nevertheless, TEs have long been dismissed as "junk" DNA, and are only now beginning to receive the attention they deserve. TEs comprise approximately half of the human genome. Their proliferation and evolution have had multiple impacts on the vertebrate genome, and these are only gradually being discovered. For example, it has been shown that TEs have substantially contributed to the expansion of binding sites for specific TFs, such as CTCF. The repetitive nature of TEs makes them difficult to analyze using currently available sequencing technologies and alignment algorithms.

Despite the widespread contribution of TEs to the mammalian genome, the distribution of TEs is not uniform. To further investigate the role of TEs in regulatory innovation, we analyzed a dataset of ~6,000 adjacent pairs of non-coding sequences in the human genome widely conserved across mammals. In particular, we calculated the fraction of the sequence in between the pairs overlapping with DNA, LTR, LINE or SINE transposons. We observed that relatively short sequences comprised smaller fractions of TEs than expected by chance. In contrast, relatively long sequences comprised greater fractions. Based on the histone signatures associated with the conserved non-coding delimiting these regions, we hypothesize that TE-related contractions and expansions in the mammalian genome often underlie the evolution of novel synergistic and antagonistic interactions between regulatory sequences.

INVESTIGATING NON-CANONICAL DNA MODIFICATION IN AN ANIMAL MODEL USING SINGLE-MOLECULE REAL-TIME SEQUENCING

Yusuke Takahashi[1], Massa Shoura[2], Andrew Fire[2], Shinichi Morishita[1]

[1]University of Tokyo, Department of Computational Biology and Medical Sciences, Kashiwa, Japan, [2]Stanford University, Departments of Pathology and Genetics, Stanford, CA

Although *C. elegans* had been thought to be free from DNA modification, Greer *et al.* reported 6-methyladenine (6mA) sites in *C. elegans* genome using the single-molecule real-time (SMRT) sequencing, a crosstalk between H3K4me2 and 6mA, and epigenetic inheritance of 6mA in *spr-5* mutant worms. In SMRT sequencing, one observes fluorescence signals from incorporated nucleotides during DNA copying on a nano-device called a zero-mode wave guide. The duration between neighboring pulse signals, called the inter-pulse duration (IPD), has been shown to increase when a base is methylated due to slower incorporation of nucleosides by the polymerase. In a situation such as *E. coli* where a fraction of bases are modified at high frequency, the ratio of an IPD to the average IPD of unmethylated bases (called IPD ratio) is useful in checking whether the focal base is methylated or not.

To further examine the characteristics of the well-studied *C. elegans* genome, we used five *C. elegans* strains (three closely-related laboratory strains and two wild strains) and predicted 6mA sites using the standard SMRT Pipe protocol (v1.8.139483). The coverage of reads anchored to the reference genome (ce10/WS220) ranged from 25-fold to 195-fold. In the strain with 195-fold read coverage, the ratio of estimated 6mA sites to whole adenines was 0.7 %, while the ratio was 0.03 % in the strain with 25-fold coverage, indicating that the incidence of putative 6mA sites was highly dependent on read coverage.

Each of the *C. elegans* DNA samples also carries DNA from the methylation-competent *E. coli* strain OP50 as a reference and positive control. From these analysis, IPD data from *E. coli* showed clear evidence of the *well-studied DAM methylation*, with unmethylated adenines 6mAs clearly separated according to IPD ratios. By contrast, IPD ratios for *C. elegans* were distributed along more of a continuum. This could be due to incomplete modification of a subset of bases, or alternatively to sequence-specific effect on SMRT-sequencing that could be unrelated to physical base modification in the template. We are currently working to resolve these possible situations.

CIRCADIAN OSCILLATIONS IN THE HUMAN SALIVARY MICROBIOME

Lena Takayasu[1,4], Wataru Suda[1,2,4], Rina Kurokawa[1], Elica Iioka[1], Yasue Hattori[1], Chie Shindo[3], Masahira Hattori[1,3,4]

[1]The University of Tokyo, Department of Computational Biology and Medical Sciences, Chiba, Japan, Japan, [2]Keio University School of Medicine, Department of Microbiology and Immunology, Tokyo, Japan, Japan, [3]Waseda University, Faculty of Science and Engineering, Tokyo, Japan, Japan, [4]RIKEN, Center for Integrative Medical Sciences (IMS), Yokohama, Japan, Japan

Human microbiomes throughout the body interact with various signals in response to biogeographical physiological conditions. We investigated how the salivary microbiome in the oral cavity is regulated by host-related signals. We found that microbial abundance and genes participating in maintaining the human salivary microbiome exhibited a global circadian rhythm. Analysis of 16S rRNA sequences of salivary microbial samples from six healthy adults collected at 4-h intervals for three days revealed that the microbial genera accounting for 68.4–89.6% of the total abundance significantly oscillated over ~24 h. These oscillation patterns varied widely between individuals, and the extent of circadian variations in individuals was lower than that of interindividual variations. Among the microbial categories showing oscillation, Firmicutes including Streptococcus and Gemella, and Bacteroidetes including Prevotella which were classified by aerobic/anaerobic growth and Gram staining were highly associated with circadian oscillation. Circadian oscillation was abolished by incubating the saliva in vitro, suggesting that host physiological changes are the strong contributor to microbial oscillation. Metagenomic sequencing analysis showed that circadian oscillation enriched the functions of environmental responses such as various transporters and two-component regulatory systems in the evening and those of metabolism such as vitamin and fatty acid biosynthesis in the morning.

SELECTION ACTS TO SUPPRESS STRUCTURAL POLYMORPHISM IN HUMAN Y CHROMOSOME AMPLICONS

Levi S Teitz[1], David C Page[1,2]

[1]Whitehead Institute and Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, [2]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA

Mammalian Y chromosomes contain large segmental duplications, called amplicons, that contain genes essential for fertility and that vastly differ between species in both genetic content and genomic structure. Previous work has shown that amplicons are highly susceptible to structural variation through non-allelic homologous recombination (NAHR) between amplicon copies. However, little is known about the selective constraints acting on those regions. Here, we used a novel method to detect amplicon copy number variants (CNVs) in 1214 men, using whole genome sequencing data primarily from the 1000 Genomes Project.

16.6% of men analyzed have at least one deleted or duplicated amplicon relative to the reference sequence. Most of these CNVs correspond to amplicon structures that are caused by one or more NAHR events, including several previously predicted but unobserved structures. In addition, we observed CNVs that are not explained by NAHR between amplicon copies, suggesting that other mechanisms also lead to amplicon copy number change.

We gained further insight by using a detailed evolutionary tree of the Y chromosomes that has been built from SNP data, made possible due to the Y chromosome's unique nature as a single haplotype. CNVs are present in almost all of the major branches of the Y chromosome phylogeny, called haplogroups. However, the reference copy number has been extraordinarily maintained, even in very divergent branches: both men in our sample from the A00 haplogroup, which diverged from the reference haplogroup over 200,000 years ago, contain the reference copy number of each amplicon. We observed several cases of amplicon "rescue," in which amplicons deleted in a group of men are restored in a member of the group through duplication of other nearly identical amplicon copies. Finally, the distribution of men with CNVs within the tree—many CNVs caused by recent mutations, and few ancient mutations—is incompatible with a model of neutral evolution.

Taken together, these results suggest that the Y chromosome's contribution to reproductive fitness is deeply sensitive to amplicon copy number. Reconciling this observation with the tremendous diversity of amplicon structure and gene content among species is the next step towards a better understanding of these crucial but mysterious regions of the genome.

GRAMENE: UNIFYING COMPARATIVE GENOMICS AND PATHWAY RESOURCES FOR PLANT COMMUNITIES

Marcela K Tello-Ruiz[1], Joshua Stein[1], Sharon Wei[1], Justin Preece[2], Sushma Naithani[2], Andrew Olson[1], Yinping Jiao[1], Parul Gupta[2], Sunita Kumari[1], Kapeel Chougule[1], Justin Elser[2], Bo Wang[1], James Thomason[1], Peter D'Eustachio[3], Robert Petryszak[4], Paul Kersey[4], Pankaj Jaiswal[2], Doreen Ware[1,5]
[1]Cold Spring Harbor Laboratory, Plant Genomics, Cold Spring Harbor, NY, [2]Oregon State University, Dept Botany & Plant Pathology, Corvallis, OR, [3]NYU School of Medicine, Dept Biochemistry & Molecular Pharmacology, New York, NY, [4]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, [5]USDA ARS NEA Plant,, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Understanding the relationships among different plant genomes and their constituent gene sets facilitates the identification of DNA sequences preserved over time in different organisms, e.g., genes that are essential to life and genomic signals that control gene function across species.
Gramene (http://www.gramene.org) is a powerful online resource for plants researchers and educators that provides easy access to reference data, visualizations and analytical tools for conducting cross-species comparisons. Learn the benefits of using Gramene (http://www.gramene.org) to enrich your lectures, accelerate your research goals, and respond to your organismal community needs.
Gramene's genomes portal hosts browsers for 44 complete reference genomes, each displaying functional annotations, gene-trees with orthologous and paralogous gene classification, and whole-genome alignments. Comparative plant gene trees are derived from a pre-computed phylogenetic analysis of protein-coding genes from all Gramene species, plus five representative vertebrate genomes used as outgroups. Build 52 included 61,582 gene family trees with nearly 1.5 million genes. SNP and structural diversity data, available for 11 species, are displayed in the context of gene annotation, protein domains and functional consequences on transcript structure (e.g., missense variant). Browsers from multiple species can be viewed simultaneously with links to community-driven organismal databases. Thus, while hosting the underlying data for comparative studies, the portal also provides unified access to diverse plant community resources, and the ability for communities to upload and display private data sets in multiple standard formats. Our BioMart data mining interface enables complex queries and bulk download of sequence, annotation, homology and variation data.
Gramene's pathway portal, the Plant Reactome, hosts over 240 pathways curated in rice and inferred in 66 additional plant species by orthology projection. Users may compare pathways across species, query and visualize curated expression data from EMBL-EBI's Expression Atlas in the context of pathways, analyze genome-scale expression data, and conduct pathway enrichment analysis.
Our integrated search database and modern user interface leverage these diverse annotations to facilitate finding genes through selecting auto-suggested filters with interactive views of the results.

Gramene is supported by NSF grant IOS-1127112, and partially from USDA-ARS (1907-21000-030-00D).

HOW TO ANALYZE 37,607 CANCER AND NORMAL GERMLINE EXOMES: A SPARK-BASED COMPUTATIONAL WORKFLOW TO UNCOVER NEW CANCER PREDISPOSITION GENES

Grace Tiao[1], Adam Kiezun[1], Mykyta Artomov[1,2,3], Paz Polak[1], Vijay Joseph[4], Namrata Gupta[1], Lauren Margolin[1], Ayellet Segre[1], Kenneth Offit[4], Mark J Daly[1,2], David Altshuler[5], Gad Getz[1,2]

[1]The Broad Institute, -, Cambridge, MA, [2]Massachusetts General Hospital, -, Boston, MA, [3]Harvard University, Department of Chemistry and Chemical Biology, Cambridge, MA, [4]Memorial Sloan-Kettering Cancer Center, -, New York, NY, [5]Vertex Pharmaceuticals, -, Boston, MA

Comprehensive case-control analysis of all the germline exomes available in the Cancer Genome Project (TCGA) has not yet been performed, despite the potential for this analysis to uncover heritable factors that predispose individuals to cancer. This is because analysis of the TCGA germline data is computationally challenging, due to the large sample and dataset size. Additionally, in case-control study designs, which compare samples from cancer patients with many thousands of appropriately matched “normal” (i.e., population) controls, samples must be matched computationally on both biological and technical characteristics — e.g., ancestry, sequencing coverage, and sequencing depth — and subjected to ordinary quality control (QC) analysis. Coordinating matching and QC across tens of thousands of samples ordinarily takes hundreds of hours of compute and custom, hands-on analysis to execute. Here, we have jointly genotyped 37,607 exomes, including most of the TCGA germline samples, as well as 26,388 population controls from the 1000 Genomes Project, the Exome Sequencing Project, the Type II Diabetes Consortium, and the Myocardial Infarction Genetics Consortium. We have developed an efficient computational pipeline in Hail, a new, scalable, open-source Spark-based tool developed at the Broad Institute for germline analysis, to quickly and efficiently filter out low-confidence variants, perform case/control sample matching, and run case-control association analyses. Some important features of this workflow include: (i) its relative speed; (ii) automatic customization to each of 29 unique tumor types and 29 meta-tumor type cohorts in the callset; (iii) the minimal manual intervention required to run the workflow; (iv) its reproducibility; and (v) its parallelization. Variants and samples that have been matched and filtered by the workflow are then subjected to case-control association of rare coding variants, with a focus on variants with predicted deleterious effects on protein function and expression. For each of the 58 tumor-specific and meta-tumor cohorts, we have performed gene-based burden tests to detect genes enriched for rare, deleterious variants, identifying several new candidate predisposition genes, which we are currently assessing. This study demonstrates the power of analyzing large cohorts of exome sequences across many different tumor types, made feasible by an enhanced computational infrastructure, for discovering new risk genes for cancer.

COORDINATION OF ALTERNATIVE RNA PROCESSING EVENTS IN HUMAN CELL-LINES, TISSUES, NEURONS AND SINGLE CELLS.

Fereshteh Jahanbani[1], Morten Rasmussen[1], Erich Jaeger[2], Steven Sloan[3], Dana Wyman[1], Ali Moshrefi[2], Xihe Xie[4], Chuying Xia[4], Fatemeh Jahanbani[1], Feng Chen[2], Carlos Bustamante[1], Ben Barres[3], Michael Snyder[1], Hagen Tilgner[4]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]Illumina, Inc, San Francisco, CA, [3]Stanford University, Department of Neurobiology, Stanford, CA, [4]Weill Cornell Medicine, Brain and Mind Research Institute, New York City, NY

Alternative RNA processing can affect multiple sites in a gene with alternative outcomes at each site. We have recently employed long read RNA sequencing to monitor the combinations of such variable sites on RNA molecules. We showed that important genes of the nervous system employ dependent (or non-random) ways of combining alternative exons that are distant of one another in the mature RNA into full-length molecules[I]. Here, we show that this phenomenon is ten times more frequent than we had previously shown. We find such non-random exon combinations in cancerous cell lines like MCF7 and K562 but also in a non-cancerous cell line (GM12878). By developing and using very low input protocols, that produce 0.5-12 kilobase reads we furthermore trace this phenomenon into sorted fetal cortical neurons and single neuronal cells. Coordination correlates across samples and in some cases alternative exon pairs follow similar rules of coordination across samples, although percent-spliced-in (PSI) values change strongly, showing that both splicing decisions are regulated in a coordinated manner. We furthermore show the existence of coordination events involving variable sites that are not defined by the spliceosome, some of which show different patterns than coordination events between pairs of alternative exons. A number of genes employ coordination events, in which one alternative exon introduces a frame shift that is corrected by the second alternative exon, so that the constitutive exon(s) in between encode a double-frame open reading frame (dfORF).

I. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, Bustamante CD, Rasmussen M, Snyder MP. Nat Biotechnol. 2015 Jul;33(7):736-42. doi: 10.1038/nbt.3242.

DEVELOPMENTALLY PROGRAMMED CHROMOSOME ELIMINATION IN LAMPREYS: REVEALING 40 MILLION YEARS OF SHARED ANCESTRY

Vladimir A Timoshevskiy, Jeramiah J Smith

University of Kentucky, Biology, Lexington, KY

The sea lamprey (*Petromyzon marinus*) represents one of the few vertebrate species, known to undergo large-scale programmed genome rearrangement (PGR) over the course of its early embryonic development, loosing ~20% its genome from all somatic cell lineages. Previously we determined that PGR in the sea lamprey is accompanied by extensive chromosome lagging during anaphase. Lagging chromosomes are first observed after the 6-th embryonic cleavage (late 1st day post fertilization – dfp) and are no longer observed after 2.5 dfp. Development of DNA-probes that specifically label eliminated chromosomes and concurrent hybridization with differentially labeled repetitive DNAs allow us identify key subcellular features of eliminating cell divisions in the sea lamprey. 1) At early metaphase stages germline-restricted chromosomes exhibit decelerated motion toward the equatorial plate; 2) to the time of equatorial alignment, both germ-line restricted and somatic, show dense equatorial location; 3) in anaphase, eliminated chromosomes exhibit slower motion and acquire a stretched antiparallel morphology originating from telomere-telomere contacts; 4) upon reaching interphase lagging chromatin is packaged into micronuclei and ultimately degraded. Here we report first observation of lagging anaphases and micronuclei in early-stage embryos of a second lamprey species, the Pacific lamprey (*Entosphenus tridentatus*), which diverged from the sea lamprey lineage 40 million years ago. Anaphase lagging and micronuclei formation pattern suggests, that pacific lamprey also undergoes chromosome elimination but eliminates less DNA than sea lamprey. Interspecific comparative hybridization experiments indicate that DNA eliminated in Pacific lamprey shares homology with a subset of sequences eliminated in Sea lamprey and further reveals broad evolutionary changes in the repeat content of both retained and eliminated chromatin over the last ~40 Million years of lamprey evolution. Overall, these analyses support the idea that PGR represents an ancient and evolutionarily stable strategy for regulating inherent developmental/genetic conflicts between germline and soma.

TRANSPOSABLE ELEMENTS ARE THE PRIMARY SOURCE OF NOVELTY IN THE PRIMATE GENE REGULATION

Marco Trizzino*[1], YoSon Park*[1], Marcia Holsbach-Beltrame[1], Katherine Aracena[1], Katelyn Mika[2], Minal Caliskan[1], George H Perry[3], Vincent J Lynch[2], Christopher D Brown[1]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]University of Chicago, Human Genetics, Chicago, IL, [3]Pennsylvania State University, Anthropology and Biology, University Park, PA

Gene regulation plays a role in the evolution of phenotypic diversity. We investigated the evolution of liver promoters and enhancers in six primate species, performing ChIP-seq for two histone modifications (H3K27ac and H3K4me1) to profile cis-regulatory elements (CREs), and RNA-seq to characterize gene expression in the same individuals. While the evolution of mammalian CREs has been investigated, previous studies primarily focused on binary outcomes (absence or presence of a ChIP-seq peak in each species), thus omitting crucial information provided by high-throughput sequence data. To maximize our discovery power, we compared CRE activity across species by testing differential ChIP-seq read depths directly measured for orthologous sequences. We show that the primate regulatory landscape is largely conserved across the lineage, with over 60% of the tested human liver CREs showing similar activity in all of the species. Conserved CRE function is associated with sequence conservation, proximity to coding genes, cell-type specificity of CRE function, and transcription factor binding. Moreover, newly evolved CREs are enriched in immune response and neurodevelopmental functions, while conserved CREs bind master regulators, demonstrating that CREs contribute to species adaptation to the environment, while maintaining essential functions intact. The primate CREs are enriched for transposable elements (TEs). In particular, newly evolved CREs are enriched in young TEs, mostly Long-Terminal-Repeats (LTR) and SINE-VNTR-Alus (SVAs), that affect gene expression, as showed by RNA-seq. On the other hand, only 17% of conserved CREs overlap a TE, suggesting that expression variations of their target genes might be under selection. We identified specific genomic features driving the functional recruitment of newly inserted TEs, and tested the cis-regulatory activity of 69 TE subfamilies, covering all of the main TE classes, by luciferase reporter assays. These assays showed that the majority of the tested TEs (95.6%) are functional, acting as either transcriptional activators or repressors. In conclusion, by incorporating multiple validation methods following high-throughput ChIP-seq and RNA-seq data from six primate species, we demonstrated the crucial role of TEs in the primate gene regulation, and illustrated the potential mechanisms underlying evolutionary divergence among the primate species through the dynamic noncoding genome.

SOCIAL HIERARCHIES AND THE DETERMINANTS OF THE IMMUNE RESPONSE IN WILD BABOONS

Amanda J Lea[1], Mercy Y Akinyi[1], Ruth Nyakundi[2], Peter Mareri[2], Fred Nyundo[2], Thomas Kariuki[2], Elizabeth A Archie[3], Susan C Alberts[1,4], Jenny Tung[1,4]

[1]Duke University, Biology, Durham, NC, [2]National Museums of Kenya, Institute of Primate Research, Nairobi, Kenya, [3]University of Notre Dame, Biological Sciences, Notre Dame, IN, [4]Duke University, Evolutionary Anthropology, Durham, NC

Social status strongly predicts health and survival in humans and other social animals. Recent evidence in captive primates suggests that this phenomenon arises in part through its effects on gene regulation of the immune response, a fundamental determinant of Darwinian fitness. In the absence of other environmental and demographic variation, these effects are pronounced. However, whether social status is a major predictor of immune gene regulation in natural populations, in which behavioral, genetic, and demographic factors are all at play, remains completely unexplored.

To address this gap, we conducted a paired control and ex vivo lipopolysaccharide (LPS) challenge (a model for bacterial infection) in whole blood from 67 wild adult baboons, subjects of a 46-year longitudinal study in the Amboseli ecosystem of Kenya (n=134 mRNA-seq samples total). Controlling for cell type heterogeneity, dominance rank was a major predictor of immune gene expression for male baboons (1501 rank-associated genes), but had no detectable effects in females. Prime age, high-ranking males were particularly distinct from other males (2125 differentially expressed genes), likely reflecting their unique physiological profiles and the energetic costs of competing for dominance. In support of this idea, high rank-associated genes were strongly enriched for genes that respond to glucocorticoid treatment (p=9.53 x 10-4). In contrast, genes more highly expressed in low ranking individuals also tended to be upregulated with old age and in response to infection.

To put the rank effects we observed in context, we compared their predictive power to that of LPS exposure, age, sex, and genotype (genetic effects are common in this population: 1017 genes had detectable cis-eQTL in this data set). Variable selection using LASSO revealed that genotype and exposure are the most consistently important pieces of information needed to predict variation in gene expression levels (retained in 86% and 82% of models for 8031 genes, respectively). However, social status was also retained in the majority of models (60%), indicating that it, too, has widespread effects on immune gene regulation. Finally, rank-associated genes were strongly enriched for genes that respond to LPS stimulation (p<10-15), and both sets were concentrated in key innate immune pathways such as NFkB signaling and natural killer cell mediated cytotoxicity. Together, our results provide the first demonstration that social status is strongly linked to immune gene regulation in a natural population. In particular, males that reach the top of baboon social hierarchies have a distinct gene regulatory profile, suggesting that social interactions are a key factor to consider in explaining variation in immune defense.

FUNCTION AND REGULATION OF INTRON RETENTION IN THE HUMAN BLOOD CELL LINAGE

Sebastian Ullrich[1,2], Alessandra Breschi[1,2], Dmitri Pervouchine[3], Roderic Guigo[1,2]

[1]CRG, Bioinformatics and Genomics, Barcelona, Spain, [2]UPF, Biomedicine, Barcelona, Spain, [3]Skoltech, Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo, Russia

As a new putative layer of gene regulation between transcript production and translation, Intron retention (IR) has been found in various vertebrate species and tissues. In particular it has been observed during blood cell differentiation. A recent study suggests it regulates granulocyte differentiation. Furthermore it is also found in erythrocytes and megakaryocytes. Using the deep transcriptome data produced by the Blueprint project during haematopoiesis, we examined differentiated blood cell types for intron retention. We found neutrophils, monocytes and B-cells to have highest IR levels, whereas macrophages and T-cells have the lowest. Tracing retention patterns in differentiating cells we observed an increase of IR in neutrophil development with a drastic peak when bone marrow cells are released into the blood stream. In B-cells we found a steady rise of IR levels from bone marrow precursors to marginal zone cells residing in the spleen. Memory B-cells with similar transcriptome profiles to marginal-zone B-cells display similar levels of IR. Interestingly, those populations are expected to be first exposed to pathogens circulating in the blood stream triggering an initial immune response. In contrast, cells that enter affinity maturation in the germinal centre of the spleen decrease retention while they get more proliferation active. Plasma cells that define the terminal stage of differentiation have the lowest retention levels in contrast to a qualitative increase during neutrophil maturation. As functional aspects of IR in granulocytes where reported before we focused on B-cells. The most widely affected group of genes is related with RNA processing and stability which is supported by previous publications suggesting a self-regulatory mechanism of splicing related genes. Besides we observed genes to be affected in histone deacetylation and IFκB signalling. For several histone deacetylases we observed rising IR while their transcript abundances decrease in marginal zones. Trying to understand the mechanism in which dynamic splicing processing is achieved we analysed binding sites and expression of splicing factors. While overall gene expression levels stay constant we find splicing factors to lower their transcript abundances when IR is rising. In retained introns we find binding sites for SR and hnRNP proteins to be enriched. In conclusion we find IR to be a self-regulating mechanism shaping the transcriptomes of B-cell populations involved in the primary immune response.

USING JOINT PHASING OF GERMLINE AND SOMATIC VARIANTS FOR ASSESSING THE LANDSCAPE OF REGULATORY VARIATION IN HUMAN CANCERS

Roland F Schwarz[1,2], Lara Urban[2], Stefan Dentro[3,4], Peter Van Loo[3], Oliver Stegle[2]

[1]MDC-BIMSB, Berlin, Germany, [2]EMBL-EBI, Hinxton, United Kingdom, [3]Francis Crick Institute, London, United Kingdom, [4]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Cancer is characterized by ubiquitous somatic genetic and epigenetic changes that interact with the germline genetic background and modify gene regulation, the precise mechanisms of which are poorly understood. Allele-specific expression (ASE) provides an internally controlled read out to assess allele-specific genetic effects and is particularly well suited to quantify such effects.

Here, we present a systematic analysis of the effect of somatic and germline variation on ASE using 1,200 matched RNA-DNA samples across 20 cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG). We combine evident phasing of germline variants in somatic copy-number alterations (SCNAs) with statistical phasing on regions lacking SCNAs. Through read overlap information we further phase >10M SNVs (~20% of genome-wide somatic SNVs) to the germline phasing map, for which we assess regulatory effects on 5.8M individual gene/sample pairs for which ASE could be quantified. Overall, we observe that genes with allelic expression imbalance (AEI) are substantially more frequent than in matched normal samples, and we detect cancer type-specific patterns of AEI aggregation along the chromosomes.

We integrate germline and somatic information into a joint statistical model of ASE, which accounts for sub-clonality, varying tumour ploidy and normal admixture. Our model predicts the presence of ASE and the effect direction of individual SNVs in different variant classes, as well as using CN state, mutation timing and evidence from germline eQTLs. We further employ a convolutional neural network (CNN) and leverage in-silico predictions for the effect of individual mutations on regulatory features in different cellular contexts in our model.

We find that ~80% of ASE variability can be attributed to SCNAs, with SNVs and germline variants each explaining about 1-2% of the remaining variability. We notably observe widespread nonsense-mediated decay triggered by protein truncating variants across tissue types, revealing how mutations in splice sites disrupt exon usage, and determine regulatory changes in 3’ and 5’ UTRs that consistently up- or down-regulate affected alleles. We additionally detect individual genes for which variation in ASE can be explained through associations with recurrent somatic mutations independent of SCNAs, and which are enriched for cell adhesion pathways and oncogenes such as POLQ or BCL2. We further found that early mutations are significantly more likely to impact ASE than late-occurring passengers.

Finally, we validated our model using RNA-Seq data from matched normal tissue, demonstrating the ability to accurately predict regulatory changes caused by somatic changes. In summary, this study represents a comprehensive and in-depth investigation into allelic deregulation across human cancers. The extent of rearrangement-derived AEI questions pre-conceived notions about the importance of SNVs for up- or down-regulation of cancer-related genes and, instead, suggests structural and CN changes as the more important contributors to cancer regulation and fitness.

ANNOTATIONS CAPTURING TISSUE SPECIFIC TRANSCRIPTION FACTOR BINDING EXPLAIN A LARGE FRACTION OF DISEASE HERITABILITY

Bryce van de Geijn[2], Hilary Finucane[1,2], Steven Gazal[2], Alexander Gusev[2], Farhad Hormozdiari[2], Xuanyao Liu[2], Yakir Reshef[3], Alkes Price[2]

[1]Harvard TH Chan School of Public Health, Epidemiology, Boston, MA,
[2]Massachusetts Institute of Technology, Mathematics, Cambridge, MA,
[3]Harvard University, Computed Science, Cambridge, MA

It is now clear that non-coding variation plays a major role in complex diseases and that prioritizing regulatory regions of the genome is vital. However, regulatory information is tissue or cell type specific, and may be missing in the most relevant tissues. We investigate methods to create and prioritize tissue specific annotations of gene regulation. We achieve our best results by overlapping transcription factor annotations that were not cell type specific with chromatin measurements based on ChIP-seq in relevant tissues.

Lymphoblastoid cell lines (LCLs) provide an excellent test bed as the cell type with the most complete data, while focusing on the histone modification H3K27ac as a phenotype provides maximum power to compare annotations. Thus, we first analyze the heritability of H3K27ac levels using stratified LD score regression (Finucane et al. 2015 Nat Genet). We assess each annotation based on two metrics: 1) enrichment in heritability explained by an annotation and 2) fraction of SNPs required to explain 50% of heritability when added to the non-cell type specific baselineLD model (Gazal et al. BioRxiv 2016). We jointly analyze 136,264 histone peaks using data previously collected in 65 Yoruba LCLs (Grubert et al. Cell 2015). We observed striking enrichments for heritability in the overlap of H3K27ac peaks and transcription factor binding sites, where 3.6% of the SNPs in the genome explain 36% of histone modification SNP-heritability (10.2x enrichment, P = 1.5x10-20). Adding this annotation to the baselineLD model also reduced the fraction of SNPs required to explain 50% of H3K27ac heritability by 24% (from 7.57% to 5.57%).

Motivated by the results on H3K27ac phenotypes, we applied the same approach to summary statistics for 4 autoimmune diseases (celiac disease, Crohn’s disease, lupus and rheumatoid arthritis; average N = 30,789), constructing H3K27ac-TFBS overlaps for many cell times using Roadmap H3K27ac annotations data from Roadmap Epigenome. The LCL H3K27ac-TFBS annotation was highly enriched for disease heritability (2.0% of SNPs explain 50% of SNP-heritability on average; 25x enrichment, P = 1.03x10-8). Adding this annotation to the baselineLD model also reduced the fraction of SNPs required to explain 50% of autoimmune disease heritability by 12% (from 1.30% to 1.15%).

EVOLUTION OF ENHANCERS REGULATING THE DYNAMIC TRANSCRIPTIONAL RESPONSE OF INNATE IMMUNE CELLS

Pranitha Vangala[1], Elisa Donnard[1], Shaked Afik[2], Sean McCauley[3], Barbara Tabak[1], Patrick McDonel[1], William Diehl[3], Anetta Nowosielska[3], Nir Yosef[2], Jeremey Luban[3], Manuel Garber[1,3]

[1]University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA, [2]University of California Berkeley, Center for Computational Biology, Berkley, CA, [3]University of Massachusetts Medical School, Program in Molecular Medicine, Worcester, MA

Surprisingly most of the non-coding elements like enhancers, long non-coding RNAs are poorly conserved even between the closely related species. A recent comparative genomic analysis of 20 mammals showed that loss of enhancer activity occurs at a much more rapid rate than their underlying sequence divergence [1]. Despite of extensive comparative studies of enhancer activity, there have been very few comparative analysis that integrate epigenetics with gene expression.

Here we revisit comparative analysis of enhancer elements in the context of their impact on gene regulation. Specifically we carry an integrative comparison of enhancer activity and gene expression of human-mouse innate immune cells in response to pathogen challenge with a focus on the genes that make up the regulatory network that controls innate immune responses to pathogen.

Our comparative analysis shows consistent with previous studies [1], overall enhancer conservation is low. This result is not surprising given the divergence of the basal gene expression levels: only 25% of the genes that are expressed in both species have conserved response to LPS. However, we found that conservation of enhancer activity of the genes that are strongly induced and that tend to have a complex regulatory architecture composed of many different enhancers is 2 fold greater than the background. This suggests that purifying selection may act strongly on regulatory regions that are part of cell type specific responses that are critical to the organism fitness.

We also show how successive waves of mobile element activity have reshaped the regulatory network. We find very specific mobile element families significantly enriched within regulatory regions of pathogen responsive genes. Together our results shed light on the role of conserved and species specific regulatory elements in driving the transcriptional dynamics of innate immune cells in response to pathogens.

1. Villar, D. et al. Enhancer Evolution across 20 Mammalian Species. Cell 160, 554–566 (2015).

SELECTION ON GENE DUPLICATES FOLLOWING WHOLE GENOME DUPLICATION: INSIGHTS FROM SALMONID GENOMES.

Srinidhi Varadharajan*[1], Simen R Sandve*[2], Ole K Tørresen[1], Sigbjørn Lien[2], Asbjørn L Vøllestad[1], Sissel Jentoft[1], Alexander J Nederbragt[1,3], Kjetill S Jakobsen[1]

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo NO-0316, Norway, [2]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway, [3]Biomedical Informatics Research Group, Department of Informatics, University of Oslo, Oslo NO-0316, Norway

*These authors contributed equally to this work.

Genome doubling events have played a significant role in the evolution of vertebrate genomes as the resulting functional redundancy is believed to provide a raw material for evolutionary innovation. Early vertebrates are known to have undergone at least two rounds of whole genome duplications (WGD), with an additional third round in the teleost ancestor ~320 million years ago (MYA). A more recent fourth round of WGD occurred in the ancestor of salmonids ~80-100 MYA, following which speciation preceded complete cytological rediploidization (progressive return to the diploid state), making salmonids an exemplary system to investigate the evolutionary consequences of WGD.

We assembled a 1.48 Gb draft assembly of European grayling (*Thymallus thymallus*), a member of the earliest diverging salmonid lineage, and used this novel genome resource to understand the importance of selective forces on the evolution of genome regulation following WGD. Comparative analysis of gene expression data from grayling and Atlantic salmon reveals that a large proportion of brain and neuron related duplicates have maintained conserved tissue-specific expression patterns over 60 million years of independent evolution after the basal split in Salmonidae. We also identified groups of duplicated genes reflecting expression evolution suggestive of adaptive divergence predating the grayling-salmon split (>60 MYA) and many cases of more recent lineage specific regulatory divergence. Our results demonstrate that selection is important in shaping genome function post WGD in salmonids and highlight potential differences in selective pressures on distinct biological functions.

RAPID PHENOTYPE-DRIVEN IDENTIFICATION OF CAUSATIVE GENETIC VARIANTS USING IOBIO TOOLS

Matthew Velinder[1], Alistair Ward[1,2], Tonya Di Sera[1], Chase Miller[1,2], Julie Feusier[3], Karin Chen[4], David Viskochil[5], Gabor Marth[1,2]

[1]USTAR Center for Genetic Discovery, Eccles Institute of Human Genetics, Salt Lake City, UT, [2]Frameshift Genomics, LLC, Boston, MA, [3]Department of Human Genetics, Eccles Institute of Human Genetics, Salt Lake City, UT, [4]Pediatric Immunology and Rheumatology, Department of Pediatrics, Salt Lake City, UT, [5]Pediatric Genetics, Department of Pediatrics, Salt Lake City, UT

Current sequencing technologies allow for rapid whole genome and exome sequencing, bringing the promise of personalized medicine within reach. However, current sequence analysis tools still heavily rely on trained experts to run complex, computationally intensive command line software, typically on dedicated high performance computing clusters. Additionally, the output from these programs are often large text files with highly technical and specific formatting. These data can be cumbersome to medical professionals lacking extensive computational or bioinformatics expertise. Furthermore, it remains difficult to prioritize and identify disease causing variants among the hundreds of thousands of variants typically reported, especially when disease-associated genes are not known. We have developed *gene.iobio* (http://gene.iobio.io), an intuitive, real-time, phenotype-driven, web-based application for the analysis of genetic variants. Importantly, these tools are visually-driven and capable of being run within a web browser on a typical personal computer, allowing for analysis to be performed in real time by clinicians and medical professionals with limited bioinformatics experience. Here we demonstrate the utility of this powerful and intuitive analytical tool to identify causative genetic variants in Treacher-Collins Syndrome and severe combined immunodeficiency (SCID).

Treacher Collins syndrome is a disorder of craniofacial development ranging from mild to severe clinical phenotypes. Using the built-in genotype-phenotype tools[1] of *gene.iobio* we visually inspected variants within genes associated with Treacher Collins syndrome, including *TCOF1*. We identified compound heterozygous mutations, p.pro1213arg and p.ala1427val, inherited from the proband's mother and father respectively. Additional sequencing demonstrated these variants are also present in an affected maternal aunt and an affected maternal grandfather, providing further evidence for the inheritance of variants and phenotype in the family. Similarly, we also used *iobio* tools to discover a causative variant in an individual with SCID. Phenotype information was input into *gene.iobio* and variants within candidate genes were inspected visually. Strikingly, we identified a loss of function frameshift variant in *FOXN1* (p.pro473fs), a potent transcriptional regulator of thymus development and subsequent T cell progenitor function and differentiation.

Importantly, these phenotype-driven analyses required no *a priori* knowledge of genes associated with the given disorder and were performed in real time within an intuitive interface. These findings have informed ongoing and future functional and molecular experiments. In conclusion, these use-cases demonstrate how *iobio* tools greatly improve the speed of genetic diagnosis and drastically reduce the computational expertise required for clinical genome sequencing analysis.

CLINICAL REPORTING FROM INTEGRATED GENOMIC DATA ON THE NEPTUNE PLATFORM

Eric B Venner[1], Tsung-jung Wu[1], Matthew Bainbridge[3], Christie Kovar[1], Theodore Chiang[1], Magalie S Leduc[2], Mullai Murugan[1], Kimberly Walker[1], William Salerno[1], Donna Muzney[1], Richard Gibbs[1]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, [3]Rady Children's Institute for Genomic Medicine, San Diego, CA

Future genomic discoveries will be best driven by pairing large quantities of genomic data with deep phenotype information, ideally a full electronic medical record. Achieving this requires full integration of genomic data into the clinic. This clinical integration, in turn, demands automated processing to provide timely and cost-efficient reporting. To address this need, we have developed Neptune, an automated analytical platform to sign-out and deliver clinical reports.

Initial data intake occurs in a HIPAA compliant environment on DNAnexus, and samples are de-identified before moving into the CLIA lab. After analysis with the Human Genome Sequencing Center's Mercury Pipeline, Neptune's custom annotation software identifies variants of putative clinical relevance for manual review and possible addition to a "VIP" database of clinically relevant variation. This resource draws on both public resources (ClinVar, literature review) and internal data sets accessed via Anton, the HGSC's Hadoop-based data store. The VIP database currently houses 20,872 SNPs and 3,946 indels, and contains a curated set of copy number variants (CNVs) annotated with internal frequency data.

Using Neptune's manual review interface, a clinical geneticist updates the VIP database accordingly. Once all variants have been categorized, Neptune extracts reportable, pathogenic variants using the VIP set, and Neptune outputs an automated clinical pre-report populated with prioritized variants (or a negative report if no relevant variants are found), descriptive text, and coverage statistics produced by the HGSC's ExCiD software. Clinical reports integrate SNVs, Indels, CNVs and CPIC level A pharmacogenomic variants.

Early applications include reporting for the National Institutes of Health eMERGE network where more than 14,000 samples and a panel of 109 genes will be processed in less than three years, as well as the Right10k pharmacogenomics project in collaboration with the Mayo Clinic.

CHARACTERIZATION OF RECOMBINATION EVENTS DRIVING ANTIGENIC VARIATION IN THE LYME SPIROCHETE BY NEXT GENERATION SEQUENCING OF FULL-LENGTH *VLSE* VARIANTS

Theodore B Verhey[1,3], Mildred Castellanos[1,2,3], George Chaconas[1,2,3]

[1]University of Calgary, Biochemistry & Molecular Biology, Calgary, Canada, [2]University of Calgary, Microbiology, Immunology & Infectious Diseases, Calgary, Canada, [3]University of Calgary, Calvin, Phoebe and Joan Snyder Institute for Chronic Diseases, Calgary, Canada

Lyme disease is a chronic and debilitating infection caused by *Borrelia burgdorferi* spirochetes, transmitted by a tick vector, and maintained in a natural reservoir primarily of the white-footed mouse. Lyme borreliae encode the variable surface antigen VlsE to evade the acquired immune response and maintain persistent infection in mammals. *vlsE* undergoes unidirectional, segmental gene conversion from a series of adjacent and unexpressed cassette sequences which are homologous to the variable region of the *vlsE* expression site. Aside from a few hints as to functional requirements for antigenic switching, the mechanism is not known. In this study, we developed a new mode of PacBio sequencing for very high quality full-length variant sequencing, and have applied it to the study the variants that are generated over the time course of infection in both wild-type and immunodeficient (SCID) mice. We also developed the first unbiased and reproducible alignment method for the variants, as well as a method to reconstruct the history of recombination events for each variant from nucleotide alignments. Using these reconstructions, we measure a multitude of parameters, including the length of recombination events, their frequency and rate, and unexpressed cassette usage. Data will be presented that show clustering of recombination events in time and in space, and that supply evidence for multiple mechanisms of non-templated mutation, including polymerase slippage. Finally, we assess the relative abundance of SNPs in wild-type and immunodeficient mice and will present evidence for the existence and tension of both a diversifying selection and functional constraints on the VlsE protein variants.

SINGLE CELL TRANSCRIPTOMICS IN MICE WITH A HUMANIZED VERSION OF FOXP2

Benjamin Vernot*, Gray Camp*, Wulf Hevers, Barbara Treutlein, Svante Pääbo

Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany

Foxp2 is known to affect speech and language in humans, and is conserved in the mammalian lineage. A striking exception to this conservation are two amino acid changes which occurred and fixed in the modern human lineage, suggesting the possibility that these changes could contribute to the unique ability of humans to speak. Mouse models are consistent with this hypothesis - mice carrying the human version of Foxp2 learn faster in the presence of environmental cues, and have differences in their vocalization patterns. However, the molecular function of these two amino acid changes have been difficult to untangle, and may vary between brain regions and even cell types. To this end, we performed single cell RNA-seq on striatal and cortical neurons in humanized and wildtype mice, and identify significant expression changes associated with the humanized version of Foxp2. Differentially regulated genes are enriched for Foxp2 ChIP-seq occupancy in promoter regions and synapse-related cellular components (GO), and include genes associated with visual learning. Interestingly, many of these changes are cell-type specific, and some involve promoter-switching, suggesting interaction partners play a role in these human-specific changes.

* authors contributed equally

BASE CALLING AND INDEXING OXFORD NANOPORE READS

Vladimir Boza, Brona Brejova, Tomas Vinar

Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Bratislava, Slovakia

Recently, we have developed an open source DeepNano (Boža et al. 2016) base caller for Oxford Nanopore reads based on recurrent neural networks. On R7 data, our base caller outperforms alternatives, while on R9 data, accuracy of DeepNano is slightly worse than Albacore and Nanonet base callers released by Oxford Nanopore.

The advantage of DeepNano, however, is in its flexibility. Under the default settings, DeepNano is faster, and by adjusting the size of the underlying network, it is possible to further trade accuracy for speed. Fast base calling is essential in applications such as selective on-device sequencing (ReadUntil, Loose et al. 2016) and in settings where using cloud services, as supported by Oxford Nanopore, is impractical. It is also possible to adaptively retrain the network, which can be used to leverage data that is otherwise impossible to base call through standard means (e.g., due to modifications or damage to the DNA).

Finally, we examine the dynamic-time-warp (DTW, Sankoff and Kruskal 1983) scheme for classification of reads and show that for applications such as ReadUntil, the method suffers from low specificity at high sensitivity. We demonstrate that by adjusting methods for scaling raw data, the sensitivity vs. specificity tradeoff can be much improved.

References:

Vladimír Boža, Broňa Brejová, Tomáš Vinař. DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads (2016). arXiv:1603.09195

Matthew Loose, Sunir Malla, Michael Stout. Real-time Selective Sequencing Using Nanopore Technology (2016). Nature Methods 13:751-754

David Sankoff, Joseph Kruskal. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison (1983). Addison-Wesley

GENETIC VARIATION MODULATES MULTIPLE MOLECULAR PHENOTYPES IN T2D SUBJECTS: A DIRECT STUDY

A Viñuela[1], J Fernandez[2], A Kurbasic[3], MG Hong[4], S Sharma[5], C Brorsson[6], J Adamski [5], J Schwenk[4], E Pearson[7], S Brunak[6], P Franks[3], M I McCarthy[2], E T Dermitzakis[1]

[1]University of Geneva, Genetic Medicine and Development, Geneva, Switzerland, [2]University of Oxford, WTCHG, Oxford, United Kingdom, [3]Lund University, Clinical Sciences, Malmö, Sweden, [4]Karolinska Institutet, SciLifeLab, Solna, Sweden, [5]Helmholtz Zentrum, Genome Analysis Center, Munich, Germany, [6]Technical University of Denmark, Systems Biology, Lyngby, Denmark, [7]University of Dundee, Molecular and Clinical Medicine, Dundee, United Kingdom

While GWAS have produced many associations for T2D, these have proved of limited use in understanding disease aetiology. Instead, a study design that combines multiple dimensions of data could help explain the genetic basis of complex traits. The DIRECT study involves pre-diabetic individuals and newly diagnosed patients with T2D with extensive phenotyping. Blood and plasma samples from ~3,100 individuals were used to produce genotypic, transcriptomic (RNA-seq), proteomic (multiplexed immunoassays) and metabolomic (both targeted and non-targeted) data. Clinical and lifestyle quantitative phenotypes were also measured at time of recruitment and in follow up visits.

QTL were mapped in each molecular data-type. Using RNA-seq we identified 17,714 eQTL (FDR 5%, Π1=0.89%), we found a significant eQTL for every protein coding gene and lincRNA. Out of 263 proteins, we found 32 (12%) to be associated with a local genetic variant (pQTL, FDR 5%, Π1=0.07). Using targeted metabolite data we observed that 136 of 154 metabolites (89%) were significantly associated to at least one SNP (metQTL, FDR 5%, Π1=0.83). Combining these results, we estimated that 23.8% of metQTL and 15.4% of the pQTL were also associated with the expression of a gene. One strong example of a genetic variant acting in more than one trait is rs174530, which affects expression of *FADS1* and the levels of circulating diacyl-phosphatidylcholines C36:1. This metabolite has previously been associated to decreased T2D risk. In contrast, we identified rs10445391 as a pQTL for *CCL16* but not an eQTL, as the relevant gene is expressed at very low levels in whole blood. The GTEx data shows the gene is expressed and its protein secreted only in liver, where rs10445391 is a significant eQTL.

Finally, we also looked at how relationships between genotype and molecular phenotypes can be perturbed by disease status. A scan for eQTL with different expression effects in T2D subjects and pre-diabetic individuals identified 250 significant interactions (FDR 1%). By understanding how genetic effects are perturbed by disease, we hope to better understand and stratify those at highest risk of developing T2D and other complications.

THE PROMOTER LANDSCAPE OF INFLAMMATORY BOWEL DISEASE (IBD).

Morana Vitezic*[1,2], Mette Boyd*[1,2], Jette Bornholdt*[1,2], Malte Thodberg*[1,2], Kristoffer Vitting-Seerup[1,2], Yun Chen[1,2], Mehmet Coscun[1,2,3], Yuan Li[2,3], Anders Gorm Pedersen[4], Kerstin Skovgaard[5], Robin Andersson[1], Thilde Bagger Terkelsen[1,2], Berit Lilje[1,2], Jesper Troelsen[6], Jakob Benedict Seidelin[3], Ole Haagen Nielsen[3], Jacob Tveiten Bjerrum*[3], Albin Sandelin*[1,2]

[1]University of Copenhagen, Department of Biology, Copenhagen, Denmark, [2]University of Copenhagen, Biotech Research and Innovation Center, Copenhagen, Denmark, [3]University of Copenhagen, Herlev Hospital, Copenhagen, Denmark, [4]Technical Univeristy of Denmark, Department of Systems Biology, Copenhagen, Denmark, [5]Technical University of Denmark, National Veterinary Institute, Fredriksberg C, Denmark, [6]Roskilde Univeristy, Department of Science, Roskilde, Denmark

Inflammatory bowel disease is a chronic inflammatory bowel disorder. It is classified into two major entities: Ulcerative colitis (UC) and Crohn's disease (CD), whose subtype identification is critical for correct disease management, especially for surgery and personalized treatment. Diagnosis is challenging, with approximately 10% of patients being classified as 'indeterminate colitis'. Therefore, novel biomarkers for stratifying patients and improving diagnostics are highly needed.

Here, we have applied a unique RNA sequencing technique, Cap Analysis of Gene Expression (CAGE), on intestinal biopsies from 94 patients with IBD as well as healthy controls. This provided a genome wide atlas of active transcription start sites and enhancers. We show that highly expressed immune cell related transcripts were powerful predictors of the degree of inflammation (controls vs CD & UC), while lowly expressed epithelial cell related transcripts were far more powerful predictors of UC vs. CD. Using the enhancer atlas we show that enhancer transcription can distinguish the inflammatory state of patients and that similarly regulated enhancers and promoters share transcription factor binding sites. Utilizing GWAS data we show that enhancers are more enriched for the heritability of IBD than promoters.

*These authors contributed equally

GENOME WIDE SEQUENCING REVEALS NEW INSIGHTS INTO AGE RELATED HEARING LOSS: CUMULATIVE EFFECTS AND THE ROLE OF SELECTION

Dragana Vuckovic[1], Massimo Mezzavilla[2], Massimiliano Cocca[1], Anna Morgan[1], Martina La Bianca[1], Paolo Gasparini[1,2], Giorgia Girotto[1]

[1]University of Trieste, Department of Medical, Surgical and Health Sciences, Trieste, Italy, [2]Sidra Medical and Research Centre, Experimental Genetics Division, Doha, Qatar

Introduction

Recently, as high-throughput technology became more cost-effective, the possibility of analyzing whole genome sequencing (WGS) and whole exome sequencing (WES) data at population level has been highlighted [Kilpinen & Barrett 2013]. However, it was shown that GWAS studies loose in power due to the allele frequency spectrum targeted by sequencing with realistic sample sizes [King and Nicolae 2014]. Hence, it is necessary to explore new methodologies and approaches for dealing with WGS and WES data. Here, a WGS study of cases and controls has been conducted to unravel the genetic determinants of Age Related Hearing Loss (ARHL), a highly heterogeneous disease.

Materials and Methods

A discovery cohort of 156 subjects and a replication cohort of 56 were enrolled. All individuals were older than 50 years and classified as cases or controls based on their hearing thresholds at high frequencies. Both groups were matched for sex, age and genetic background. Total variation load per gene was compared between cases and controls by means of linear regression to detect outliers, which were then analyzed for Gene Ontology (GO) enrichment with PANTHER web-tool. Replicated genes were investigated for natural selection along the Europe-EastAsian axis, using a PCA-based method. Finally expression studies by RT-PCR were performed in mouse cochlea cDNA for a selection of candidates.

Results

Two groups of outlier genes were detected: 375 more variable in cases and 371 less variable in cases. The largest GO enrichment for both groups of genes (fold>5,p<0.05) was the “sensory perception of sound” biological process, suggesting cumulative genetic effects involved in ARHL. 141 genes were replicated in the independent cohort, among which we identified 21 genes putatively under selection and thus with a possible strong phenotypic effect. After expression studies in the inner-ear, 20 out of 21 genes were positively expressed and two of them (CSMD1 and PTPRD), were previously detected by GWAS studies as involved in hearing function [Girotto et al.2011&2014].

Conclusions

We show that this novel multistep strategy provides major insights into the molecular characterization of complex diseases such as ARHL and could be applied to other phenotypes/diseases, where the paucity of samples makes the GWAS approach not feasible.

Supported by RBSI14AG8P-SIR2014 to GG.

A PHYLOGENETICALLY BASED COMPARATIVE TRANSCRIPTIONAL LANDSCAPE BETWEEN MAIZE AND SORGHUM USING SINGLE-MOLECULE SEQUENCING

Bo Wang[1], Michael Regulski[1], Elizabeth Tseng[2], Sara Goodwin[1], Richard W McCombie[1], Doreen Ware[1,3]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY, [2]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA, [3]USDA ARS NEA Robert W, Holley Center for Agriculture and Health Cornell University, Ithaca, NY

Maize and sorghum are related grain crops with similar plant architectures. Yet, maize is an ancient tetraploid having many retained duplicate genes and highly re-arranged genome compared to sorghum. To better understand these two species on a molecular level, we performed a comparison of the expression profiles of 11 developmentally-matched tissues from each organism by Single-Molecule Long-Read (SMRT) sequencing and deep RNA sequencing. We examined the similarities and differences in the transcriptome for both protein-coding and non-coding transcripts. As a result, we identified tremendous novel isoforms in both maize and sorghum, found that young genes were likely to be generated in reproductive tissues, and usually have fewer isoforms than old genes. We saw similarities and differences of alternative splicing patterns among different tissues and species. We also identified a number of conserved splicing events between maize and sorghum. In addition, we generated a comprehensive high-resolution map of poly(A) sites in both organisms, revealing similarities and differences between mRNA cleavage in these two species. Overall, our results indicate that there is considerable isoform and RNA expression diversity between sorghum and maize, well beyond previous studies, possibly reflecting the architecture differences between these two species. This work was supported by NSF grant #1127112 and NSF grant #1238014.

COMPREHENSIVE QUALITY CONTROL OF MANY SAMPLES USING IOBIO

Alistair Ward[1,2], Chase Miller[1,2], Nielson Phu[1,2], Yi Qiao[1], Gabor Marth[1,2]

[1]University of Utah, Institute of Human Genetics, Salt Lake City, UT,
[2]Frameshift Genomics, Boston, MA

Next-generation sequencing is becoming standard for many research and clinical projects, which now commonly comprise hundreds or thousands of samples. Ensuring the data quality of each sample in the project meets minimum requirements, and that there is uniformity to the data is increasingly important. Filling the need for “at-a-glance” yet comprehensive sequencing data quality assessment in small, medium, or large sequencing projects, we developed a new application, multibam.iobio, built on the open-source iobio platform, that allows users to compare quality metrics of sequencing datasets at the project level; identify outliers; and investigate these samples in greater detail to identify the mode of failure. This application has been developed to ensure this pan-project analysis can be performed by bioinformatician experts familiar with DNA sequencing data; or medical experts trying to understand their own data sets, but who do not have experience with sequencing data; and everyone in between. Current tools used to analyze quality metrics for sequencing alignment files typically operate on the command line, and generate a static text or image file for each sample. As the number of samples included in an analysis grows, it becomes less and less likely that the quality metrics for each sample will be interrogated. Even when the quality metrics are interrogated, it is not always immediately obvious what constitutes a “good” or “bad” sample. In contrast, rather than focusing on quality metrics for a single alignment file, multibam.iobio focuses on visualizing quality metrics for an arbitrary number of samples simultaneously. In this way, multibam.iobio addresses a number of outstanding issues; 1) it is a quick and easy task to check the quality of sequencing data for an entire project, without the need to interrogate each file individually; 2) the data is presented in an intuitive, interactive web application, using clear visualizations to ensure that quality analysis can be performed by experienced bioinformaticians, or analysts with limited computational expertize; 3) problematic samples can be rapidly identified by quick comparison with all other samples in the project; 4) newly sequenced samples can be analyzed in real-time and instantaneously evaluated in the context of large, existing datasets.

We have used multibam.iobio to compare quality metrics of over 1,000 samples in the Heritage 1K project at the University of Utah. Within seconds, several samples immediately stood out as outliers from the “average” sample in the project, in one or more relevant quality metrics. By clicking in these samples, the bam.iobio app is launched to provide a more focused assessment of the individual sample. Ultimately, the discrepancies discovered in this process were genuine errors in the samples that required reprocessing of the data by the sequencing vendor.

MAIZE A COMPLEX GENOME INSIGHTS REVEALED BY SINGLE MOLECULE TECHNOLOGIES

Yinping Jiao[1], Paul Peluso[2], Jinghua Shi[3], Michelle C Stitzer[4], Bo Wang[1], Michael Campbell[1], Joshua C Stein[1], Xuehong Wei[1], Chen-Shan Chin[2], Michael Regulski[1], Sunita Kumari[1], Richard McCombie[1], Gernot G Presting[5], Jeffrey Ross-Ibarra[4], Kelly Dawe[6], Alex Hastie[3], David R Rank[2], Doreen Ware[1,7]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Pacific Biosciences, Menlo Park, CA, [3]BioNano Genomics, San Diego, CA, [4]University of California, Davis, Davis, CA, [5]University of Hawaii, Honolulu, HI, [6]University of Georgia, Athens, GA, [7]USDA-ARS, Cold Spring Harbor, CA

Complete and accurate reference genomes and annotations provide fundamental tools for characterization of genetic and functional variation, as well as insights into genome evolution. Many reference genomes for crop plants have been generated over the past decade, but these assemblies are often fragmented and missing complex repeat regions. Here, we report the assembly and annotation of maize, a genetic and agricultural model organism, using Single Molecule Real-Time (SMRT) sequencing and high-resolution optical map. Our assembly consists of 2,958 contigs with N50 of 1.2 Mb. After gap-filling and error correction using short reads, the total size of maize B73 RefGen_v4 pseudomolecules was 2,106 Mb. Characterization of the repetitive portion, ~ 85% of the genome revealed over 130,000 intact transposable elements (TEs), allowing us to identify TE lineage expansions unique to maize. Gene annotations were updated using 111,000 full-length transcripts, doubling the number of alternative transcripts. Comparative optical mapping of two other inbreds, confirmed the high degree of genetic diversity, with less than 40% of the optical maps aligning, as compared to ~95% typically seen in humans. After its divergence from Sorghum, the maize lineage underwent genome doubling followed by diploidization and gene loss. Previous work showed that gene loss is biased toward one of the parental genomes, but our new assembly and annotation paint a more dramatic picture, revealing that 56% of syntenic sorghum orthologs map uniquely to the dominant maize subgenome. In addition to the gene loss in the context of polyploidy and functional redundancy, we found that despite its polyploidy, maize has lost a larger proportion (14%) of the 22,048 ancestral gene orthologs than any of the other four grass species. Nearly one-third of these losses are specific to maize, with many of the genes involved in biotic and abiotic stresses involved in pathogen defense and programmed cell death.

DISCOVERY OF VARIABLE LYMPHOCYTE RECEPTORS TO CROSS THE HUMAN BLOOD-BRAIN BARRIER

Elizabeth A Waters[1], Brantley R Herrin[2], Jason M LaJoie[1], Eric V Shusta[1]

[1]University of Wisconsin - Madison, Chemical & Biological Engineering, Madison, WI, [2]Emory University, Department of Pathology and Laboratory Medicine, Emory Vaccine Center, Atlanta, GA

Background: Jawless invertebrates, such as lamprey, diverged from jawed invertebrates over 500 million years ago. This evolutionary distance resulted in differences in their immune system, including lamprey with variable lymphocyte receptors (VLRs) instead of IgG antibodies. We hope to raise VLRs from lampreys to bind human proteins that mammalians cannot raise antibodies toward due to immunological self tolerance. We are specifically interested in discovering VLRs that bind novel receptor-mediated transport (RMT) receptors for drug transport across the brain endothelial cells, also known as the blood-brain barrier (BBB).

Methods: VLRs have been collected after immunization of lamprey with human pluripotent stem cell-derived brain endothelial cells (hPSC-derived BECs). A library of the VLRs in a yeast surface display format will be constructed and screened via biopanning on hPSC-derived BECs. Unique VLR clones from the screened library will be made into a soluble format and characterized by their binding and transport properties.

Results: We expect to find unique VLRs that bind and transcytose human brain endothelial cells. We will determine the antigen, which we expect to be a RMT receptor. In vivo studies should demonstrate the VLR distribution into the brain and lack of VLR binding in other organs.

Significance: The discovery of a novel RMT system in the BBB, a major brain drug delivery challenge, would allow for drugs to be delivered to the brain to treat neurodegenerative diseases, such as stroke, Alzheimer's disease, and brain tumors. A brain-selective RMT system would be beneficial to lower peripheral organ uptake of drugs, thus decreases side effects and increase efficacy of the drugs.

A NOVEL SCANNING MODEL UTILIZED TO THE PREDICTION OF BRANCH POINT SEQUENCE IN HUMAN PRE-mRNA SPLICING

Jia Wen

The Chinese University of Hong Kong, School of Life Science, Hong Kong, Hong Kong

Associated with the secondary structure of intron, a novel scanning model is utilized to improve the computational prediction of branch point sequence (BPS) in human pre-mRNA splicing, in which the nucleotide preference at each site in BPS and its binding energy with U2 snRNP are jointly used to quantify the splicing strength of putative BPS. The candidate BPS is exclusively restricted in the BPS search region, which can effectively avoid the influences of other elements in the intron and further strengthen the prediction efficiency. We illustrate the effectiveness of this model on two sets of the experimentally verified human introns, and the improved accuracies demonstrate that our proposed method is better than other current implements on the human BPS prediction. In addition, we propose that the binding energy of BPS-U2 snRNP could contribute the molecular recognition in splicing process, and nucleotide

COVARIATE-AWARE MODELS FOR HIGH-ORDER EPISTASIS ANALYSIS IN YEAST

Jia Wen, Xinghua Shi

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

Epistasis plays a pivot role in human complex traits and common diseases. Previous studies have shown that accounting for epistasis leads to better predictions of complex traits. Current genomic datasets are typically high dimensional, where the number of features is usually much larger than sample size. The high dimensionality of data makes it challenging to design efficient computational methods for high order epistasis analysis. In this study, we leverage a scalable model, i.e. empirical Bayesian Elastic Net (EBEN)1, to detect both pairwise and high-order epistasis in an iterative process with consideration of covariates (e.g. population structure, age, gender). Specifically, we incorporate three complementary strategies for incorporating covariates in the EBEN model, namely an orthogonal strategy, a least square estimation, and a likelihood ratio test. Simulations show that incorporating covariates greatly improve the performance of EBEN model for epistasis analysis. Applying to an experimental yeast cross dataset2, we identify a set of pairwise and high-order epistatic interactions associated with 20 quantitative traits in yeast. We then comprehensively quantify the contribution of main effect, pairwise epistatic effect and high-order epistatic effect on each trait variation. We conduct pathway analysis to help us better understand the molecular mechanism underlining these traits3. We further visualize these main and epistatic effects in a network view for traits (e.g. Indolacetic Acid level) relevant to yeast fitness.

Reference:
1. Anhui Huang, Shizhong Xu, and Xiaodong Cai. Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity, 114(1):107-115 (2015).
2. Joshua S. Bloom, Iulia Kotenko, Meru J. Sadhu, Sebastian Treusch, Frank W. Albert and Leonid Kruglyak. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat. Commun. 6, 8712 (2015).
3. Simon K G Forsberg, Joshua S. Bloom, Meru J. Sadhu, Leonid Kruglyak, and Örjan Carlborg. Accounting for genetic interactions is necessary for accurate prediction of extreme phenotypic values of quantitative traits in yeast. bioRxiv, 059485 (2016).

FINE-MAPPING IDENTIFIES RA AND T1D FUNCTIONAL CAUSAL VARIANTS IN *DNASE1L3, MEG3, TNFAIP3* AND *CD28/CTLA4* LOCI

Harm-Jan Westra[1,2], Marta Martinez Bonet[3,4], Suna Onengut[5,6], Anette Lee[7], Steve Eyre[8], John Todd[9], Peter A Nigrovic[3,4], Peter K Gregersen[7], Stephen S Rich[5,6], Soumya Raychaudhuri[1,2,3,4]

[1]Brigham and Women's Hospital, Harvard Medical School, Division of Genetics and Rheumatology, Boston, MA, [2]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA, [3]Brigham and Women's Hospital, Division of Rheumatology, Immunology, and Allergy, Boston, MA, [4]Boston Children's Hospital, Division of Immunology, Boston, MA, [5]University of Virginia, Center for Public Health Genomics, Charlottesville, VA, [6]University of Virginia, Department of Public Health Sciences, Charlottesville, VA, [7]Northwell Health, The Feinstein Institute for Medical Research, Manhasset, NY, [8]University of Manchester, Arthritis Research UK Centre for Genetics and Genomics, Musculoskeletal Research Centre, Institute for Inflammation and, Manchester, United Kingdom, [9]University of Cambridge, Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for MedicCambridge, United Kingdom

We fine-mapped 76 type 1 diabetes (T1D) and rheumatoid arthritis (RA) non-MHC loci. After imputation, using a sequencing experiment of 568 individuals targeting 799 1kb regions within these loci, we observed 96% coverage of common variants. We performed Bayesian fine-mapping in autoimmune disease loci in an RA (11,475 cases, 15,870 controls), T1D (9,334 cases and 11,111 controls) and combined dataset. We consequently narrowed down the potential number of causal variants to 5 or less in 7 loci for RA and 11 loci for T1D. Detailed analysis, accounting for multiple signals, identified likely causal coding variants in four loci (*DNASE1L3, PTPN22, IFIH1*, and *TYK2*), likely causal indel variants in *MEG3, TNFAIP3*, and *ANKRD55* and causal SNP variants in *ATXN2/SH2B3*, *REL/PUS10* and *CD28/CTLA4*). Functional analysis identified allele specific binding and luciferase activity at three of these loci: rs117701653 in *CD28/CTLA4*, rs34552516 in *MEG3*, and rs35926684 in *TNFAIP3*. This study demonstrates the potential for dense genotyping and imputation to pinpoint individual causal alleles, including those with allele specific enhancer activity.

A NONSENSE MUTATION IN *CEP55* DEFINES A NEW LOCUS FOR A MECKEL-LIKE SYNDROME, AN AUTOSOMAL RECESSIVE LETHAL FETAL CILIOPATHY

Maria Wilbe[1], Marie-Louise Bondeson[1], Katharina Ericson[1,2], Sanna Gudmundsson[1], Adam Ameur[1], Fredrik Pontén[1], Jan Wesström[3,4], Carina Frykholm[1]

[1]Uppsala University, Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala, Sweden, [2]Uppsala University Hospital, Department of Pathology and Cytology, Uppsala, Sweden, [3]Center for Clinical Research, Department of Obstetrics and Gynecology, Dalarna, Sweden, [4]Uppsala University, Department of Women´s and Children´s Health, Uppsala, Sweden

Mutations in genes involved in the cilium–centrosome complex are called ciliopathies. Meckel-Gruber syndrome (MKS) is a ciliopathic lethal autosomal recessive syndrome characterized by genetically and clinically heterogeneous signs, such as renal cystic dysplasia, occipital encephalocele and polydactyly. Several genes have previously been associated in MKS and MKS-like phenotypes, but there are still genes remaining to be discovered. We have used whole exome sequencing (WES) to uncover the genetics of a suspected autosomal recessive Meckel syndrome phenotype in a family with two affected fetuses. RNA studies and histopathological analysis was performed for further investigations.
WES lead to identification of a homozygous nonsense mutation c.256C>T (p.Arg86*) in *CEP55* (centrosomal protein of 55 kDa) in the affected fetus. The variant has previously been identified in carriers in low frequencies, and segregated in the family.
Conclusively, we describe a family with recurrent fetal loss, with a homozygous mutation (c.256C>T) in *CEP55* causing an autosomal recessive Meckel-like syndrome phenotype in affected fetuses. CEP55 is an important centrosomal protein required for the mid-body formation at cytokinesis. Our results expand the list of centrosomal proteins implicated in human ciliopathies and provide evidence for an essential role of CEP55 during embryogenesis and development of disease.

INTEGRATION OF GWAS AND EPIGENETIC FEATURES FOR TOTAL IRON BINDING CAPACITY IDENTIFIES NEW LOCI AND NEW HYPOTHESES REGARDING CLINICAL IMPLICATIONS

Cristen Willer[1], Jonas Nielsen[1], Lars Fritsche[2], Wei Zhou[3], Maoxuan Lin[1], Maiken Elvestad[2], Anne Heidi Skogholt[2], Hyun Min Kang[4], Michael Boehnke[4], Ketil Thorstensen[5],Goncalo Abecasis[4], Kristian Hveem[2]

[1]University of Michigan, Department of Internal Medicine, Ann Arbor, MI, [2]Norwegian University of Science and Technology, HUNT Research Centre, Department of Public Health and General Practice, Trondheim, Norway, [3]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [4]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [5]St. Olav Hospital, Dept. of Medical Biochemistry, Trondheim, Norway

Iron is an essential component of hemoglobin which distributes oxygen to tissues, but iron levels are tightly controlled because excess iron is toxic. Iron-deficiency causes anemia which can disrupt cognitive development and the immune system. Excess iron is associated with a variety of health problems including liver damage, neurodegenerative diseases such as Alzheimer's disease, type 2 diabetes, and cardiovascular disease and related complications.

We recently performed a genome-wide association scan (GWAS) in 70,000 Norwegian individuals from the HUNT study. Of the many phenotypes examined, some of the potentially most interesting results were observed for total iron binding capacity, highly correlated with ferritin levels. Twenty-two loci reached genome-wide significance, including four novel loci. Two loci reached association P-values $< 10^{-800}$: the missense variant p.C282Y in *HFE* that is known to cause recessive hemochromatosis (rs1800562; MAF 9.2%) and a common non-coding variant intronic to the transferrin (*TF*) gene which binds iron for transport from the intestine to proliferating tissues. After examining 709 additional phenotypes, we demonstrated that six of the 22 index variants are also associated with serum triglyceride levels ($P < 10^{-4}$), two variants with liver enzymes alanine aminotransferase and alkaline phosphatase indicating liver damage, and several variants with indicators of metabolic disease including serum glucose levels. Tests for enrichment of associated variants in a variety of regulatory features identified significant enrichment in histone modifications in lymphoblastoid cell lines, DNaseI hypersensitivity peaks in blood tissues, fetal thymus and fetal intestine. eQTLs and missense variants allowed us to identify candidate functional genes at approximately half of the loci. This quantitative phenotype provides an exemplary model for testing a variety of approaches for functional characterization of GWAS loci, identifying biological mechanisms and determining clinical implications using phenome-wide and Mendelian randomization approaches.

INFERRING PARENTAL GENOMES USING DATA FROM A SET OF SIBLINGS

Sayatani Basu-Roy[1], John Blangero[2], Amy L Williams[1]

[1]Cornell University, Biological Statistics and Computational Biology, Ithaca, NY, [2]University of Texas Rio Grande Valley, South Texas Diabetes and Obesity Institute, Brownsville, TX

Children inherit two chromosome copies, one from each parent, with both formed via recombination. While each child inherits only half of the genomes of each parent, independent segregation and recombination are randomized such that *n* siblings will inherit on average a proportion of $1-1/2^n$ of both parents' genomes. Analyses of sibling data can recover (partial) parental haplotypes, yet which parent—mother or father—each pair of haplotypes belongs to is ambiguous. Moreover, reconstruction of each chromosome is independent, leading to $2^{22} > 4$ million possible parental haplotype configurations in humans.

Male- and female-generated crossovers differ in location and frequency, a fact we exploit to infer which parent carried each haplotype. Specifically, we use a hidden Markov model to analyze sibling data, inferring the parental haplotypes (without determining which parent each belongs to) together with the positions of crossover events. Crucially, since crossovers only occur between haplotypes carried by a single parent, in most cases, linkage enables the partitioning of the haplotypes and all their associated crossovers into two sets corresponding to the two parents. A subset of haplotype transmission patterns are ambiguous and linkage is unknown across these patterns, but analysis of the sets of transmitted crossovers detected on both sides of these patterns enables inference. Next, from the two sets of regional or chromosome-wide crossovers transmissions, we compute the joint likelihood of the crossovers being transmitted by a male or female parent, determining the probability of each parental assignment. We adopt a Poisson model of crossover with the probabilities of the number and location of the events based on Morgan positions in sex-specific genetic maps.

We evaluated this framework using 69 families from the San Antonio Family Studies that include data for three or more siblings and genotypes for 918,917 SNPs. To learn about the feasibility of the approach, we currently phase the siblings with parental data and infer the probabilities of the parent configurations using the identified two sets of transmitted crossovers. This inference yielded near perfect accuracy, with the correct parental assignment for 1,515 out of 1,518 chromosomes (three mistakes). Analyses using data from ≥6 siblings only will achieve nearly this fidelity, with smaller families only inhibited near the relatively rare ambiguous patterns. These results glimpse a future in which association studies use data from deceased individuals for greater power and use full biomedical life histories.

NANOCONFINEMENT BASED APPROACHES TO CHROMOSOME SYNTHESIS

Eamon M Winden[1], Samuel Krerowicz[2], David C Schwartz[1]

[1]University of Wisconsin- Madison, Genetics, Madison, WI, [2]University of Wisconsin- Madison, Chemistry, Madison, WI

Very large DNA molecules are incredibly informative substrates that present unique challenges which must be met in order to utilize their advantages. Past advances here have included Pulsed Field Gel Electrophoresis, Yeast Artificial Chromosomes (YACs) and Optical Mapping, which have leveraged many intrinsic physical and genomic advantages. Given this context, we aim to synthesize DNA molecules on the order of mammalian chromosomes. Our group has previously pioneered systems that manipulate and confine DNA molecules using very low ionic strength buffers and engineered charge interactions between physical walls of a device and the molecule in ways that create an “electrostatic bottle.” Here, a molecule is confined by not just physical boundaries, but by its electrostatic interactions. By advancing and harnessing this effect, we are enabling use of nanoscale control to direct serial hybridizations within a usable synthesis cycle supporting the fabrication of entire chromosomes. This technology will enable cell-free whole genome synthesis.

USING GENOTYPED RELATIVES OF UNGENOTYPED TYPE 2 DIABETES CASES AS PROXY-CASES IN A COHORT BASED GENOME WIDE ASSOCIATION STUDY

Brooke N Wolford[1], Seunggeun Lee[2], Wei Zhou[1], Jonas B Nielsen[3], Lars G Fritsche[4], Maoxuan Lin[3], Hyun Min Kang[2], Maiken Gabrieldsen[5], Oddgeir Holmen[5], Kristian Hveem[5], Michael Boehnke[2,1], Cristen J Willer[1,3]

[1]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [3]University of Michigan, Department of Internal Medicine, Ann Arbor, MI, [4]Norwegian University of Science and Technology, K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and General Practice, Levanger, Norway, [5]Norwegian University of Science and Technology, HUNT Research Centre, Department of Public Health and General Practice, Levanger, Norway

Type 2 diabetes (T2D) is a complex disease with genetic and environmental factors contributing to its onset. To date, over 100 independent SNPs have been associated with T2D through genome wide association studies (GWAS). A recently published method by Liu et. al. introduces the concept of GWAS by proxy (GWAX): performing case-control genetic association studies using unaffected first degree relatives of cases (proxy-cases) in the absence or near absence of true cases. Here, we have extended this method to model genetic liability in cases, proxy-cases, and controls in a T2D GWAS in the Norwegian Nord-Trøndelag Health Study (HUNT). We partitioned our genotyped and imputed sample of 69,716 European individuals into 5,302 T2D cases, 9,880 proxy-cases, and 45,086 controls using health questionnaires, electronic health records, and genetic relatedness in the population-based sample. Proxy-cases were determined by self-report of having an affected first degree relative. We tested for association between T2D and 25 million genotyped and imputed genetic variants. We compared the standard GWAS model to GWAX and several methods of accounting for proxy-cases. Preliminary analysis at 108 known T2D GWAS loci using standard GWAS identified two SNPs reaching genome-wide significance (GWS, p<5x10-8) in our sample. The standard GWAS model compares cases to controls and includes proxy-cases, typically undetected, as controls. By simply removing proxy-cases from controls, we identified an additional four SNPs at genome-wide significance. Finally, by appropriately modeling proxy-cases with cases and controls we identified a total of 11 SNPs at GWS. By modeling genetic liability in proxy-cases we increase power to detect signals at known loci without the cost of collecting more samples. For example, modeling genetic liability in proxy-cases at rs9936385, which is in high LD (r^2=0.99) with known causal variant rs1421085 of the T2D risk locus *FTO*, results in a more significant p-value (p=6.09x10^{-14}) over traditional GWAS (p=4.15x10^{-11}). This method can also be applied to the evaluation of genetic risk for other traits such as myocardial infarction and asthma. With the increasing availability of biobank data, this work demonstrates the advantage of statistically modeling proxy-cases in cohort based GWAS.

GENOME-WIDE CHROMOSOMAL CONFORMATION ELUCIDATES REGULATORY RELATIONSHIPS IN HUMAN BRAIN DEVELOPMENT AND DISEASE

Hyejung Won[1], Luis de la Torre-Ubieta[1], Jason L Stein[2], Neelroop N Parikshak[1], Jason Ernst[3], Daniel H Geschwind[1]

[1]David Geffen School of Medicine, University of California Los Angeles, Neurogenetics Program, Department of Neurology, Los Angeles, CA, [2]University of North Carolina, Chapel Hill, Department of Genetics & Neuroscience Center, Chapel Hill, NC, [3]University of California Los Angeles, Department of Computer Science, Los Angeles, CA

The demonstration that chromatin exhibits a complex 3 dimensional organization, whereby short and long distance physical interactions correspond to complex gene regulatory processes has opened a new window on understanding the functional organization of the human genome. Recently, chromatin remodeling has also been causally implicated in several neurodevelopmental disorders, including autism and schizophrenia. However, it remains unclear whether knowledge of chromosome organization in a tissue specific manner might inform our understanding of gene regulation in brain development or disease. Here we determined the genome-wide landscape of chromosome conformation during early human cortical development by performing Hi-C analysis in the mitotically active and post mitotic laminae of human fetal brain. We integrate Hi-C data with transcriptomic and epigenomic data and utilize chromosome contact information to delineate physical gene-gene regulatory interactions for non-coding regulatory elements. We show how these data permit large-scale functional annotation of non-coding variants identified in schizophrenia GWAS and of human specific enhancers. These data provide a rubric that illustrates the power of tissue-specific annotation of non-coding regulatory elements, as well as novel insights into the pathogenic mechanisms of neuropsychiatric disorders and the evolution of higher cognition.

RAMBOUILLET SHEEP GENOME AND GENOMIC RESOURCES

Yue Liu[1], R. A Harris[1,2], Xiang Qin[1,2], Stephen Richards[1,2], Jeffrey Rogers[1,2], Yi Han[1], Vanesa Vee[1], Min Wang[1], Qingchang Meng[1], Mike P Heaton[3], Timothy P Smith[3], Brian P Dalrymple[4], Stephen N White[5], Brenda Murdoch[6], James Kijas[7], Noelle E Cockett[8], Donna M Muzny[1,2], Kim C Worley[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, [3]USDA Agricultural Research Service, U.S. Meat Animal Research Center, Clay Center, NE, [4]University of Western Australia, Institute of Agriculture, Perth, Australia, [5]Washington State University, Veternary Microbiology and Pathology, Pullman, WA, [6]University of Idaho, Animal and Veternary Science, Moscow, ID, [7]CSIRO, St. Lucia, Australia, [8]Utah State University, President's Office, Logan, UT

High quality reference genomes are fundamental resources for biology. We report here a new sheep reference genome for the Rambouillet breed generated using the latest methods for producing and *de novo* assembling long reads and scaffolding to chromosomes. High quality annotation will be enabled by assays of the FAANG quality samples that have been collected from the reference animal.

A total of 200 Gb of sequence was generated from a single ewe, Benz2616, using the Pacific Biosciences (PacBio) technology. The data have a 12.6 kb N50 and 8.9 kb mean subread length. We used both Falcon and Celera Assembler to error correct and assemble the error corrected reads and selected the more complete product of the Celera Assembler for further processing. The preliminary assembly has a contig N50 of 2.2 Mb, a total length of 2.85 Gb, 365 contigs contain half of the genome sequence, and the longest contig is 16.3 Mb. The majority (89%) of 338,551 EST sequences align to the genome, with most (90%) having nearly complete alignments, aligning over more than 95% of their length. This version of the genome is more complete and has more contiguous alignment to these expressed sequences than the Texel breed Oar4.0 reference. Base quality of this assembly is high, with error rates less than 1% following assembly polishing using Arrow.

Hi-C proximity ligation data from the same individual is being used for scaffolding the preliminary contigs. An initial scaffolding attempt incorporated 97.4% of the assembly into 32 large scaffolds and 2,900 smaller scaffolds (<100kb). This initial scaffold version captures more complete, single copy BUSCO genes than Oar-v4.0 or the input initial contigs. Further scaffold refinement is ongoing. Additional PBJelly gap filling and Pilon base error correction are planned prior to release.

Over 100 tissues from the reference animal have been collected for additional FAANG assays. This large effort involved over 35 people in the sample collection, advance planning and coordination. In addition to the PacBio genome sequence and Hi-C data for the genome assembly, PacBio IsoSeq, miRNAseq, ATAC-Seq and other assays are planned.

BINDING SITES WITHIN LONG NON-CODING RNAS DISCRIMINATE BETWEEN RNA- AND TRANSCRIPTION MEDIATED MECHANISM

Tomasz Wrzesinski, Wilfried Haerty

Earlham Institute, Norwich Research Park, Colney Lane, Haerty Group, Norwich, United Kingdom

Tens of thousands of long non-coding RNAs (lncRNAs) have now been annotated in the human genome. They represent a highly heterogeneous class of transcribed elements with respect to their genomic position, molecular mechanisms, cellular localization and potential function. Only few lncRNAs has been experimentally characterized showing functions in dosage compensation, genomic imprinting and gene expression regulation. Despite active research, little is still known for the vast majority of lncRNAs including the proportion that are biologically functional. Previous reports highlighted a dichotomy between lncRNAs, identifying loci whose function was solely conveyed by the act of transcription the transcript being functionally inert (*Bendr*) and loci with a RNA based function (*Xist*). We previously identified significant signals of purifying selection for splicing regulatory elements within a subset of lncRNAs supporting a RNA mediated mechanism.

Here we further investigate transcript vs transcription-mediated role of lncRNAs in human aiming to identify the proportion of loci belonging to either or both classes. We tested the selective constraints acting on binding sites (AGO, TFs, RBP) within lncRNA exons, introns, as well as 1kb upstream and downstream sequences. LncRNAs with binding sites for RNA binding proteins are more highly expressed than those with either TF binding sites only or without any binding sites. We report increased conservation of the binding sites relative to matched sequences. Interestingly, lncRNAs with AGO sites are depleted within annotated enhancers whereas the opposite was found for lncRNAs with TFBSs. The joint analysis of purifying selection acting on functional elements within lncRNAs and of the loci genomic context help in distinguishing loci with different mechanisms of action.

IDENTIFICATION OF DRUG METABOLIZING ENZYME ASSOCIATED TO AFRICAN ANCESTRY

Yilin Xu, Tanima De, Cristina Alarcon, Minoli Perera

Northwestern University, Pharmacology, Chicago, IL

Differences among ethnic groups in drug metabolizing enzymes (DME) and transporters may account for differences in pharmacokinetics, resulting in variability in response to drug therapy (Yasuda, Zhang, & Huang, 2008). Expression of DME is influenced by a unique combination of factors including genetic polymorphisms, medications taken, age, sex and others. Single Nucleotide polymorphisms (SNPs), which may have population frequency differences, play a major role for the function of many DMEs, such as CYPs 2D6, 2C19, 2C9, 2B6, 3A5 and 2A6 (Zanger & Schwab, 2013). The aim of this study was to identify important DME and transporters whose expression level is associated with African ancestry.

In this study, we generated RNA-Seq and genotype data from hepatocytes isolated from 44 different African American (AA) liver tissues. Association between gene expression and African ancestry percentage was analyzed using DESeq2 Bioconductor package after correcting for sample heterogeneity and technical variation. Among all genes, 11 genes were positively correlated with African ancestry: GUCY1B3, NOL4, CAV2, ELMOD1, DNAH6, HLX, IL18, MUM1L1, MEGF6, HSPA4L, RNF135 while 10 genes were negatively associated: LTF, ST6GALNAC1, TLL2, IL13RA2, HILPDA, S100A8, HIC1, WASH1, AF064858.6, KRTAP5-9. Furthermore, we examined relationship between African ancestry and DME/transporters genes. Among the 63 PharmGKB very important pharmacogenomics (VIP) genes, we identified 6 genes (ADH1A, ADH1B, ADRB1, ALK, ALOXT and CYP2C19) that were significantly associated with African ancestry.

In addition, applying similar methods to the GTEx dataset of AA subjects, we identified 259 genes globally associated with African ancestry and 6 Pharm GSK genes (ALDH1A1, CYP2A6, CYP2B6, CYP2C19, CYP2C8 and CYP3A4) associated with African ancestry. Due to differences in sample size, library preparation and sequencing platform between the two cohort, the identified gene sets were distinct. However, we replicated the correlation to ancestry in LTF, S100A8 and CYP2C19's expression, which was negatively correlated with African ancestry. It has been also reported that frequency of many clinically relevant variants of CYP2C19 differ between Caucasian and AA populations. Our study suggests that African ancestry may contribute to population differences in drug response, specifically for drug that use the CYP2C19 enzyme pathway.

Yasuda, S. U., Zhang, L., & Huang, S. M. (2008). The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. Clin Pharmacol Ther, 84(3), 417-423. doi:10.1038/clpt.2008.141
Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther, 138(1), 103-141. doi:10.1016/j.pharmthera.2012.12.007

GLOBAL EFFECTS OF HUMAN GENETIC DISORDERS AND HUMAN ADAPTION ON MOLECULAR NETWORKS

Anupama Yadav[1,2], Tong Hao[1,2], David E Hill[1,2], Marc Vidal[1,2]

[1]Dana-Farber Cancer Institute, Center for Cancer Systems Biology and Department of Cancer Biology, Boston, MA, [2]Department of Genetics, Harvard Medical School, Boston, MA

Large numbers of genetic variants have been identified for many Mendelian and complex diseases. However, this deluge of genetic information is not being followed by a comparatively paced identification of the molecular characterization of these genetic variants, thereby hampering the ability to utilize this information for scientific or clinical purposes. Here, I present a systematic and unbiased orthogonal approach to study the molecular effects of these genetic variants. Genes and their products do not function in isolation and instead interact with each other in the context of molecular networks. Disease-causing variants can perturb these networks causing detrimental effects. Our lab (1) showed that although disease-causing variants perturb protein-protein interactions (PPIs), half of the variants perturbed only a subset of PPIs and left the others intact. The PPI profiles of proteins from different variants of the same gene followed patterns of their disease specificity, indicating that these PPIs could provide biological insights into disease mechanisms and identify novel disease-specific targets.

The same disease-causing variant can have differential effects like severity, age of onset, response to drugs etc. in different populations. This population-specificity is likely due to heterogeneity of the underlying cellular networks mediating the effect of the variant on the phenotype. Cellular networks are shaped over the course of adaptation, and hence differ across human populations that have encountered diverse selection pressures due to migrations and changes in diet and climate. While genetic diversity across populations has been estimated in detail identifying many alleles under selection in different populations, their effects on the cellular networks and phenotypes remain largely unknown. We propose mapping PPIs of alleles under selection in different populations with reference and disease-causing variants to understand their molecular effects. Five hundred genes under strong positive selection show an average of 10 PPI in our dataset including 170 disease-causing genes. Moreover, several genes under selection without characterized functions show large number of PPIs in our data. Our study will elucidate how the molecular interactome rewires in response to adaptation and how this rewiring effects the molecular dynamics of disease phenotypes.

1. Sahni N, et al. 2015. Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161, 647-660.

METAGENOMIC BINNING THROUGH LOW DENSITY HASHING

Yunan Luo[1,4], Y. William Yu[2,3], Jianyang Zeng[1], Bonnie Berger[2,3], Jian Peng[4]

[1]Tsinghua University, Institute for Interdisciplinary Information Sciences, Beijing, China, [2]MIT, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [3]MIT, Department of Mathematics, Cambridge, MA, [4]University of Illinois at Urbana-Champaign, Department of Computer Science, Urbana, IL

Bacterial microbiomes of incredible complexity are found throughout the world, from exotic marine locations to the soil in our yards to within our very guts. With recent advances in Next-Generation Sequencing (NGS) technologies, we have vastly greater quantities of microbial genome data, but the nature of environmental samples is such that DNA from different species are mixed together. Here, we present Opal for metagenomic binning, the task of identifying the origin species of DNA sequencing reads. Opal brings low-density hashing to metagenomic binning, enabling quick and accurate binning; the family of hash functions we use, based on Gallager Low Density Parity Check (LDPC) codes, ensures even coverage of all locations using low density hashing. Our tool has comparable speed as other compositional binners and better accuracy, and can be used to as a first pass coarse filter to give an order of magnitude speedup to alignment based methods. Because we better encode long-range dependencies using low density hashing, we achieve much better accuracy in the high high mutation/sequencing error rate regime, as well as better capturing of higher taxonomic level identification.

DETECTING THE SOURCE OF DNA CONTAMINATION IN GENOTYPING ARRAYS

Gregory JM Zajac[1,2], Lars G Fritsche[3], Susan L Dagenais[4], Robert H Lyons[4], Chad M Brummett[5], Gonçalo Abecasis[1,2]

[1]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [2]University of Michigan, Center for Statistical Genetics, Ann Arbor, MI, [3]Norwegian University of Science and Technology, Department of Public Health and Nursing, Trondheim, Norway, [4]University of Michigan, Department of Biological Chemistry and DNA Sequencing Core, Ann Arbor, MI, [5]University of Michigan, Department of Anesthesiology, Division of Pain Medicine, Ann Arbor, MI

Genotyping arrays are a cost-effective tool to assay thousands of individuals at known genetic markers for many types of studies including genome-wide association studies which have contributed to the understanding of the genetics of hundreds of complex traits. Contamination, the mixture of DNA samples from multiple individuals prior to or during genotyping, increases the probability that the true genotype is called incorrectly or as missing, reducing the power and accuracy of follow-up genetic analyses. Current methods for contamination detection do not try to identify the contaminating sample. We propose a two-step process. First, we use a particular sample's alternative allele intensity and the called genotypes for all potential contamination sources to identify a likely set of DNA donors. Then, we compute a final estimate of the amount of contamination from all donors and unidentified sources. Experimental results from intentionally mixed HapMap samples show that our method estimated contamination more accurately than existing methods. Applying our method to an ongoing study of more than 20,000 individuals successfully identified the source of contamination and pointed to likely processing steps where it occurred. When compared to existing methods, this approach has several advantages, including more accurate estimation of the proportion of contaminating DNA, no need for external databases of allele frequencies, and less sensitivity to the ancestral origin of the contaminating sample. Detecting the source of contamination provides useful information to guide improvements in sample processing and preparation protocols.

ALTERNATIVE POLYADENYLATION DRIVES DYNAMIC GENE EXPRESSION THROUGHOUT EMBRYOGENESIS

Harel Zalts[1], Natalia Mostov[1], Eitan Winter[1], Tamar Hashimshony[1], Itai Yanai[2]

[1]Technion - Israeli Institute of Technology, Department of Biology, Haifa, Israel, [2]New York University, School of Medicine, Institute for Computational Medicine, New York City, NY

Transcriptomic studies typically examine expression at the gene level, though it has been known for some time that genes produce multiple isoform types, following splicing and alternative polyadenylation. Specifically, the expression dynamics and biological significance of 3'UTR variants during embryonic development has not been characterized. Here, we describe a method for the analysis of single-cell RNA-Seq data which allows for the determination of specific alternative polyadenylation (APA) sites, as well as the expression level of the different variants. We use this method to study 3'UTR variant usage throughout the embryonic development of the nematode *C. elegans*. Surprisingly, we find that genes with constitutive expression throughout development are enriched for dynamic isoform usage. These genes tend to participate in cellular as opposed to developmental functions. The same phenomenon is also evident in a closely related nematode, *C. japonica,* providing evidence that the manner by which these cellular processes are regulated during embryogenesis is under selection. Finally, we report evidence for stronger miRNA regulation upon genes with dynamic isoform usage. Collectively, our results reveal the importance of alternative polyadenylation throughout embryogenesis. Our methodology provides a context for the incorporation of miRNAs and mRNA regulation in the modeling of biological pathways.

CHORUS: A FEDERATED DATABASE ENABLING THE SHARING OF GENOMIC DATA ASSOCIATED WITH SUBJECTS FROM VARIED ETHNIC GROUPS FROM DIFFERENT SINGAPORE INSTITUTIONS

Samantha L Zarate[1], Jack Ow[2], Yih-Chii Hwang[1], Yifei Men[1], Preston Lim[1], Pauline Ng[2]

[1]DNAnexus, Science, Mountain View, CA, [2]Genome Institute of Singapore, POLARIS, Singapore

The multi-ethnic population of Singapore makes the clinical application of genomics challenging. Variants that seem causal for a condition due to their rarity in the global population may, in fact, occur at high frequencies among those from the same ethnic group as the patient. The diversity of Singapore's population places great import on aggregating information about the genomic backgrounds of many individuals without compromising their privacy.

Community of Health and Omics Resources Unique to Singapore (CHORUS) is a federated database managed by the Genome Institute of Singapore. Comprised of de-identified allele frequencies from control subjects in case-control studies, CHORUS allows more than five independent institutions in Singapore to share genomic data with DNAnexus users. In collaboration with DNAnexus, a cloud-based genomic analysis platform serving as a neutral third party, users can query the CHORUS database to find variants present in Singaporean individuals by ethnic group and disease phenotypes. However, CHORUS prevents its users from downloading an entire data set corresponding to an individual in order to maintain security. This data federation allows each contributing entity to both manage its own separate database and query others' databases, thus upholding privacy while keeping in the spirit of scientific collaboration.

Currently, control subjects from seven different case-control studies provide data for CHORUS. However, making the case subjects' data available may be useful in the future to allow clinical researchers to compare their variants of interest with those in CHORUS.

CHORUS is actively seeking collaboration with hospitals and research centers outside of Singapore in order to broaden and diversify the database. This work can serve as a model for how to build knowledge about ethnic heterogeneity within clinical genomics and how to facilitate the widespread collection of genomic information while preserving data confidentiality. More information about CHORUS can be found online at https://www.a-star.edu.sg/polaris/RESOURCES/CHORUS.aspx.

ENSEMBLE BASED PREDICTION OF CAUSAL REGULATORY VARIANTS.

Haoyang Zeng, Matthew D Edwards, Yuchun Guo, David K Gifford

Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA

We present a novel ensemble-based computational framework, EnsembleExpr, that identifies eQTLs with higher accuracy than published methods. It achieved the best performance in the Fourth Critical Assessment of Genome Interpretation (CAGI4) “eQTL-causal SNPs” challenge for identifying eQTLs and prioritizing their gene expression effects. Expression quantitative trait loci (eQTLs) are genome sequence variants that result in gene expression changes and thus are prime suspects in the search for contributions to the causality of complex traits. When EnsembleExpr is trained on data from massively parallel reporter assays it accurately predicts reporter expression levels from unseen regulatory sequences and identifies sequence variants that exhibit significant changes in reporter expression. EnembleExpr uses features derived from the output of deep learning models trained on multiple high-throughput experimental data sets as well motif scores from a new representation of transcription factor binding affinities called the K-Mer set Model. The K-mer Set Model is able to predict transcription factor binding more accurately than motifs from MEME, Homer, and TTFM. We envision EnsembleExpr to be a resource to help interpret non-coding regulatory variants and prioritize disease-associated mutations for downstream validation.

CIS-REGULATORY ANNOTATION OF GENOMES IN ENSEMBL

Daniel R Zerbino[1], Thomas Juettemann[1], Ilias Lavidas[1], Michael Nuhn[1], Steven Wilder[1], Avik Datta[1], Ernesto Löwy-Gallego[1], Kieron Taylor[1], William Jones[2], Myrto Kostadima[1], Laura Clarke[1], Magali Ruffier[1], Paul R Flicek[1]

[1]EMBL, EMBL-EBI, Hinxton, United Kingdom, [2]Wellcome Trust Sanger Institute, 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 that brings together genome sequences, genes, non-coding RNAs, known variants, etc.

In particular, Ensembl’s Regulatory Build synthesises public epigenomic datasets produced by large-scale projects such as ENCODE, Roadmap Epigenomics or BLUEPRINT. We process them through a unified pipeline and make them available through a single interface. We also define functionally active regions across 68 human cell types (on both the GRCh37 and GRCh38 assemblies) and 6 mouse cell types, assigning them a function wherever possible. We are currently expanding our annotation to more cell types. In particular, we maintain the International Human Epigenome Consortium’s (IHEC) Epigenome Reference Registry (EpiRR) where teams from around the world record the metadata describing their epigenomic datasets before they are incorporated into IHEC’s Data Portal. We hope to expand soon to more species in collaboration with the Functional Annotation of Animal Genomes (FAANG) project.

Regulatory elements are chiefly of interest because of their action on genes; we are therefore simultaneously developing a database of cis-regulatory interactions attaching them to their target genes. Currently, two main approaches are being used to detect these interactions: genetics (e.g. eQTLs) and chromatin conformation (e.g. Hi-C). We have developed new technologies to store and display these datasets, using in particular HDF5 indexing for fast retrieval. This technology allows us to store and display all of the GTEx summary eQTL data, as opposed to only significant correlations, thus avoiding interval censoring. Over the next year, we will be integrating more eQTL and Promoter Capture Hi-C datasets, organized by tissue.

Finally, we are also developing high performance tools for high-throughput analysis. We foresee that in the future most genomic data will be confined by local legislation and it will be necessary to deploy applications across remote data centers. Using our RESTful APIs, it is already possible to retrieve our data simply and efficiently for any gene, variant or region, along with all other Ensembl annotations such as LD calculations from the 1000 Genomes dataset Phase 3, conservation scores, etc. It is thus possible to quickly develop advanced functional analysis pipelines without having to download or process massive data files, as we demonstrate with a post-GWAS analysis pipeline called POSTGAP.

THE USAGE OF LOCAL ANCESTRY TO INFORM eQTL MAPPING IN AFRICAN AMERICANS

Yizhen Zhong[1], Eric Gamazon[2], Minoli Perera[1]

[1]Northwestern University, Department of Pharmacology, Chicago, IL,
[2]Vanderbilt University, Division of Genetic Medicine, Nashville, TN

Expression quantitative trait loci (eQTL) are genetic variants that are significantly associated with gene expression. Since the majority of trait-associated variants identified by genome-wide association study (GWAS) reside in non-coding regions, eQTLs are important resource to explain the potential mechanism of GWAS variants because of their regulatory effects. Unfortunately, recent eQTLs mappings are mostly performed in European populations and have excluded populations of African ancestry. The eQTL mapping in admixed population (i.e. African Americans) is not only urgently needed to decrease the disparity in knowledge of gene expression regulatory mechanisms across populations, but also has great potential to reveal novel population-specific regulatory variants because the African Americans harbor more variants than other populations. However, there are challenges in conducting eQTL mapping in admixed population. First, since the effect size of individual eQTL is usually small, the variation in allele frequency and LD structure may lead to spurious association. For admixed population, in which the genome is mosaic of different origins, the global genomic structure adjustment such as principal component is shown to be insufficient to remove the population stratification at local regions, and the adjustment of local ancestry at each test SNP is necessary in GWAS. Second, the admixture linkage disequilibrium may mask the true association and lead to the missingness of true causal variants. Here we investigated the effects of the local ancestry adjustment on controlling the false positive rate and the ability to reveal true associated variants in eQTL mapping. We first developed a highly efficient algorithm to make it computationallymanageable to incorporate local ancestry into the associations between millions of SNPs and genes. This method was then tested in simulations and applied to publicly available AA lymphoblastoid cell line (LCL) data. Our algorithm can finish 2,576,024 association tests in 10 minutes running on Northwestern Quest. In the simulation, we compared the false positive rate and genomic control factor without adjusting for population stratification, adjusting for principle components, and adjusting for ancestry status at the test SNP. We investigated the effect size estimation assuming the gene expression is associated with genotype for the type-2 error evaluation. The similar comparison was done in the AA LCL data. Our results suggest that the principal components and local ancestry are both powerful to control the population stratification and the differential expression across population is mostly due to the overall population genomic structure instead of the ancestry at local genomic regions.

PROTEIN-ALTERING AND REGULATORY GENETIC VARIANTS NEAR *GATA4* IMPLICATED IN BICUSPID AORTIC VALVE

Wei Zhou[1], Bo Yang[2,3], Jiao Jiao[2], Jonas B Nielsen[4], Michael Mathis[5], Simon C Body[6], Gonçalo Abecasis[7], Kim Eagle[3,4], Alan P Boyle[1,8], Bicuspid Aortic Valve Consortium[6], Cristen J Willer[1,3,4,8]

[1]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Department of Cardiac Surgery, Ann Arbor, MI, [3]University of Michigan, Frankel Cardiovascular Centre, Ann Arbor, MI, [4]University of Michigan, Department of Internal Medicine, Division of Cardiovascular Medicine, Ann Arbor, MI, [5]University of Michigan, Department of Anesthesiology, Ann Arbor, MI, [6]Brigham and Women's Hospital, Harvard Medical School, Department of Anesthesiology, Perioperative, and Pain Medicine, Boston, MA, [7]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [8]University of Michigan, Department of Human Genetics, Ann Arbor, MI

Bicuspid aortic valve (BAV) is a heritable congenital aortic valve defect characterized by fusion of two of the normal three leaflets. It is the most common cardiovascular malformation in humans, with a prevalence of ~1-2% in the population. As an important risk factor for severe diseases of the valve and aorta, BAV accounts for ~40% of the annual >50,000 aortic valve replacements performed in the US. Despite the prevalence, importance, and heritability of BAV, its genetic origins remain elusive. With a goal of identifying genetic variants associated with BAV, leading to biological insight of the underlying causes, we performed one of the largest GWAS for BAV with 466 cases and 4,660 controls, with replication in additional samples of up to 1,326 cases and 8,103 controls.

We identified two genetic variants that reached or were near genome-wide significance levels ($P < 5x10^{-8}$). The strongest result was observed for a low-frequency intergenic variant rs6601627 (odds ratio (OR) = 2.38, $P_{after-replication} = 3x10^{-15}$) with a substantially higher frequency in BAV cases (8.3%) than in controls (4.2%). We also identified an independent association signal at a common protein-altering variant p.Ser377Gly (rs3729856) in *GATA4*, which encodes a cardiac-specific transcription factor that is 151 kilobases(kb) away from rs6601627 ($P_{after-replication} = 8.8x10^{-8}$). The distal region near rs6601627 was shown to be brought in close proximity to GATA4 by chromatin interaction loops in K562 and GM12878 cells that were identified using ChIA-PET and Hi-C data. To examine the role of GATA4 in valve development, we used sgRNA guided Cas9 to disrupt *GATA4* in iPSCs from a healthy human donor with normal tricuspid aortic valves and demonstrated impaired transition of endothelial into mesenchymal cells, a critical step in valve formation.

INTEGRATIVE PERSONAL OMICS PROFILES DURING PERIODS OF WEIGHT GAIN AND LOSS

Brian Piening*[1], Wenyu Zhou*[1], Kevin Contrepois*[1], Hannes Röst*[1], Gucci Gu[1], Tejaswini Mishra[1], Blake Hansen[2], Eddy Bautista[2], Shana Leopold[2], Christine Yeh[1], Daniel Spakowicz[2], Kimberly Kukurba[1], Dalia Perelman[3], et al.[1,3], Erica Sodergren[2], Tracey McLaughlin[3], George Weinstock*[2,4], Michael Snyder*[1,4]

[1]Stanford University School of Medicine, Genetics Department, Stanford, CA, [2]The Jackson Laboratory for Genomic Medicine, Farmington, CT, [3]Stanford University School of Medicine, Division of Endocrinology, Stanford, CA, [4]National Institute of Health (NIH), The iHMP consortium, Bethesda, MA

Obesity is a worldwide health epidemic and a major cause for cardiovascular disease and acquired insulin resistance leading to type 2 diabetes mellitus (T2DM). Moreover, the human body has the remarkable ability to rapidly increase body fat stores with excess caloric intake, and short-term dieting via caloric restriction is often quickly undone (i.e. “yo-yo dieting”). Despite this, the detrimental health effects of short-term excess weight gain are poorly understood. Here we performed a longitudinal study combining multiple omics strategies (genomics, transcriptomics, proteomics, metabolomics and microbiomics) to comprehensively characterize biomolecular changes in the blood and microbiomes of healthy and insulin resistant human subjects during periods of experimental weight gain and loss. Longitudinal multi-omic profiling revealed a wealth of biomolecular changes concomitant with weight gain, including pathways associated with glucose regulation/metabolism, and the activation of strong inflammatory and hypertrophic cardiomyopathy signatures in the blood. A subset of these changes returned to baseline after weight loss whereas other biomolecules remained activated multiple months following completion of the weight-loss phase, indicative of long-term physiologic changes. From these data we reconstructed biological networks associated with metabolic changes that span gene expression, proteins and the metabolite output of these cellular pathways, as well as remodeling of the microbiome. We also identify important regulatory molecules that differ among IR and IS participants as well as those that differ greatly among individuals and yet are stable to perturbations (i.e. stable personalized markers). In total, these large-scale longitudinal data offer a novel view of the rapidly-changing biomolecular landscape associated with weight gain/loss and may offer new strategies for predicting and preventing obesity-associated metabolic and/or cardiovascular disease. The data also serve as a unique resource for the scientific community.

DEEP LEARNING APPROACHES TO DENOISE, IMPUTE, INTEGRATE AND DECODE FUNCTIONAL GENOMIC DATA

Chuan Sheng Foo[1], Johnny Israeli[2], Avanti Shrikumar[1], Peyton Greenside[3], Chris Probert[4], Irene Kaplow[1], Pang Wei Koh[1], Emma Pierson[1], Anna Scherbina[3], Anshul Kundaje[1,4]

[1]Stanford University, Computer Science, Palo Alto, CA, [2]Stanford University, Biophysics, Palo Alto, CA, [3]Stanford University, Biomedical Informatics, Palo Alto, CA, [4]Stanford University, Genetics, Palo Alto, CA

We present interpretable deep learning approaches to address three key challenges in integrative analysis of functional genomic data.

(1) Data denoising: Data quality of functional genomic data is affected by a myriad of experimental parameters. Making accurate inferences from chromatin profiling experiments that involve diverse experimental parameters is challenging. We introduce a convolutional denoising algorithm to learn a mapping from suboptimal to high-quality datasets that overcomes various sources of noise and variability, substantially enhancing and recovering signal when applied to low-quality chromatin profiling datasets across individuals, cell types, and species. Our method has the potential to improve data quality at reduced costs.

(2) Data imputation: It is largely infeasible to perform 100s of genome-wide assays targeting diverse transcription factors and epigenomic marks in 100s of cellular contexts due to cost and material constraints. We have developed multi-task, multi-modal deep neural networks to predict chromatin marks and *in vivo* binding events of 100s of TFs by integrating regulatory DNA sequence with just two assays namely ATAC-seq (or DNase-seq) and RNA-seq performed in a target cell type of interest. We train our models on large reference compendia from ENCODE/Roadmap Epigenomics and obtain high prediction accuracy in new cellular contexts thereby significant expanding the context-specific annotation of the non-coding genome.

(3) Decoding the context-specific regulatory architecture of the genome: Finally, we develop novel, efficient interpretation engines for extracting predictive and biological meaningful patterns from integrative deep learning models of TF binding and chromatin accessibility. We obtain new insights into TF binding sequence affinity models (e.g. significance of flanking sequences and fusion motifs), infer high-resolution point binding events of TFs, dissect higher-order cis-regulatory sequence grammars (including density and spatial constraints), learn chromatin architectural features correlated with chromatin marks, unravel the dynamic regulatory drivers of cellular differentiation and score the regulatory influence of non-coding genetic variants.

We provide early access to all associated code and frameworks at https://github.com/kundajelab

INTERSECTING PATHOLOGY IMAGES AND GENE EXPRESSION DATA TO UNDERSTAND DRIVERS OF COMPLEX PHENOTYPES

Jordan T Ash[1], Daniel B Munro[2], Barbara E Engelhardt[1,3]

[1]Princeton University, Dept. of Computer Science, Princeton, NJ, [2]Princeton University, Dept. of Quantitative and Computational Biology, Princeton, NJ, [3]Princeton University, Center for Statistics and Machine Learning, Princeton, NJ

Understanding the correlations between genotype, gene expression levels, and high dimensional complex traits has been essential to studying the drivers of human complex disease and identifying effective therapeutic strategies for these traits. However, some complex traits are difficult to characterize in such a way as to make the quantification of correlations possible; one such complex trait is pathology imaging data. In this work, we use a type of deep learning, a convolutional autoencoder, to automatically extract one thousand features from each pathology image, and we use sparse canonical correlation analysis to correlate these pathology images with paired gene expression data on the same samples. Across three data sets, including two cancer tissue data sets and the GTEx data that include paired pathology imaging data and gene expression data, we find that our approach identifies the subset of genes that are differentially expressed with respect to specific image features, including cell size, extracellular matrix organization, cell wall thickness, and cell shape. We also pursue genotype association with pathology features in the GTEx data. We validate these associated genes and genotypes correlated with pathology image features using various approaches including gene ontology enrichment, tissue specific expression, and Mendelian randomization, allowing us to identify the drivers of cellular phenotypes. This work begins to explore the possibility of association mapping with phenotype data automatically derived from images.

ESTIMATION OF NUCLEOTIDE- AND ALLELE-SPECIFIC SELECTION COEFFICIENTS FOR PERSONAL GENOMICS USING DEEP LEARNING AND POPULATION GENETICS

Yi-Fei Huang, Adam Siepel

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

A central problem in human genomics is to understand the functional, clinical, and evolutionary significance of variants identified by genome sequencing studies. Recently, several computational methods have been developed to estimate the strength of negative selection on genomic sequences by integrating numerous weak predictors of evolutionary constraints, such as conservation scores, histone modifications, and chromatin accessibility. Methods of this kind, such as LINSIGHT, fitCons, and CADD, not only provide insights into evolutionary constraints but they also help to prioritize putative disease variants for follow-up study. However, none of the existing methods is able to estimate selection coefficients, the most interpretable measures of natural selection. On the other hand, statistical methods based on the Poisson random field model in population genetics have been widely used to estimate distributions of selection coefficients for pre-identified genomic regions, but these methods are unable to estimate the effects of individual mutations.

Here, we describe a novel statistical framework, DeepINSIGHT, that unifies methods for variant prediction and the estimation of distributions of selection coefficients, providing estimates of selection coefficients for all possible point mutations in the human genome. In particular, we formulate the estimation of selection coefficients as a regression problem in which the covariates are genomic features and the response is the observed derived allele frequency. DeepINSIGHT employs a deep-learning strategy to solve this regression problem. The method uses the log likelihood function of the Poisson random field model in place of conventional cost functions used in deep-learning models. This evolution-based cost function allows DeepINSIGHT to infer selection coefficients in a nucleotide- and allele-specific manner. In addition, as a neural network model, DeepINSIGHT is able to learn the potentially complicated nonlinear relationship between selection coefficients and genomic features. We apply DeepINSIGHT to a large number of genomic features and the high-coverage 1000 Genomes data, and show that it produces highly accurate estimates of variant-specific selection coefficients, unmatched by any existing computational methods. Using known disease variants from public databases, we show that DeepINSIGHT is a powerful method both for obtaining insights into natural selection and for prioritizing functional variants.

MODELING HUMAN GUT MICROBIOME COMMUNITY STRUCTURE ACROSS HEALTHY AND DISEASED STATES IN 2,500 TWINS

Emily R Davenport[1], Tim D Spector[2], Ruth E Ley[3], Andrew G Clark[1]

[1]Cornell University, Department of Molecular Biology and Genetics, Ithaca, NY, [2]King's College London, Department of Twin Research and Genetic Epidemiology, London, United Kingdom, [3]Max Planck Institute for Developmental Biology, Department of Microbiome Science, Tübingen, Germany

Historically, microbes were considered either pathogenic if they caused illness or benign if they lived commensally within a human host. In many cases, however, single organisms have not been identified that consistently confer disease, but rather it is believed that bacterial community structure and interactions differ between healthy and diseased individuals. Investigating gut microbiome differences between healthy and diseased individuals using a systems biology framework could lead to insight into the processes that underlie dysbiosis. Although methodology has been developed to identify co-occurrence networks from compositional microbiome data, the focus has remained on healthy microbiome datasets and we still lack an understanding of the common properties of dysbiosis. To address these gaps, we built microbiome co-occurrence networks using 16S rRNA data from ~2,500 individuals from the United Kingdom Adult Twins Registry stratified by health status for 53 diseases and 127 quantitative phenotypes, most of which are immune-related. First, disease-associated taxa were identified using generalized and linear mixed models, which take into account twin relationships between individuals. Next, microbiome co-occurrence networks were built separately for individuals i) with and without disease or ii) from opposite tails of quantitative phenotype distributions. The networks were used to identify community differences across healthy and diseased states, including comparing general network statistics (modularity and diversity), characterizing the properties of disease-associated nodes (degree, betweenness, and closeness centrality), and identifying modules of co-occurring taxa. Using these data, we are conducting one of the first large scale comparisons of microbiome dynamics across healthy and diseased individuals.

EVOLUTION OF GENE EXPRESSION AND REGULATION ACROSS THE MAMMALIAN LINEAGE

Jenny Chen[1,2], Jeremy Johnson[2], Kerstin Lindblad-Toh[2,3], Wilfried Haerty[4], Federica DiPalma[2,4], Aviv Regev[2,5]

[1]MIT, Health Sciences and Technology, Cambridge, MA, [2]Broad Institute, Cambridge, MA, [3]IMBIM, Science for Life Laboratory, Uppsala, Sweden, [4]Earlham Institute, Norwich, United Kingdom, [5]Howard Hughes Medical Institute, Biology, MIT, Cambridge, MA

The evolution of gene expression and its relationship to species phenotype is of considerable interest to the scientific community. However, no standard model for the evolution of expression currently exists to test the nature of selective forces acting on gene regulation. As such, the interpretation of comparative expression data has been highly disputed. This problem is further compounded by the lack of phylogenetically rich datasets. Using a comprehensive set of published and newly generated RNA-seq data spanning 7 tissues types across 15 species, we show that expression evolution is accurately modeled by an Ornstein-Uhlenbeck (OU) process, which enables us to identify genes under constrained, neutral, and accelerated expression evolution and characterize their corresponding regulatory properties.

We model expression evolution of 14,012 gene families using an OU model governed by two parameters: (**1**) *drift*, which drives expression away from the ancestral expression level, and (**2**) *negative selection*, which maintains expression within an optimal range. We find that within each tissue, ~20% of robustly expressed genes evolve neutrally within the mammalian lineage, diverging linearly with respect to time. Across the remaining genes, the selection parameter is directly related to function: Genes under high selective constraints are enriched for core cellular processes and are twice as likely to be essential. Conversely, transmembrane or secreted proteins carrying out tissue-specific processes such as sensory perception (brain) and muscle contraction (heart) are enriched among those displaying accelerated evolution. Surprisingly, we find that expression conservation is only weakly correlated to nucleotide conservation (R^2 = .1).

We also identify 2,859, 2,017 and 1,350 genes with lineage-specific expression changes in primates, rodents, and carnivores, respectively. These genes are enriched for functions related to lipid transport (liver), peroxisomal metabolism (liver/kidney), and spermatogenesis (testis). These changes in expression are accompanied by concomitant shifts in transcription factor binding and histone modifications, giving us clues to molecular processes for tuning gene expression.

Taken together, our work provides a theoretical framework for characterizing expression evolution and introduces an empirically-fitted selection parameter that provides a crucial context for interpreting functional genomics data across species. This framework paves the way for deeper understanding of evolutionary changes in expression and their underlying regulatory mechanisms.

MULTI-TISSUE POLYGENIC MODELS FOR TRANSCRIPTOME-WIDE ASSOCIATION STUDIES

Yongjin Park[1,2], Abhishek Sarkar[1,2], Kunal Bhutani[3], Manolis Kellis[1,2]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]University of California, Department of Bioinformatics & Systems Biology, San Diego, CA

Transcriptome-wide association studies (TWAS) have proven to be a powerful tool to identify genes associated with human diseases by aggregating cis-regulatory effects on gene expression. However, TWAS relies on building predictive models of gene expression, which are sensitive to the sample size and tissue on which they are trained. The Gene Tissue Expression Project has produced reference transcriptomes across 53 human tissues and cell types; however, the data is highly sparse, making it difficult to build polygenic models in relevant tissues for TWAS. Here, we propose `fQTL`, a multi-tissue, multivariate model for mapping expression quantitative trait loci and predicting gene expression. Our model decomposes eQTL effects into SNP-specific and tissue-specific components, pooling information across relevant tissues to effectively boost sample sizes. In simulation, we demonstrate that our multi-tissue approach outperforms single-tissue approaches in identifying causal eQTLs and tissues of action. Using our method, we fit polygenic models for 13,461 genes, characterized the tissue-specificity of the learned cis-eQTLs, and performed TWAS for Alzheimer's disease and schizophrenia, identifying 107 and 382 associated genes, respectively.

ACCURATE CLASSIFICATION OF CELLS AND TISSUES FROM HIGH THROUGHPUT MICROSCOPY IMAGES USING DEEP LEARNING

Tanel Pärnamaa[1], William Jones[2], Oliver Stegle[3], Leopold Parts[2]

[1]University of Tartu, Department of Computer Science, Tartu, Estonia, [2]Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [3]EMBL, European Bioinformatics Institute, Hinxton, United Kingdom

igh throughput microscopy of many single cells generates high-dimensional data that are far from straightforward to analyze. One important problem is automatically detecting the cellular compartment where a fluorescently tagged protein resides, a task relatively simple for an experienced human, but difficult to automate on a computer. We trained an 11-layer neural network on data from mapping thousands of yeast proteins, achieving per cell localization classification accuracy of 91%, and per protein accuracy of 99% on held out images. We confirm that low-level network features correspond to basic image characteristics, while deeper layers separate localization classes. Using this network as a feature calculator, we train standard classifiers that assign proteins to previously unseen compartments after observing only a small number of training examples. Our results are the most accurate subcellular localization classifications to date, and demonstrate the usefulness of deep learning for high throughput microscopy. We also apply the same ideas to the publicly available tissue histology images from the Genotype-Tissue Expression (GTEx) consortium.

TECHNICAL ABSTRACTS FOR WORKSHOPS

ILLUMINA WORKSHOP

Britt Flaherty, PhD, Sr. Sequencing at Illumina:

This seminar will introduce you the new Illumina | Bio-Rad Single-Cell Sequencing Solution. This platform pairs Bio-Rad's Droplet Digital Technology with Illumina's NGS library preparation and analysis technology to provide a comprehensive workflow for single-cell analysis. We will discuss the importance of single cell analysis, an overview of the key technical elements of the workflow, and address data analysis options through Illumina's BaseSpace(r) applications. Data will also be presented which demonstrates the ability of this methodology to distinguish cells in a heterogeneous population from a variety of cell sources.

NOVASEQ: THE NEXT ERA OF SEQUENCING STARTS NOW

David Miller, Manager Product Marketing at Illumina

Unveiled at the 2017 J.P. Morgan Healthcare Conference, the NovaSeq Series of Sequencing Systems is the most powerful sequencer Illumina has ever launched, redefining what's possible with next-generation sequencing.

Here we will provide both an introduction to and update on the NovaSeq systems, a single instrument that's capable of sequencing from up to 48 whole human genomes in a single run. In addition, recent application and performance data will be presented, highlighting how the NovaSeq Systems can make a wide range of applications routine, from ultra-deep sequencing of matched tumor-normal pairs, to large-scale variant discovery studies associated with complex diseases, or even low-pass sequencing of seed banks to select for specific traits.

We will also present on how NovaSeq has been designed from the ground up with over 70 innovations to allow access to next-generation sequencing technology with scalability and flexibility, enabling researchers to utilize high throughput sequencing for virtually any genome, sequencing method, and scale.

https://www.youtube.com/watch?v=y7uqvGRRqEw

https://www.illumina.com/company/news-center/feature-articles/meet-the-novaseq-series.html

https://www.illumina.com/systems/sequencing-platforms/novaseq.html

NANOPORE SEQUENCING OF A HUMAN GENOME

Hear about the latest technology updates from Oxford Nanopore as well as hearing from **Adam Phillippy, National Human Genome Research Institute,** about his latest project on the MinION.

Speakers:

- **James Brayer, Oxford Nanopore Technologies**
- **Adam Phillippy, National Human Genome Research Institute**

It is now possible to sequence and assemble human genomes using only nanopore sequencing. I will describe the results of assembling two human genomes from publicly available nanopore data, including the genome of Oxford Nanopore CTO Clive Brown and the human reference sample NA12878. The extreme length of nanopore sequencing reads, now exceeding hundreds of kilo-bases, allowed for highly continuous assemblies including the complete reconstruction of some chromosome arms. These results suggest that nanopore sequencing is now applicable to vertebrate genome sequencing and assembly, albeit with some caveats. Ongoing work aims to improve the accuracy of these assemblies using newer base-calling algorithms and complementary data types, and to detect base modifications directly from nanopore signal data.

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Prof. Hiroyuki Aburatani
The University of Tokyo
haburata-tky@umin.ac.jp

Dr. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Mr. Shaked Afik
UC Berkeley
safik@berkeley.edu

Dr. Frank Albert
University of Minnesota
falbert@umn.edu

Dr. Babak Alipanahi
23andMe
balipanahi@23andme.com

Dr. Felicity Allen
Wellcome Trust Sanger Institute
fa9@sanger.ac.uk

Dr. C. Eduardo Amorim
Stony Brook University
carloseduardo.guerraamorim@stonybrook.edu

Ms. Armande Ang Houle
University of Toronto
armande.ang.houle@gmail.com

Dr. Alexander Arguello
National Institute of Mental Health
merriwetherr@mail.nih.gov

Dr. Georgios Athanasiadis
Aarhus University
athanasiadis@birc.au.dk

Dr. Elizabeth Atkinson
Stony Brook University
elizabeth.g.atkinson@gmail.com

Dr. Mickey Atwal
Cold Spring Harbor Laboratory
atwal@cshl.edu

Dr. Adam Auton
23andMe, Inc
aauton@23andme.com

Mr. Robert Autry
St. Jude Children's Research Hospital
robert.autry@stjude.org

Ms. Gal Avital
NYU Langone
Gal.Avital@nyumc.org

Dr. Julien Ayroles
Princeton University
jayroles@princeton.edu

Dr. Md Badsha
University of Idaho
mdbadsha@uidaho.edu

Dr. Taejeong Bae
Mayo Clinic
bae.taejeong@mayo.edu

Dr. Orli Bahcall
Nature
o.bahcall@us.nature.com

Mr. Zachary Baker
Columbia University
ztb2002@columbia.edu

Dr. Benoit Ballester
INSERM TAGC
benoit.ballester@inserm.fr

Mr. Alvaro Barbeira
University of Chicago
abarbeira3@medicine.bsd.uchicago.edu

Ms. Maayan Baron
NYU Langone
Maayan.Baron@nyumc.org

Mr. Will Barr
Wesleyan University
wbarr@wesleyan.edu

Dr. Luis Barreiro
University of Montreal
luis.barreiro@umontreal.ca

Dr. Elizabeth Bartom
Northwestern University
elizabeth.bartom@gmail.com

Dr. Anindita Basu
University of Chicago
onibasu@uchicago.edu

Dr. Alexis Battle
Johns Hopkins University
ajbattle@cs.jhu.edu

Dr. Serafim Batzoglou
Illumina, Inc.
sbatzoglou@illumina.com

Dr. Jordan Beard
St. Jude Children's Research Hospital
jordan.beard@stjude.org

Dr. Paola Benaglio
University of California Dan Diego
pbenaglio@ucsd.edu

Dr. Eyal Ben-David
UCLA
ebd@ucla.edu

Dr. Andres Bendesky
Harvard University
bendesky@fas.harvard.edu

Dr. Pamela Bennett-Baker
University of Michigan
bennett@umich.edu

Dr. Jeremy Berg
Columbia University
jeremy.jackson.berg@gmail.com

Ms. Emelie Berglund
KTH Royal Institute of Technology
eberglun@kth.se

Dr. Camille Berthelot
Institut de Biologie de l'ENS
cberthel@biologie.ens.fr

Dr. Claude Bherer
New York Genome Center
cbherer@nygenome.org

Dr. Minou Bina
Purdue University
bina@purdue.edu

Dr. Rebecca Birnbaum
Icahn School of Medicine at Mount Sinai
rebecca.birnbaum@mssm.edu

Mr. Alex Bishara
Stanford University
abishara@stanford.edu

Ms. Lauren Blake
University of Chicago
leblake@uchicago.edu

Dr. Ran Blekhman
University of Minnesota
blekhman@umn.edu

Mr. Evgeni Bolotin
Faculty of Medicine, Technion
seb85il@gmail.com

Dr. Giulia Bonciani
Illumina Inc.
gbonciani@illumina.com

Ms. Amanda Bonner
Stowers Institute for Medical Research
amw@stowers.org

Dr. Mark Borowsky
Novartis Institutes for BioMedical Research
mark.borowsky@novartis.com

Mr. Aritra Bose
Purdue University
bose6@purdue.edu

Mr. Evan Boyle
Stanford University
eaboyle@stanford.edu

Ms. Margot Brandt
Columbia University
mbrandt@nygenome.org

Dr. James Brayer
Oxford Nanopore Technologies Ltd
james.brayer@nanoporetech.com

Dr. Alessandra Breschi
Centre for Genomic Regulation (CRG)
alessandra.breschi@crg.eu

Dr. Andrew Brown
University of Geneva
andrew.brown@unige.ch

Dr. Christopher Brown
University of Pennsylvania
chrbro@mail.med.upenn.edu

Ms. Ilana Buchumenski
Bar-Ilan University
ilana.sychikov@gmail.com

Dr. Gregory Buck
Virginia Commonwealth University
gregory.buck@vcuhealth.org

Dr. David Burt
Roslin Institute, University of Edinburgh
dave.burt@roslin.ed.ac.uk

Ms. Katherine Buxton
University of Wisconsin - Madison
kbuxton@chem.wisc.edu

Ms. Alicia Byrne
University of South Australia
alicia.byrne@sa.gov.au

Dr. Scott Cain
OICR
scott@scottcain.net

Dr. Gray Camp
Max Planck Institute for Evolutionary Anthropology
graycamp@gmail.com

Mr. C. Ryan Campbell
Duke University
c.ryan.campbell@duke.edu

Dr. Tom Campbell
University of Cambridge
Tom.Campbell@cruk.cam.ac.uk

Dr. Michael Campbell
Cold Spring Harbor Laboratory
mcampbel@cshl.edu

Dr. Brandi Cantarel
UTSW Medical Center
brandi.cantarel@utsouthwestern.edu

Dr. Han Cao
BioNano Genomics
han@bionanogenomics.com

Dr. Francesco Carelli
University of Cambridge
fnc21@cam.ac.uk

Dr. Piero Carninci
RIKEN Center for Life Science Technologies
carninci@riken.jp

Dr. Stephane Castel
Columbia University
scastel@nygenome.org

Ms. Monika Cechova
Pennsylvania State University
biomonika@psu.edu

Dr. Jonatas Cesar
University of Sao Paulo
jonataseduardo@gmail.com

Dr. Ernest Chan
Case Western Reserve University
erc@case.edu

Dr. Esther Chan
Stanford University
etchan@stanford.edu

Dr. Howard Chang
Stanford University School of Medicine
howchang@stanford.edu

Dr. Lesley Chapman
National Institute of Standards and Technology
lesley.chapman@nist.gov

Dr. Carole Charlier
University of Liège
carole.charlier@ulg.ac.be

Dr. Cai Chen
yangzhou university
chencai9596@163.com

Dr. Noel Chen
Novogene Corporation Inc.
noel.chen@novogene.com

Dr. Yibu Chen
University of Southern California
yibuchen@usc.edu

Dr. Beibei Chen
UT Southwestern Medical Center
beibei.chen@utsouthwestern.edu

Ms. Jenny Chen
MIT
jjenny@mit.edu

Dr. Hans Cheng
USDA, ARS
hans.cheng@ars.usda.gov

Mr. Colby Chiang
Washington University
colbychiang@wustl.edu

Dr. Nelson Chuang
University of Maryland
nchuang@umaryland.edu

Dr. Deanna Church
10X Genomics
deanna.church@10xgenomics.com

Dr. Andrew Clark
Cornell University
ac347@cornell.edu

Mr. Brian Cleary
MIT / Broad Institute
bcleary@mit.edu

Dr. Joseph Coolon
Wesleyan University
jcoolon@wesleyan.edu

Dr. Montserrat Corominas
University of Barcelona
mcorominas@ub.edu

Dr. Marc Corrales
Center for Genomic Regulation
corralesmarc@gmail.com

Dr. Justin Cotney
UConn Health
cotney@uchc.edu

Dr. Sarah Craig
Penn State University
sarah@bx.psu.edu

Ms. Ciara Curtin
GenomeWeb
ccurtin@genomeweb.com

Prof. Christina Curtis
Stanford University
cncurtis@stanford.edu

Dr. Khanh Dao Duc
University of Pennsylvania
khanhd@sas.upenn.edu

Dr. Amy Dapper
University of Wisconsin - Madison
dapper@wisc.edu

Ms. Charlotte Darby
Johns Hopkins University
cdarby@jhu.edu

Dr. Emily Davenport
Cornell University
ed379@cornell.edu

Dr. Emma Davenport
BWH, Harvard Medical School
edaven@broadinstitute.org

Dr. Jose Davila Velderrain
MIT
jdavilav@mit.edu

Ms. Jessica Davis
UCLA
jessdavis@g.ucla.edu

Dr. Aaron Day-Williams
Merck Research Laboratories
aaron.day-williams@merck.com

Dr. Jacob Degner
Abbvie
jacob.degner@abbvie.com

Dr. Olivier Delaneau
University of Geneva
olivier.delaneau@unige.ch

Dr. Laura DeMare
Cold Spring Harbor Laboratory Press
ldemare@cshl.edu

Dr. Scott Devine
Institute for Genome Science
sdevine@som.umaryland.edu

Dr. Dhananjay Dhokarh
Mayo Clinic
dhokarh.dhananjay@mayo.edu

Dr. Federica Di Palma
Ealrham Institute
federica.di-palma@earlham.ac.uk

Ms. Tonya Di Sera
University of Utah
tonyads@genetics.utah.edu

Dr. Diane Dickel
Lawrence Berkeley National Laboratory
dedickel@lbl.gov

Dr. Alexander Dobin
CSHL
dobin@cshl.edu

Dr. Elisa Donnard
Umass Medical School
elisa.donnard@umassmed.edu

Prof. Peter Donnelly
University of Oxford
donnelly@well.ox.ac.uk

Dr. Anthony Doran
Welcome Trust Sanger Institute
ad19@sanger.ac.uk

Dr. Britt Drogemoller
University of British Columbia
bdrogemoller@cmmt.ubc.ca

Mr. Noah Dukler
Cold Spring Harbor Labs
ndukler@cshl.edu

Dr. Richard Durbin
Wellcome Trust Genome Campus
rd@sanger.ac.uk

Dr. Krysta Engel
HudsonAlpha Institute for Biotechnology
kengel@hudsonalpha.org

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Dr. Yaniv Erlich
New York Genome Center/Columbia University
yaniv@cs.columbia.edu

Dr. Umut Eser
Harvard Medical School
eser@genetics.med.harvard.edu

Ms. Susan Fairley
EMBL-EBI
fairley@ebi.ac.uk

Dr. Khalid Fakhro
Sidra Medical and Research Center
KFAKHRO@SIDRA.ORG

Mr. Han Fang
Cold Spring Harbor Laboratory
hanfang.cshl@gmail.com

Dr. Andrew Farrell
University of Utah
jandrewrfarrell@gmail.com

Dr. Catherine Farrell
NIH/NLM/NCBI
farrelca@ncbi.nlm.nih.gov

Dr. Elise Feingold
NIH/National Human Genome Research Institute
Elise_Feingold@nih.gov

Dr. Ester Feldmesser
Weizmann Institute
ester.feldmesser@weizmann.ac.il

Dr. Adam Felsenfeld
National Institutes of Health
adam_felsenfeld@nih.gov

Dr. Elliott Ferris
University of Utah
elliott77@gmail.com

Mr. Jonathan Fischer
University of California, Berkeley
jrfischer@berkeley.edu

Dr. Britt Flaherty
Illumina
bflaherty@illumina.com

Dr. Paul Flicek
EMBL-EBI
flicek@ebi.ac.uk

Dr. Alexandre Fort
University of Geneva
Alexandre.Fort@unige.ch

Dr. Jingyuan Fu
University Medical Centre Groningen
fjingyuan@gmail.com

Dr. Audrey Fu
University of Idaho
audreyf@uidaho.edu

Dr. Naoko Fujito
SOKENDAI (The Graduate University for Advanced Stu
fujito_naoko@soken.ac.jp

Dr. Arkarachai Fungtammasan
DNAnexus
chai@dnanexus.com

Dr. Nicolo Fusi
Microsoft Research
fusi@microsoft.com

Dr. Pedro Galante
Hospital Sirio Libanes
pgalante@mochsl.org.br

Dr. Andrea Ganna
MGH
aganna@broadinstitute.org

Ms. Julia Garcia
Stanford University
garciajt@stanford.edu

Ms. Raquel Garcia
Universitat Pompeu Fabra
judit.sainz@upf.edu

Dr. Eugene Gardner
University of Maryland School of Medicine
egardner@umaryland.edu

Mr. Rohit Garg
Harvard Medical School
rohitgarg@g.harvard.edu

Dr. Mathew Garnett
Wellcome Trust Sanger Institute
mg12@sanger.ac.uk

Mr. Kyle Gellatly
UMASS Medical School
kyle.gellatly@umassmed.edu

Mr. Giulio Genovese
Broad Institute
giulio@broadinstitute.org

Dr. Michel Georges
ULg
michel.georges@ulg.ac.be

Dr. Mark Gerstein
Yale University
pi@gersteinlab.org

Dr. Sulagna Ghosh
The Broad Institute
ghoshs@broadinstitute.org

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Dr. David Gifford
Massachusetts Institute of Technology
gifford@mit.edu

Dr. Casey Gifford
Gladstone Institutes
casey.gifford@gladstone.ucsf.edu

Dr. Yoav Gilad
The University of Chicago
gilad@uchicago.edu

Dr. Daniel Gilchrist
NIH/National Human Genome Research Institute
daniel.gilchrist@nih.gov

Dr. Thomas Gingeras
Cold Spring Harbor Laboratory
gingeras@cshl.edu

Dr. Dominik Glodzik
Lund University/Sanger Institute
dg17@sanger.ac.uk

Dr. Andreas Gnirke
Broad Institute
gnirke@broadinstitute.org

Dr. Jonathan Goeke
Genome Institute of Singapore
gokej@gis.a-star.edu.sg

Dr. Sara Goodwin
Cold Spring Harbor Laboratory
sgoodwin@cshl.edu

Dr. Bettie Graham
National Human Genome Research Institute
grahambj@mail.nih.gov

Dr. Christopher Gregg
University of Utah
chris.gregg@neuro.utah.edu

Ms. Kristina Grigaityte
Cold Spring Harbor Laboratory
kgrigait@cshl.edu

Dr. Jeremy Grushcow
Sequence Bio
jeremy@sequencebio.com

Ms. Li Guan
University of Michigan
guanli@umich.edu

Ms. Sanna Gudmundsson
Uppsala University, SciLifeLab
sanna.gudmundsson@igp.uu.se

Dr. Adam Gudys
Silesian University of Technology
adam.gudys@gmail.com

Mr. Wilfried Guiblet
Penn State University
wilfried.guiblet@gmail.com

Dr. Rodrigo Gularte Merida
Memorial Sloan Kattering Cancer Center
gularter@mskcc.org

Dr. Brad Gulko
Cornell University/CSHL
bgulko@cs.cornell.edu

Dr. Li Guo
Xi'an Jiao Tong University
guo_li@xjtu.edu.cn

Dr. Alex Gutteridge
GSK
alex.x.gutteridge@gsk.com

Dr. Melissa Gymrek
University of California San Diego
mgymrek@ucsd.edu

Dr. Wilfried Haerty
Earlham Institute
wilfried.haerty@earlham.ac.uk

Dr. Ira Hall
Washington University School of Medicine
ihall@wustl.edu

Mr. Bob Handsaker
Broad Institute
handsake@broadinstitute.org

Dr. Drew Hardigan
HudsonAlpha Institute for Biotechnology
hardigan@uab.edu

Dr. Ross Hardison
The Pennsylvania State University
rch8@psu.edu

Dr. Timothy Harkins
Swift Biosciences
zaborski@swiftbiosci.com

Mr. Arbel Harpak
Stanford University
arbelh@stanford.edu

Dr. Kelley Harris
Stanford University
harris.kelley@gmail.com

Dr. Alan Harris
Baylor College of Medicine
rharris1@bcm.edu

Dr. Christopher Hart
Ionis Pharmaceuticals
chart@ionisph.com

Prof. Masahira Hattori
Waseda University
m-hattori@aoni.waseda.jp

Dr. Abbas Hawwari
National Guard Health Affairs
hawwariab@ngha.med.sa

Mr. Yupeng He
Salk Institute
yuhe@salk.edu

Dr. Edgar Hernandez
University of Utah
edgarjavi@gmail.com

Dr. Javier Herrero
UCL Cancer Institute
javier.herrero@ucl.ac.uk

Dr. Jason Hilton
Stanford University
jahilton@stanford.edu

Ms. Josephine Ho
Wesleyan University
jholokar@wesleyan.edu

Mr. Larson Hogstrom
Broad Institute
hogstrom@broadinstitute.org

Dr. Eurie Hong
Ancestry
ehong@ancestry.com

Dr. Julie Horvath
NC Central University/NC Museum of Natural Science
jhorvath@nccu.edu

Dr. Kerstin Howe
Wellcome Trust Sanger Institute
kj2@sanger.ac.uk

Dr. Zheng Hu
Stanford University
zhu1@stanford.edu

Prof. Lusheng Huang
Jiangxi Agricultural University
lushenghuang@hotmail.com

Dr. Yi-Fei Huang
Cold Spring Harbor Laboratory
yihuang@cshl.edu

Dr. Bernice Huang
Virginia Commonwealth University
huangb2@vcu.edu

Dr. Xiaomeng Huang
University of Utah
xm01.huang@gmail.com

Mr. Jack Humphrey
University College London
jack.humphrey@ucl.ac.uk

Dr. Matthew Hurles
Wellcome Trust Genome Campus
meh@sanger.ac.uk

Dr. Julie Hussin
Universite of Montreal
Julie.hussin@umontreal.ca

Dr. Carolyn Hutter
National Human Genome Research Institute
huttercm@mail.nih.gov

Dr. Lilia Iakoucheva
University of California San Diego
lilyak@ucsd.edu

Dr. Hae Kyung Im
The University of Chicago
haky@uchicago.edu

Dr. Marcin Imielinski
Weill Cornell Medicine
mai9037@med.cornell.edu

Ms. Kim Jaederkvist Fegraeus
Swedish University of Agricultural Sciences
kim.jaderkvist@slu.se

Dr. Andrew Jaffe
Lieber Institute for Brain Development
andrew.jaffe@libd.org

Dr. David Jaffe
10X Genomics
david.jaffe@10xgenomics.com

Mr. Karthik Jagadeesh
Stanford University
kjag@stanford.edu

Dr. Asif Javed
Genome Institute of Singapore
javeda@gis.a-star.edu.sg

Dr. Zhe Ji
Harvard Medical School & Broad Institute
zhe_ji@hms.harvard.edu

Ms. Shan Jiang
University of California, Irvine
jiangs2@uci.edu

Dr. Yinping Jiao
Cold Spring Harbor Lab
yjiao@cshl.edu

Dr. Puthen Veettil Jithesh
Sidra Medical and Research Center
pjithesh@sidra.org

Dr. Vijai Joseph
Memorial Sloan-Kettering Cancer Center
josephv@mskcc.org

Dr. Goo Jun
University of Texas Health Science Center Houston
goo.jun@uth.tmc.edu

Mr. James Kaminski
Yosef Lab - UC Berkeley
jimkaminski@berkeley.edu

Dr. Aurelie Kapusta
University of Utah
4urelie.k@gmail.com

Ms. Yukie Kashima
The University of Tokyo
0799458414@edu.k.u-tokyo.ac.jp

Dr. Alon Keinan
Cornell University
alon.keinan@cornell.edu

Dr. Eimear Kenny
Icahn School of Medicine at Mount Sinai
eimear.kenny@mssm.edu

Dr. Ekaterina Khramtsova
The University of Chicago
eakhram@gmail.com

Prof. Ekta Khurana
Weill Cornell Medicine
ekk2003@med.cornell.edu

Mr. Sid Kiblawi
University of Wisconsin-Madison
kiblawi@wisc.edu

Dr. Helena Kilpinen
University College London
helena.kilpinen@ucl.ac.uk

Dr. Seok-Won Kim
RIKEN
seokwon.kim@riken.jp

Dr. Daehwan Kim
Johns Hopkins University
infphilo@gmail.com

Dr. Charissa Kim
MD Anderson Cancer Center
ckim4@mdanderson.org

Dr. Anthony Kirilusha
National Institutes of Health
anthony.kirilusha@nih.gov

Dr. Stefan Kirov
Bristol Myers Squibb
stefan.kirov@bms.com

Dr. Cecilia Klein
Center For Genomic Regulation (CRG)
cecilia.klein@crg.es

Prof. Sung Cheol Koh
Korea Maritime University
skoh@kmou.ac.kr

Dr. Miriam Konkel
Clemson University
mkonkel@clemson.edu

Dr. Athanasios Kousathanas
Institut Pasteur
athanasios.kousathanas@pasteur.fr

Mr. Samuel Kovaka
Johns Hopkins University
skovaka1@jhu.edu

Dr. Anat Kreimer
UCB
anat.kreimer@berkeley.edu

Dr. Alper Kucukural
University of Massachusetts Medical
heidi.beberman@umassmed.edu

Dr. Martin Kuhlwilm
Universitat Pompeu Fabra
martin.kuhlwilm@upf.edu

Dr. Vivek Kumar
CSHL
vkumar@cshl.edu

Dr. Vijay Kumar
10x Genomics
vijay.kumar@10xgenomics.com

Dr. Sushant Kumar
Yale University
sushant.kumar@yale.edu

Ms. Mahalakshmi Kumaran
University of Alberta
mahalaks@ualberta.ca

Dr. Anshul Kundaje
Stanford University
akundaje@stanford.edu

Dr. Kasper Lage
MGH / Harvard / Broad Institute
lage.kasper@mgh.harvard.edu

Dr. Avantika Lal
Stanford University
avlal@stanford.edu

Dr. Fabien Lamaze
Ontario Institute for Cancer Research
fabien.lamaze@gmail.com

Mr. Tianming Lan
BGI-Shenzhen
lantianming@genomics.cn

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. Jean-Marc Lassance
Harvard University/HHMI
lassance@fas.harvard.edu

Dr. Ryan Layer
University of Utah
ryan.layer@gmail.com

Ms. Amanda Lea
Duke University
amandalea7180@gmail.com

Mr. Dillon Lee
University of Utah
dlee123@gmail.com

Dr. Sandra Lee
Stanford University
sandra.lee@stanford.edu

Mr. Arthur Lee
Washington University School of Medicine
arthurlee@wustl.edu

Dr. Su-In Lee
University of Washington
suinlee@cs.washington.edu

Dr. Ellen Leffler
University of Oxford
leffler@well.ox.ac.uk

Dr. Kalle Leppala
Aarhus University
kalle.m.leppala@gmail.com

Dr. Stephen Levene
The University of Texas at Dallas
stephen.levene@utdallas.edu

Dr. Mingkun Li
Fondation Merieux
fengzys@163.com

Dr. Ting Li
Genomic Health, Inc
tingli2010@gmail.com

Dr. Dawei Li
University of Vermont
dawei.li@uvm.edu

Dr. Stephen Lincoln
Invitae
steve.lincoln@me.com

Dr. Yunlong Liu
Indiana University Purdue University - Indianapolis
yunliu@iu.edu

Dr. Qinwen Liu
Grail, Inc.
qliu@grailbio.com

Mr. Dianbo Liu
MIT/University of Dundee
dianbo@mit.edu

Dr. Dajiang Liu
Pennsylvania State University
dajiang.liu@outlook.com

Dr. Tzu-Yu Liu
University of Pennsylvania
tzuyuliu@sas.upenn.edu

Dr. Yu Liu
St. Jude Children's Research Hospital
yu.liu@stjude.org

Dr. Paul Lloyd
University of California San Francisco
paul.lloyd@ucsf.edu

Dr. Jason Lloyd-Price
Broad Institute
jasonlp@broadinstitute.org

Ms. Jennifer Lu
Johns Hopkins University
jlu26@jhmi.edu

Dr. Francesca Luca
Wayne State University
fluca@wayne.edu

Dr. Elise Lucotte
Aarhus University
elucotte@gmail.com

Dr. Ruibang Luo
The Johns Hopkins University
rluo5@jhu.edu

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Ms. Heather Machado
Stanford University
machadoheather@gmail.com

Mr. Sho Maekawa
The University of Tokyo
smaekawa@hgc.jp

Dr. Gabriele Magris
Universita degli Studi di Udine
gmagris@appliedgenomics.org

Dr. Thomas Mailund
Aarhus University
mailund@birc.au.dk

Dr. Vladimir Makarov
Swift Biosciences
zaborski@swiftbiosci.com

Dr. Kateryna Makova
Penn State University
kdm16@psu.edu

Mr. Venkat Malladi
University of Texas Southwestern Medical Center
Venkat.Malladi@utsouthwestern.edu

Dr. Elaine Mardis
Nationwide Children's Hospital Research Institute
elaine.mardis@nationwidechildrens.org

Dr. Marco Mariotti
Brigham and Women's Hospital; Harvard Med School
mmariotti@bwh.harvard.edu

Prof. Gabor Marth
University of Utah
gmarth@genetics.utah.edu

Dr. Alicia Martin
Massachusetts General Hospital
armartin@broadinstitute.org

Dr. Teresa Martinez
Stony Brook University
teresa.martinez@stonybrook.edu

Dr. Alexander Martinez Fundichely
Weill Cornell Medical College
alm2069@med.cornell.edu

Dr. Davis McCarthy
EMBL-EBI
davis@ebi.ac.uk

Dr. William McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Dr. Brian McLoone
University of Wisconsin–Madison
brianbmcloone@gmail.com

Dr. Tarang Mehta
Earlham Institute
tarang.mehta@earlham.ac.uk

Dr. Maria Mejia Guerra
Cornell University
MM2842@cornell.edu

Ms. Katherine Melville
Oxford Nanopore Technologies
katherine.melville@nanoporetech.com

Dr. Jason Merkin
KSQ THERAPEUTICS, INC.
jmerkin@ksqtx.com

Dr. Matthias Meyer
Max Planck Institute for Evolutionary Anthropology
mmeyer@eva.mpg.de

Mr. Zong Miao
University of California, Los Angeles
zmiao@ucla.edu

Mr. Sjors Middelkamp
UMC Utrecht
s.h.a.middelkamp-2@umcutrecht.nl

Dr. Chase Miller
University of Utah, Center for Genetic Discovery
chmille4@gmail.com

Mr. David Miller
Illumina
dmiller3@illumina.com

Dr. Michael Montague
University of Pennsylvania
mike.j.montague@gmail.com

Dr. Sandrine Moreira
Columbia University
sandrine.moreira.rousseau@gmail.com

Dr. Alexander Morgan
Khosla Ventures
alex@khoslaventures.com

Dr. Leonid Moroz
University of Florida
moroz@whitney.ufl.edu

Mr. Hakhamanesh Mostafavi
Columbia University
hsm2137@columbia.edu

Dr. Jonathan Mudge
Wellcome Trust Sanger Institute
jm12@sanger.ac.uk

Dr. Swagatam Mukhopadhyay
Ionis Pharmaceuticals
SMukhopadhyay@ionisph.com

Mr. Rabi Murad
UC Irvine
rmurad@uci.edu

Mr. David Murphy
Columbia University
dam2214@columbia.edu

Dr. Gemma Murray
University of California Santa Cruz
gemurray@ucsc.edu

Dr. Ramaiah Nagaraja
National Institute on Aging
nagarajar@mail.nih.gov

Dr. Yuya Nakajima
Keio University
notrefraisier@gmail.com

Dr. Narisu Narisu
NIH
narisu@mail.nih.gov

Mr. Alexander Nash
Imperial College London
alexander.nash13@imperial.ac.uk

Dr. Waleed Nasser
Baylor College of Medicine
nasser@bcm.edu

Mr. Fabio Navarro
Yale University
fabio.navarro@yale.edu

Mr. Yoav Naveh
MyHeritage
yoav.naveh@myheritage.com

Dr. Nicholas Navin
MD Anderson Cancer Center
nnavin@mdanderson.org

Dr. Tal Nawy
Nature Methods
t.nawy@us.nature.com

Dr. Benjamin Neale
Massachusetts General Hospital
neale@atgu.mgh.harvard.edu

Dr. Vishwa Nellore
Duke University
vishwa.nellore@duke.edu

Dr. Alondra Nelson
Columbia University
alondra.nelson@columbia.edu

Dr. Tan-Hoang Nguyen
The Icahn School of Medicine at Mount Sinai
tan-hoang.nguyen@mssm.edu

Dr. Jonas Nielsen
University of Michigan
jbnie@umich.edu

Dr. Suguru Nishijima
National Institute of Advanced Industrial Science
nishijima.suguru@aist.go.jp

Mr. Conor Nodzak
University of North Carolina, Charlotte
cnodzak@uncc.edu

Dr. Sahar Nohzadeh
Fabricgenomics, Inc.
snohza@fabricgenomics.com

Dr. Tsviya Olender
The Weizmann Institute
tsviya.olender@weizmann.ac.il

Dr. Morten Olesen
Rigshospitalet
morten.salling.olesen@gmail.com

Ms. Meritxell Oliva
University Of Chicago
meritxellop@uchicago.edu

Dr. Hanna Ollila
Stanford University School of Medicine
hannao@stanford.edu

Dr. Louise Ormond
Ecole Polytechnique Federale de Lausanne
louise.ormond@epfl.ch

Mr. Omead Ostadan
Illumina, Inc.
oostadan@illumina.com

Dr. Toshio Ota
Kyowa Hakko Kirin Co., Ltd.
toshio.ota@kyowa-kirin.co.jp

Dr. Francis Ouellette
GENOME QUEBEC
francis@genomequebec.com

Prof. Svante Paabo
Max Planck Institute for Evolutionary Anthropology
paabo@eva.mpg.de

Dr. Athma Pai
Massachusetts Institute of Technology
athma@mit.edu

Prof. Paivi Pajukanta
University of California, Los Angeles (UCLA)
ppajukanta@mednet.ucla.edu

Dr. Luisa Pallares
Princeton Univeristy
pallares@princeton.edu

Dr. Yongjin Park
MIT
ypp@csail.mit.edu

Ms. Princy Parsana
Johns Hopkins University
princy@jhu.edu

Dr. Leopold Parts
Wellcome Trust Sanger Institute
lp2@sanger.ac.uk

Mr. Anthony Payne
University of Oxford
apayne@well.ox.ac.uk

Dr. Brent Pedersen
University of Utah
bpederse@gmail.com

Dr. Dana Pe'er
Sloan Kettering Institute/MSKCC
peerster@gmail.com

Dr. Elizabeth Pennisi
Science
epennisi@aaas.org

Dr. Minoli Perera
Northwestern University
minoli.perera@northwestern.edu

Dr. Marc Perry
University of California, San Francisco
marc.perry@ucsf.edu

Dr. Dmitri Pervouchine
Skolkovo Institute of Science and Technology
d.pervouchine@skoltech.ru

Dr. Nitin Phadnis
University of Utah
nitin.phadnis@utah.edu

Dr. Lon Phan
NIH
hullja@mail.nih.gov

Dr. Adam Phillippy
National Human Genome Research Institute
adam.phillippy@nih.gov

Ms. Lenore Pipes
Cold Spring Harbor Laboratory
lpipes@cshl.edu

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Dr. Milton Pividori
The University of Chicago
miltondp@uchicago.edu

Mr. Jason Pizzollo
University of Massachusetts Amherst
jpizzollo@umass.edu

Dr. Linda Polfus
University of Kentucky
bgoodley@uky.edu

Dr. Dimitris Polychronopoulos
Imperial College London
dpolychr@imperial.ac.uk

Ms. Hagit Porath
Bar-Ilan University
Hagit_br@hotmail.com

Dr. Daniil Prigozhin
University of Cambridge
prigozhin@gmail.com

Dr. Jonathan Pritchard
Stanford University
ttrim@stanford.edu

Dr. Kay Prufer
Max Planck Institute for Evolutionary Anthropology
pruefer@eva.mpg.de

Dr. Yi Qiao
University of Utah
qiaoy01@gmail.com

Prof. Francis Quetier
GIP GENOPOLE
francis.quetier@genopole.fr

Dr. Jorge Quintana Kageyama
Max Planck Institute For Evolutionary Anthropology
jorge_kageyama@eva.mpg.de

Dr. Fernando Racimo
New York Genome Center
fernandoracimo@gmail.com

Dr. Towfique Raj
Icahn School of Medicine at Mount Sinai
towfique.raj@mssm.edu

Mr. Ryne Ramaker
HudsonAlpha/University of Alabama at Birmingham
ryneramaker@gmail.com

Dr. Srividya Ramakrishnan
Johns Hopkins University
srividya.ramki@gmail.com

Dr. Daniele Ramazzotti
Stanford University
daniele.ramazzotti@stanford.edu

Ms. Shweta Ramdas
University of Michigan
sramdas@umich.edu

Dr. Sarah Ratzel
American Journal of Human Genetics
sratzel@ajhg.net

Dr. Martin Reese
Omicia Inc.
mreese@omicia.com

Dr. Aviv Regev
Broad Institute of MIT & Harvard
aregev@broadinstitute.org

Prof. Jun Ren
Jiangxi Agricultural University
renjunjxau@hotmail.com

Mr. Andre Rendeiro
CeMM Research Center for Molecular Medicine of the
arendeiro@cemm.oeaw.ac.at

Dr. Sarah Rennie
University of Copenhagen
sarah@binf.ku.dk

Dr. Armando Reyes-Palomares
EMBL
armando.reyes@embl.de

Ms. Charlotte Rich
University of Warwick
c.s.rich@warwick.ac.uk

Dr. Jose Rodriguez-Martinez
University of Puerto Rico - Rio Piedras
jose.rodriguez233@upr.edu

Dr. Mostafa Ronaghi
Illumina, Inc.
mronaghi@illumina.com

Dr. Jeffrey Rosenfeld
Rutgers University
jeffrey.rosenfeld@rutgers.edu

Dr. Maxime Rotival
Institut Pasteur
maxime.rotival@pasteur.fr

Dr. Tanmoy Roychowdhury
Mayo Clinic
roychowdhury.tanmoy@mayo.edu

Ms. Mariana Ruiz Velasco
EMBL Heidelberg
mariana.ruiz@embl.de

Mr. Jorge Ruiz-Orera
Hospital del Mar Medical Research Institute Foundation (FIMIM)
jruiz@imim.es

Dr. Tina Saey
Science News
tsaey@sciencenews.org

Dr. Kan Saito
EMD Millipore
kan.saito@emdmillipore.com

Dr. Marie Saito
State University of New York at Buffalo
mariesaitou@gmail.com

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Dr. Adam Santanasto
University of Pittsburgh
ajs51@pitt.edu

Dr. Aliya Saperstein
Stanford University
asaper@stanford.edu

Mr. Thomas Sasani
University of Utah
tom.sasani@utah.edu

Mr. Nathan Schaefer
University of California, Santa Cruz
nkschaefer@soe.ucsc.edu

Dr. Michael Schatz
Cold Spring Harbor Laboratory and Johns Hopkins
mschatz@cshl.edu

Dr. Melanie Schirmer
The Broad Institute of MIT and Harvard
melanie@broadinstitute.org

Dr. Joshua Schmidt
Max Planck Institute for Evolutionary Anthropology
joshua_schmidt@eva.mpg.de

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@ncbi.nlm.nih.gov

Dr. Olga Schubert
University of California, Los Angeles
olga.schubert@gmail.com

Dr. Molly Schumer
Harvard University/Columbia University
schumerm@gmail.com

Dr. David Schwartz
University of Wisconsin - Madison
dcschwartz@wisc.edu

Dr. Roland Schwarz
Max Delbrueck Center for Molecular Medicine
roland.schwarz@mdc-berlin.de

Dr. Alan Scott
Johns Hopkins University
afscott@jhmi.edu

Dr. Laura Scott
University of Michigan
ljst@umich.edu

Dr. Jonathan Sebat
UC San Diego
jsebat@ucsd.edu

Dr. Alisa Sedghifar
Princeton University
asedghifar@princeton.edu

Dr. Fritz Sedlazeck
Johns Hopkins University
fritz.sedlazeck@gmail.com

Dr. Ayellet Segre
Broad Institute
asegre@broadinstitute.org

Dr. Masahide Seki
The University of Tokyo
mseki@edu.k.u-tokyo.ac.jp

Dr. Myrna Serrano
Virginia Commonwealth University
mgserrano@vcu.edu

Ms. Afrah Shafquat
Cornell University
as3397@cornell.edu

Dr. Maxwell Shapiro
Stony Brook University
maxwell.shapiro@stonybrook.edu

Dr. Shehzad Sheikh
UNC Chapel Hill
shehzad_sheikh@med.unc.edu

Dr. Jay Shendure
University of Washington
shendure@uw.edu

Ms. Rachel Sherman
Johns Hopkins University
rsherman@jhu.edu

Ms. Ruchi Sheth
Wesleyan University
rsheth@wesleyan.edu

Dr. Atsushi Shimizu
Iwate Medical University
ashimizu@iwate-med.ac.jp

Dr. Massa Shoura
Stanford University
massa86@stanford.edu

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Mr. Nasa Sinnott-Armstrong
Stanford University
nasa@stanford.edu

Dr. Cristina Sisu
Yale University
cs784@gersteinlab.org

Dr. Michael Snyder
Stanford University
mpsnyder@stanford.edu

Dr. Noah Snyder-Macker
Duke University
nsmack@gmail.com

Prof. Nicole Soranzo
Wellcome Trust Sanger Institute
ns6@sanger.ac.uk

Mr. Pieter Spealman
Carnegie Mellon University
pspealman@cmu.edu

Dr. Noah Spies
NIST/Stanford University
nspies@stanford.edu

Ms. Elena Stamenova
Broad Institute
stamenov@broadinstitute.org

Dr. Arnold Stein
Purdue University
steina@purdue.edu

Dr. Joshua Stein
Cold Spring Harbor Laboratory
steinj@cshl.edu

Dr. Thomas Stoeger
Northwestern University
thomas.stoeger@northwestern.edu

Dr. Marcus Stoiber
Lawerence Berkeley Labs
mhstoiber@lbl.gov

Dr. Tomasz Stokowy
University of Bergen
tomasz.stokowy@k2.uib.no

Dr. Barbara Stranger
University of Chicago
bstranger@medicine.bsd.uchicago.edu

Dr. J Seth Strattan
Stanford University
jseth@stanford.edu

Dr. Wataru Suda
The University of Tokyo
wataru_suda@cb.k.u-tokyo.ac.jp

Dr. Peter Sudmant
MIT
psudmant@mit.edu

Dr. Qi Sun
Cornell University
qisun@cornell.edu

Prof. Shamil Sunyaev
Brigham & Women's Hospital, Harvard Medical School
ssunyaev@rics.bwh.harvard.edu

Dr. Hillary Sussman
Cold Spring Harbor Laboratory Press
hsussman@cshl.edu

Dr. Yoshihiko Suzuki
The University of Tokyo
suzuki_yoshihiko_15@stu-cbms.k.u-tokyo.ac.jp

Dr. Leila Taher
University of Erlangen-Nuremberg
leila.taher@fau.de

Mr. Yusuke Takahashi
The University of Tokyo
hisakatha@gmail.com

Dr. Lena Takayasu
The University of Tokyo
lena_takayasu@cb.k.u-tokyo.ac.jp

Dr. Michael Talkowski
Harvard Medical School
mtalkowski@mgh.harvard.edu

Dr. Jiaying Tan
CELL PRESS
jtan@cell.com

Dr. Todd Taylor
RIKEN
taylor@riken.jp

Mr. Levi Teitz
Whitehead Institute for Biomedical Research
lsteitz@mit.edu

Dr. Marcela Tello-Ruiz
Cold Spring Harbor Laboratory
mmonaco@cshl.edu

Dr. James Thomas
NIH
thomasjw4@mail.nih.gov

Ms. Grace Tiao
Broad Institute
gtiao@broadinstitute.org

Dr. Hagen Tilgner
Weill Cornell Medicine
hagen.u.tilgner@gmail.com

Dr. Vladimir Timoshevskiy
University of Kentucky
vti224@uky.edu

Dr. Richard Trembath
King's College London
dean-folsm@kcl.ac.uk

Dr. Barbara Treutlein
Max Planck Institute for Evolutionary Anthropology
barbara_treutlein@eva.mpg.de

Dr. Marco Trizzino
The Wistar Insitute
marco.trizzino83@gmail.com

Dr. Jenny Tung
Duke University
jt5@duke.edu

Mr. Sebastian Ullrich
Center for Genomic Regulation (CRG)
sebastian.ullrich@crg.eu

Ms. Lara Urban
EMBL-EBI
lurban@ebi.ac.uk

Ms. Oana Ursu
Stanford University School of Medicine
oursu@stanford.edu

Dr. Raditya Utama
Cold Spring Harbor Laboratory
Utama@cshl.edu

Dr. Anton Valouev
Grail Bio
avalouev@grailbio.com

Dr. Bryce van de Geijn
Harvard School of Public Health
vandegeijn@hsph.harvard.edu

Dr. Karine Van Doninck
University of Namur
karine.vandoninck@unamur.be

Ms. Pranitha Vangala
University Of Massachusetts Medical School
pranitha.vangala@umassmed.edu

Ms. Pajau Vangay
University of Minnesota
vanga015@umn.edu

Ms. Srinidhi Varadharajan
University of Oslo
srinidhivaradharajan@gmail.com

Dr. Matthew Velinder
University of Utah
matt.velinder@utah.edu

Prof. Byrappa Venkatesh
Institute of Molecular and Cell Biology
mcbbv@imcb.a-star.edu.sg

Dr. Oliver Venn
GRAIL, INC.
ovenn@grailbio.com

Dr. Eric Venner
Baylor College of Medicine
venner@bcm.edu

Mr. Ted Verhey
University of Calgary
tbverhey@ucalgary.ca

Dr. Benjamin Vernot
Max Planck Institute for Evolutionary Anthropology
benjamin_vernot@eva.mpg.de

Dr. Tomas Vinar
Comenius University in Bratislava
tvinar@gmail.com

Dr. Ana Vinuela
University of Geneva
ana.vinuela@unige.ch

Dr. Morana Vitezic
University of Copenhagen
mvitezic@gmail.com

Ms. Elena Vizcaya
UB/CRG
elenavizcayamolina@gmail.com

Dr. Dragana Vuckovic
University of Trieste
dragana.vuckovic@burlo.trieste.it

Dr. yong wang
Ancestry
ywang@ancestry.com

Dr. Lu Wang
NIH/NHGRI
lu.wang@mail.nih.gov

Dr. Miaoyan Wang
University of Pennsylvania
miaoyan@sas.upenn.edu

Dr. Leyao Wang
Yale University
leyao.wang@yale.edu

Dr. Bo Wang
Cold Spring Harbor Lab
bwang@cshl.edu

Ms. Pankhuri Wanjari
Northwestern University
pankhuri.wanjari@northwestern.edu

Dr. Alistair Ward
University of Utah
alistairnward@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory/USDA/ARS
ware@cshl.edu

Ms. Elizabeth Waters
University of Wisconsin-Madison
ewaters2@wisc.edu

Dr. Michael Weir
Wesleyan University
mweir@wesleyan.edu

Mr. Jia Wen
The Chinese University of Hong Kong
wenjia198021@126.com

Dr. Jia Wen
UNC Charlotte
jwen6@uncc.edu

Dr. Harm-Jan Westra
BWH, Harvard Medical School
westra.harmjan@outlook.com

Ms. Kris Wetterstrand
National Human Genome Research Institute, NIH
wettersk@mail.nih.gov

Dr. Owen White
University of Maryland School of Medicine
owhite@som.umaryland.edu

Dr. Maria Wilbe
Uppsala University
maria.wilbe@igp.uu.se

Dr. Cristen Willer
University of Michigan
cristen@umich.edu

Dr. Amy Williams
Cornell University
awilliams@cornell.edu

Dr. Eva-Marie Willing
MPI for Plant Breeding Research
willing@mpipz.mpg.de

Dr. Richard Wilson
Nationwide Children's Hospital
richard.wilson@nationwidechildrens.org

Dr. Peter Wilton
University of California, Berkeley
pwilton@berkeley.edu

Mr. Eamon Winden
University of Wisconsin, Madison
ewinden@wisc.edu

Dr. Shari Wiseman
Nature Neuroscience
shari.wiseman@us.nature.com

Ms. Brooke Wolford
University of Michigan
bwolford@umich.edu

Dr. Hyejung Won
UCLA
wonhyejung87@gmail.com

Dr. Kim Worley
Baylor College of Medicine
kworley@bcm.edu

Dr. Galen Wright
University of British Columbia
gwright@cmmt.ubc.ca

Dr. Tomasz Wrzesinski
Earlham Institute
tomasz.wrzesinski@earlham.ac.uk

Ms. Xiaoli Wu
Cold Spring Harbor Laboratory
xlw1207@gmail.com

Dr. Simon Xi
Pfizer
hualin.xi@pfizer.com

Dr. Yilin Xu
Northwestern University
yellen10@gmail.com

Ms. Chenling Xu
U.C. Berkeley
chenlingantelope@gmail.com

Dr. Anupama Yadav
Dana-Farber Cancer Institute
anupama_yadav@dfci.harvard.edu

Prof. Mark Yandell
University of Utah
myandell@genetics.utah.edu

Prof. Huanming Yang
BGI-China
yanghm@genomics.cn

Dr. Bin Yang
Jiangxi Agricultural University
binyang@live.cn

Dr. Moran Yassour
The Broad Institute of MIT and Harvard
moran@broadinstitute.org

Dr. Lynn Young
National Institutes of Health
lynny@mail.nih.gov

Mr. Yun Yu
MIT
ywy@mit.edu

Dr. Jinzhou Yuan
Columbia University Medical Center
jy2756@columbia.edu

Dr. Laura Zahn
AAAS/Science
lzahn@aaas.org

Mr. Gregory Zajac
University of Michigan
gzajac@umich.edu

Mr. Harel Zalts
Technion
harelzalts@gmail.com

Ms. Samantha Zarate
DNAnexus
szarate@dnanexus.com

Dr. Judith Zaugg
EMBL
judith.zaugg@embl.de

Mr. Haoyang Zeng
Massachusetts Institute of Technology
haoyangz@mit.edu

Dr. Daniel Zerbino
EMBL-EBI
zerbino@ebi.ac.uk

Dr. Zhaojie Zhang
H3 Biomedicine
zhaojie_zhang@h3biomedicine.com

Ms. Ying Zhang
Hua Zhong University of Science and Technology
ying_zh@hust.edu.cn

Prof. Bo Zhang
Peking University
bzhang@pku.edu.cn

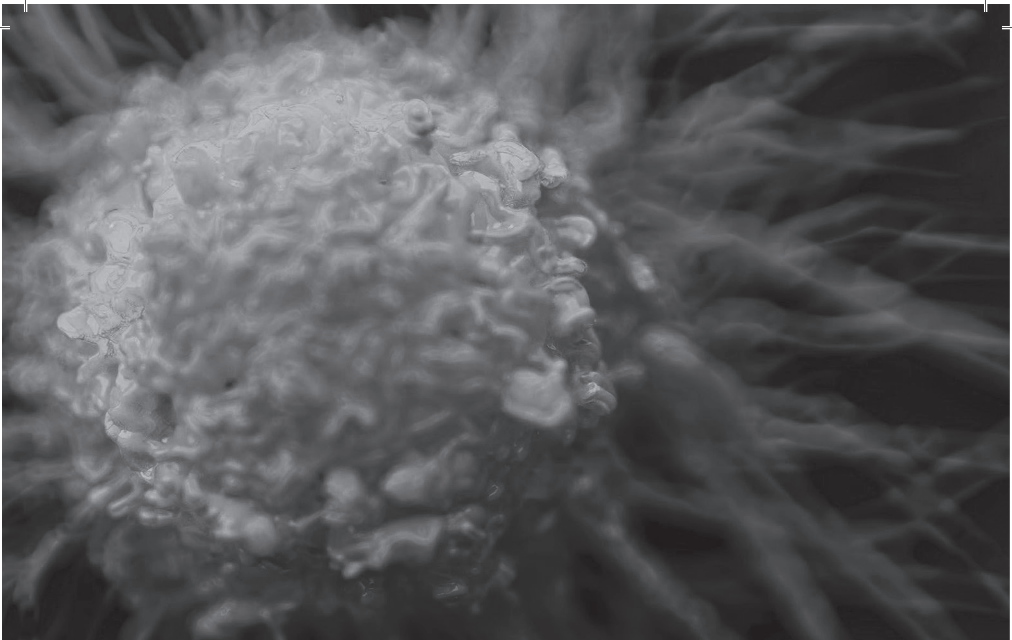
Ms. Yizhen Zhong
Northwestern University
yizhenzhong2015@u.northwestern.edu

Dr. zhengfu zhou
bri
fadsa@sdd.com

Ms. Wei Zhou
University of Michigan
zhowei@umich.edu

Dr. Wenyu Zhou
Stanford University
wenyuz@stanford.edu

Dr. Peter Zimmerman
Case Western Reserve University
paz@case.edu



ONE PLATFORM. INFINITE POSSIBILITIES.

Whether you are interested in high-output applications like whole-genome sequencing (WGS), looking to complete multi-omics studies, or simply operate more cost effectively, the NovaSeq™ Series empowers you with new possibilities.

The platform is enabled by the broad portfolio of Illumina library preparation kits, providing the flexibility to run your applications on any scale—from single cells to an entire population.

Learn about the latest advancements in sequencing technology and applications with the NovaSeq Series during our luncheon workshop on May 10.

Power your research with the NovaSeq Series
www.illumina.com/novaseq

illumina®

For Research Use Only. Not for use in diagnostic procedures.

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	**631-351-2000** **(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# 61170** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access

Grace Auditorium
Upper level: E-mail and printing in the business center area
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar

Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late (Cash Only)
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes, ATM

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m.
Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, east wing, lower level
PIN#: **Press 61170 (then enter #)**

Meetings & Courses Front Office

Hours during meetings: 8am – 7pm, until 9pm on arrival day
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**

Local Interest

Fish Hatchery	631-692-6758
Sagamore Hill	516-922-4788
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($9.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
US Limousine Service	800-962-2827,ext:3 **(1047)**
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	877-690-5114
New Jersey Transit	973-275-5555

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
British Airways	800-247-9297
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Southwest Airlines	800-435-9792
United	800-241-6522
Virgin American	877-359-9792