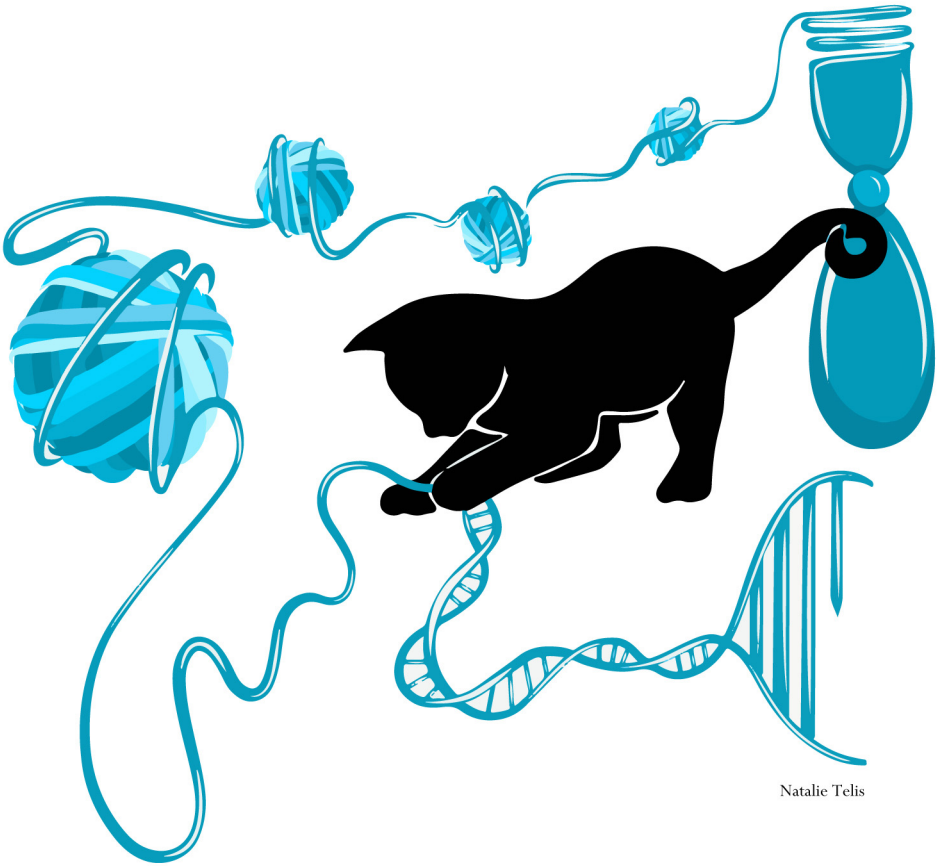
Abstracts of papers presented
at the 2018 meeting on

# THE BIOLOGY OF GENOMES

May 8–May 12, 2018




CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2018 meeting on

# THE BIOLOGY OF GENOMES

May 8–May 12, 2018

Arranged by

Matthew Hurles, *Wellcome Trust Sanger Institute*
Elaine Ostrander, *National Human Genome Research 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;** and **Oxford Nanopore Technologies.**

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
Lundbeck
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.*

---

*Cover:* "Unraveling the genome." By Natalie Telis http://www.telis.blog.

**THE BIOLOGY OF GENOMES**
Tuesday, May 8 – Saturday, May 12, 2018

| | | |
|---|---|---|
| Tuesday | 7:30 pm | **1** Genome Engineering and Editing |
| Tuesday | *Following eve. session* | *Happy Hour sponsored by Illumina* |
| Wednesday | 9:00 am | **2** Population Genomics |
| Wednesday | 1:30 pm | Introduction to the NHGRI Strategic Planning Process |
| Wednesday | 2:00 pm | **3** Poster Session I |
| Wednesday | 4:30 pm | *Wine and Cheese Party* |
| Wednesday | 7:30 pm | **4** Functional Genomics and Epigenetics |
| Thursday | 9:00 am | **5** Evolutionary and Non-human Genomics |
| Thursday | 2:00 pm | **6** Poster Session II |
| Thursday | 4:30 pm | **7** ELSI Panel and Discussion |
| Thursday | 7:30 pm | **8** Cancer and Medical Genomics |
| Friday | 9:00 am | **9** Computational 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** Complex Traits and Microbiome |

*Workshops (following morning session, Hershey Building)*
Wednesday: Oxford Nanopore *(see p. T-1)*
Thursday: Illumina

Mealtimes at Blackford Hall are as follows:
Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm
Bar is open from 5:00 pm until late

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

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

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

Any discussion via social media platforms of material presented at this meeting requires explicit permission from the presenting author(s).

*Printed on 100% recycled paper.*

PROGRAM

TUESDAY, May 8—7:30 PM

**SESSION 1** GENOME ENGINEERING AND EDITING

**Chairpersons:** **Jef Boeke,** NYU Langone Health, New York, New York
**Feng Zhang,** Broad Institute of MIT and Harvard, Cambridge, Massachusetts

***Happy Hour***
Sponsored by **Illumina**

WEDNESDAY, May 9—9:00 AM

**SESSION 2** POPULATION GENOMICS

**Chairpersons:** **Mattias Jakobsson,** Uppsala University, Sweden
**Sarah Tishkoff,** University of Pennsylvania, Philadelphia

WEDNESDAY, May 9—1:30 PM

**Introduction to the NHGRI Strategic Planning Process**

**SESSION 3** POSTER SESSION I

WEDNESDAY, May 9—4:30 PM

**Wine and Cheese Party**

WEDNESDAY, May 9—7:30 PM

**SESSION 4** FUNCTIONAL GENOMICS AND EPIGENETICS

**Chairpersons:** **Job Dekker,** HHMI, UMass Medical School, Worcester, Massachusetts
**Emma Farley,** University of California, San Diego

THURSDAY, May 10—9:00 AM

**SESSION 5** EVOLUTIONARY AND NON-HUMAN GENOMICS

**Chairpersons:** **Monica Justice,** The Hospital for Sick Children, Toronto, Canada
**Gavin Sherlock,** Stanford University, California

THURSDAY, May 10—2:00 PM

**SESSION 6** POSTER SESSION II

**Incidence of uniparental disomy in 2 million individuals from the 23andMe database**
Priyanka Nakka, Kimberly McManus, 23andMe Research Team, Anne O'Donnell-Luria, Uta Francke, Sohini Ramachandran, Joanna Mountain, Fah Sathirapongsasuti.
Presenter affiliation: Brown University, Providence, Rhode Island. 192

**Stepwise evolution of sex-biased gene expression in mammalian tissues**
Sahin Naqvi, Alexander K. Godfrey, Jennifer F. Hughes, Mary L. Goodheart, Richard N. Mitchell, David C. Page.
Presenter affiliation: Whitehead Institute, Cambridge, Massachusetts; Massachusetts Institute of Technology, Cambridge, Massachusetts. 193

**Comprehensive survey of LINE-1 transcriptional activity in human cell lines, healthy tissue, and tumors**
Fabio Navarro, Jacob Hoops, Lauren Bellfy, Eliza Cerveira, Qihui Zhu, Chengsheng Zhang, Charles Lee, Mark B. Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 194

**Identification of loci associated with embryonic and fetal loss in Holstein heifers**
Holly L. Neibergs, Jennifer N. Kiser, Joseph Dalton, Joao G. Moraes, Thomas E. Spencer.
Presenter affiliation: Washington State University, Pullman, Washington. 195

**An extended *msprime* coalescent simulation framework avoids biases from large sample sizes, and increases performance when simulating long regions**
Dominic Nelson, Jerome Kelleher, Gil McVean, Simon Gravel.
Presenter affiliation: McGill University, Montreal, Canada. 196

THURSDAY, May 10—4:30 PM

**SESSION 7** ELSI PANEL AND DISCUSSION

**"Human Gene Editing: Traversing the Germline"**

**Moderator:** **Nicole C. Lockhart, Ph.D.,** NIH/National Human Genome Research Institute

**Panelists**

**Ryan Fischer**, Parent Project Muscular Dystrophy
**Rosario Isasi**, University of Miami
**Debra JH Mathews**, Johns Hopkins Berman Institute of Bioethics
**Fyodor Urnov**, Altius Institute for Biomedical Sciences

Powerful gene editing tools are revolutionizing the field of biology and the ability to model and understand disease. The increasing efficiency and precision of gene editing tools is democratizing research in new and unexpected ways. As the field continues to advance, long standing ethical frameworks are being challenged and the development of new research guidelines is the subject of much debate. Several commentators and expert groups have drawn a strong line at the use of these tools for human germline gene editing, but are there situations in which germline gene editing is not only permissible, but perhaps ethically justified? This expert panel will explore the issue from a variety of perspectives including technology development, clinical research, bioethics, international law and policy, and the lived experience of patients and families.

THURSDAY, May 10—7:30 PM

**SESSION 8** CANCER AND MEDICAL GENOMICS

**Chairpersons:** **Trey Ideker,** University of California, San Diego
**Sharon Plon,** Baylor College of Medicine, Houston, Texas

**Decoding patient genomes through the hierarchical pathway architecture of the cancer cell**
Trey Ideker.
Presenter affiliation: University of California-San Diego, La Jolla, California. 197

**Signatures of complex structural variation across thousands of cancer whole genomes**
Kevin M. Hadi, Xiaotong Yao, Evan Biederstedt, Mahmoud Ghandi, Marcin Imielinski.
Presenter affiliation: Weill Cornell Medicine, New York, New York; New York Genome Center, New York, New York. 204

FRIDAY, May 11—9:00 AM

**SESSION 9** COMPUTATIONAL GENOMICS

**Chairpersons:** **Oliver Stegle,** EMBL/EBI, Hinxton, United Kingdom
**Christina Leslie,** Memorial Sloan-Kettering Cancer Center, New York, New York

**Methods for the joint analysis of high-dimensional traits and sample substructure in human cohorts**
Oliver Stegle.
Presenter affiliation: European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany; European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom. 205

**Dynamic effects of genetic variation on gene expression during cellular differentiation**
Benjamin Strober, Reem Elorbany, Katherine Rhodes, Nirmal Krishnan, David Knowles, Jonathan Pritchard, Yoav Gilad, Alexis Battle.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 206

**Population scale single cell sequencing to reveal context specific effects of SLE variants**
Meena Subramaniam, Lenka Maliskova, Nadav Rappoport, Cristina Lanata, Lindsey Criswell, Noah Zaitlen, Jimmie Ye.
Presenter affiliation: UCSF, San Francisco, California. 207

**Information theory analysis of ATAC-seq data predicts local chromatin kinetics and reveals novel aspects of gene regulation and genome organization**
Ricardo D'Oliveira Albanus, John Hensley, Yasuhiro Kyono, Jacob Kitzman, Stephen Parker.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 208

**Prediction of B-cell acute lymphoblastic leukemia subtypes from the expression of long noncoding RNAs and protein coding genes within common topologically associated domains**
Conor Nodzak, Gabrielle Centoducatte, J. Andrés Yunes, Xinghua Shi.
Presenter affiliation: University of North Carolina at Charlotte, Charlotte, North Carolina. 214

**Merging gene annotations enables high-resolution cell type identification**
Jim Notwell, Thomas Portmann.
Presenter affiliation: Circuit Therapeutics, Menlo Park, California. 215

**ADmiRE—Annotation of microRNA sequence variation across human population and adult cancer datasets**
Ninad Oak, Rajarshi Ghosh, Kuan-lin Huang, Deborah I. Ritter, Li Ding, Sharon E. Plon.
Presenter affiliation: Baylor College of Medicine, Houston, Texas; Texas Children's Hospital, Houston, Texas. 216

**Artificial genome rearrangement system using a restriction enzyme**
Arisa Oda, Takahiro Nakamura, Nobuhiko Muramoto, Hidenori Tanaka, Kazuto Kugou, Kunihiro Ohta.
Presenter affiliation: The University of Tokyo, Tokyo, Japan. 217

**The Rare Genomes Project—Improving our ability to diagnose rare genetic conditions through a nationwide partnership with families**
Anne O'Donnell-Luria, Melanie O'Leary, Mekdes Getaneh, Idara Ndon, Clara Williamson, Jaime Chang, Katherine Blakeslee, Julia Goodrich, Monica Wojcik, Nadya Lopez Zalba, Anzu Hakone, Jennifer Hendry Lapan, Esme Baker, Moran Cabili, Samantha Baxter, Ben Weisburd, Heidi Rehm, Daniel MacArthur.
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Boston Children's Hospital, Boston, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts; Harvard Medical School, Boston,. 218

**Biological factors strongly impact cell type abundances in human tissues**
Meritxell Oliva, Sarah Kim-Hellmuth, François Aguet, GTEx Consortium, Barbara E. Stranger.
Presenter affiliation: University of Chicago, Chicago, Illinois. 219

**Iterative reanalysis using novel genomic research tools improves clinical exome diagnostic yield in complex undiagnosed disease cases**
Matt Velinder, John C. Carey, Lorenzo D. Botto, Ryan Layer, Brent Pedersen, Andrew Farrell, Ashley Andrews, Pinar Bayrak-Toydemir, Rong Mao, Aaron Quinlan, Gabor T. Marth.
Presenter affiliation: USTAR Center for Genetic Discovery, Salt Lake City, Utah. 273

**T2D disease status alters the effects of genetic variation on molecular phenotypes—A DIRECT study**
A Viñuela, J Fernandez, A Kurbasic, MG Hong, S Sharma, C Brorsson, J Adamski, JM Schwenk, ER Pearson, S Brunak, PW Franks, MI McCarthy, ET Dermitzakis, IMI DIRECT consortium.
Presenter affiliation: University of Geneva, Geneva, Switzerland. 274

**Trans-eQTL meta-analysis in more than 31,000 blood samples associates downstream genes and pathways with polygenic risk**
Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Tõnu Esko, Lude Franke.
Presenter affiliation: University Medical Center Groningen, Groningen, Netherlands. 275

**Disruptions to RNA processing and gene regulation as convergent mechanisms associated with mutation to *CHD8***
A. Ayanna Wade, Kenneth Lim, Rinaldo Catta-Preta, Alex S. Nord.
Presenter affiliation: University of California, Davis, Davis, California. 276

**Dissecting transcriptional regulation by machine learning**
Hai Wang, Maria K. Mejía-Guerra, Karl A. Kremling, Guillaume Ramstein, Ravi Valluru, Edward S. Buckler, Jacob D. Washburn.
Presenter affiliation: Cornell University, Ithaca, New York; Chinese Academy of Agricultural Sciences, Beijing, China. 277

**Hybrid *de novo* assembly of *Branchiostoma belcheri Beihai* amphioxus genome**
Ming-Qiang Wang, Kevin Yi Yang, Junyuan Chen, Bingyu Mao, Stephen Kwok-Wing Tsui.
Presenter affiliation: The Chinese University of Hong Kong, Hong Kong SAR, China. 278

**Integration and assembly of extant sequencing technologies to enable complete structural variation discovery and phasing**
Xuefang Zhao, Mark J. Chaisson, Ashley D. Sanders, Human Genome Structural Variation Consortium.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan; Massachusetts General Hospital, Center for Genomic Medicine, Boston, Massachusetts. 292

**Identification of a beneficial, complex, genetic rearrangement in *rcan-1*, an ortholog of the down syndrome critical region protein 1 (DSCR1/RCAN1), in a laboratory strain of *C. elegans***
Yuehui Zhao, Patrick McGrath.
Presenter affiliation: Georgia Institute of Technology, Atlanta, Georgia. 293

**Mega-analysis of odds ratio (MegaOR)—A convergent method for deep understanding the genetic evidence in schizophrenia**
Peilin Jia, Zhongming Zhao.
Presenter affiliation: University of Texas Health Science Center at Houston, Houston, Texas; Vanderbilt University, Nashville, Tennessee. 294

FRIDAY, May 11—4:30 PM

**GUEST SPEAKERS**

**Wendy Bickmore**
MRC, University of Edinburgh

**David Page**
Whitehead Institute/MIT/HHMI

FRIDAY, May 11

**BANQUET**

Cocktails 6:00 PM Dinner 6:45 PM

**SESSION 11** COMPLEX TRAITS AND MICROBIOME

**Chairpersons:** **Jeff Barrett,** Wellcome Sanger Institute, Hinxton, United Kingdom
**Ami Bhatt,** Stanford University, Palo Alto, California

# WRITING GENOMES

Jef D Boeke
NYU Langone Health, Inst Systems Genetics, New York, NY

Rapid advances in DNA synthesis techniques have made it possible to engineer diverse genomic elements, pathways, and whole genomes, providing new insights into design and analysis of systems. The synthetic yeast genome project, Sc2.0, is well on its way with six synthetic Saccharomyces cerevisiae chromosomes completed by a global team. The synthetic genome features several systemic modifications, including TAG/TAA stop-codon swaps, deletion of subtelomeric regions, introns, tRNA genes, transposons and silent mating loci. Telomeres were replaced by a systematic universal telomere cap. Strategically placed loxPsym sites enable genome restructuring using an inducible evolution system termed SCRaMbLE (Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution). SCRaMbLE can be used as a novel method of mutagenesis, capable of generating millions of derived variants with predictable structures leading to complex genotypes and a variety of phenotypes. The fully synthetic yeast genome opens the door to a new type of combinatorial genetics based on variations in gene content and copy number, rather than base changes. Remarkably, the 3D structure of synthetic and native chromosomes are very similar despite the substantial changes introduced. We also describe supernumerary designer "neochromosomes" that add new functionalities to cells such as humanized pathways and complexes, such as the humanization of metabolic pathways and even chromatin. Finally, we have automated many steps in our big DNA synthesis pipeline, opening the door to massively parallel big DNA assembly, including assembly of human genomic regions of 100 kb and up. We are developing robust methods to deliver such segments. Discussions are underway to write even bigger genomes.

Dymond et al. 2011 Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. Nature, 477:471-6.

Annaluru et al. 2014 Total synthesis of a functional designer eukaryotic chromosome. Science. 344:55-8.

Richardson et al. 2017 Design of a synthetic yeast genome, Sc2.0. Science 355:1040-1044.

Mitchell et al. 2017 Synthesis, debugging and effects of synthetic chromosome consolidation: synVI and beyond. Science, 355 pii: eaaf4831.

Mercy et al. 2017 3D organization of synthetic and scrambled chromosomes. Science, 355 pii: eaaf4597.

# ACCURATE FUNCTIONAL CLASSIFICATION OF THOUSANDS OF BRCA1 VARIANTS WITH SATURATION GENOME EDITING

Gregory M Findlay[1], Riza M Daza[1], Beth Martin[1], Melissa D Zhang[1], Anh P Leith[1], Molly Gasperini[1], Joseph D Janizek[1], Lea M Starita[1,2], Jay Shendure[1,2,3]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]Brotman Baty Institute for Presicion Medicine, Seattle, WA, [3]Howard Hughes Medical Institute, Seattle, WA

Variants of uncertain significance (VUS) fundamentally limit the utility of genetic information in a clinical setting. The challenge of VUS is epitomized by *BRCA1*, a tumor suppressor gene integral to DNA repair and genomic stability. Germline *BRCA1* loss-of-function variants predispose women to early-onset breast and ovarian cancers. Although *BRCA1* has been sequenced in millions of people, the risk associated with most newly observed variants cannot be definitively assigned. Data sharing attenuates this problem but it is unlikely to solve it, as most newly observed variants are exceedingly rare. In lieu of genetic evidence, experimental approaches can be used to functionally characterize VUS. However, to date, functional studies of *BRCA1* VUS have been conducted in a post hoc, piecemeal fashion and performed outside of the gene's endogenous context.

Here, we employ saturation genome editing using CRISPR/Cas9 to assay 96.5% of all possible single nucleotide variants (SNVs) in 13 exons that encode the RING and BRCT domains of BRCA1. Our highly optimized, multiplex assay measures variants' effects on cellular fitness in a haploid human cell line whose survival is dependent on intact *BRCA1* function. The function scores we obtain for 3,893 SNVs are bimodally distributed and almost perfectly concordant with assessments of pathogenicity in the ClinVar database (sensitivity = 97%; specificity = 98%). Of the 256 VUS assayed, 25.0% score as non-functional, as do 49.2% of 122 SNVs with conflicting interpretations of pathogenicity, and 15.9% of the 3,140 SNVs absent from ClinVar. Sequence-function maps enhanced by parallel measurements of variant effects on mRNA levels reveal mechanisms by which SNVs disrupt gene function. Hundreds of missense SNVs detrimental to the protein's function are identified, as well as dozens of exonic and intronic SNVs that compromise *BRCA1* function by disrupting splicing. Comparing SNV function scores to computational predictions and to allele frequencies in population sequencing data yields additional insights.

We predict that our function scores will be directly useful for the clinical interpretation of cancer risk based on *BRCA1* sequencing. Furthermore, we propose that this paradigm can be extended to overcome the challenge of VUS in other genes in which genetic variation is clinically actionable.

# EXTRACHROMOSOMAL CIRCULAR DNA (eccDNA) IS A POSSIBLE MEDIATOR OF CHROMOSOMAL POLYMORPHISM AT MULTIPLE LOCI.

Stephen D Levene[1,2,3], Massa J Shoura[4], Andrew Z Fire[4,5]

[1]University of Texas at Dallas, Bioengineering, Richardson, TX, [2]University of Texas at Dallas, Biological Sciences, Richardson, TX, [3]University of Texas at Dallas, Physics, Richardson, TX, [4]Stanford University School of Medicine, Pathology, Stanford, CA, [5]Stanford University School of Medicine, Genetics, Stanford, CA

The existence of endogenous circular-DNA elements, termed extrachromosomal circular DNA (eccDNA), has been established in a wide range of organisms, from plants to humans, for more than thirty years. DNA-sequencing methods have progressed to the point that detailed (if not quite complete) pictures of linear genomes can be assembled in 1-D and 3-D. Despite the progress that has and will continue to come from understanding linear genomes, the genomics community would be missing a substantial opportunity by ignoring the diversification in coding capacity that circular-DNA formation can provide. We have taken a whole-genome approach to mapping endogenous circular DNA in a single organism (C. elegans), human cell lines, and human tissue samples; we denote this component of the genome the “circulome”. The information gained from extensive eccDNA profiling of diseased and normal cell lines provides insights into critical missing links between the genome and transcriptome and has the potential to advance development of biomarkers and tools having applications in molecular diagnosis. High-throughput analysis of endogenous circular DNAs identified dynamic and highly variable loci, which might otherwise defy characterization. We show that using eccDNA as an indicator of stability may illuminate a set of mechanisms that regulate functional genomic rearrangements and thereby provide new biological approaches for studying genome dynamics.

# RESURRECTION OF HISTONE H3 K27 METHYLATION IN BREWER'S YEAST BY HUMAN PRC2 AND PLANT ATXR6

David M Truong, Jef D Boeke

New York University Langone Health, Institute for Systems Genetics, New York, NY

The human Polycomb-Repressive Complex 2 (PRC2) methylates Histone H3 K27 (H3K27), a histone modification that marks transcriptionally repressed heterochromatin. It is a multisubunit protein-complex, critically involved in cell differentiation and maintenance of cell identity, by stabilizing transcriptional programs. Misregulation of this mark and mutations to the histone target or PRC2 subunits, are enriched in numerous types of cancers, and are enticing 'druggable' candidates. However, functional genomic studies in higher eukaryotes are often limited by the numerous proteins, splice isoforms, and varying compositions which make-up histone methylation complexes. Studies of any single gene reveal only partial information. Multiple knockouts lead to cell dysfunction and pleiotropy, making it tough to separate out phenotypic effects. Unique amongst eukaryotes, *Saccharomyces cerevisiae* (Brewer's Yeast) has lost the machinery for generating histone modifications associated with repressive chromatin, H3K27 and H3K9 methylation, and H2AK119ub. As a fast growing organism with a small genome and advanced genetic tools, it's an intriguing prospect for use as a "blank-slate" by which to reconstruct these pathways from the ground-up. We recently "humanized" the core histones (nucleosome) of *S. cerevisiae* (Truong & Boeke, 2017). We have now generated an accessory human "neochromosome" for yeast, that contains 5 human PRC2 proteins (EZH2, EED, SUZ12, AEBP2, RBBP4). Alone, this complex does not methylate yeast nucleosomes. However, we find that a plant protein, ATXR6, generates robust H3K27 mono-methylation. By combining these, we observe PRC2-dependent di-methylation of H3K27 by Western Blot. Thus, we conclude that an H3K27 mono-methyl may be sufficient to recruit PRC2 to generate H3K27 higher-order methylation states. This suggests that another protein or complex may serve the role of initiating H3K27 mono-methylation, and highlights an under-recognized mechanism for PRC2 recruitment. These studies set the stage for eventually determining the causal sequence of events that establish repressive chromatin, *de novo*.

# ADVANCES IN GENOME EDITING TECHNOLOGIES

Feng Zhang[1,2]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Department of Biological Engineering, Cambridge, MA

In the five years since the initial demonstration of mammalian genome editing using the Cas9 enzyme, the molecular scissors of the microbial adaptive immune CRISPR system, a number of advancements in genome editing technology have been made with astounding speed. Cas9 has been leveraged for a range of genome manipulation tools, including gene activation and repression as well as modulation of chromatin and DNA modifications. Additional DNA-targeting Cas enzymes have been discovered, broadening the possible targeting space within the human genome and offering greater activity in other species. More recently, RNA-targeting Cas enzymes have been discovered, expanding CRISPR-mediated technologies into the realm of the transcriptome modulation. We have characterized a number of these novel enzymes, known as Cas13, and identified orthologs that work in mammalian cells with high activity and specificity. We have shown that Cas13 can be used to knock down endogenous transcripts as well as serve as a programmable RNA-binding platform. Additionally, we engineered a fusion between Cas13 and the adenine deaminase ADAR to achieve RNA Editing for Precise A-to-I Replacement (REPAIR). We showed that REPAIR has the potential to correct single-base pathogenic mutations at the transcriptional level. REPAIR may be a powerful therapeutic for diseases that affect cell types and tissues not amenable to DNA-based gene therapies, such as neurons and other post-mitotic cells. We are continuing to explore microbial diversity to find new enzymes and systems that can be adapted for use as molecular biology tools and novel therapeutics.

# CRISPR-QTL MAPPING AS A GENOME-WIDE ASSOCIATION FRAMEWORK FOR CELLULAR GENETIC SCREENS

Molly Gasperini[1], Andrew Hill[1], José L McFaline Figueroa[1], Beth Martin[1], Cole Trapnell[1], Nadav Ahituv[2], Jay Shendure[1,3,4]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]University of California San Francisco, Bioengineering and Therapeutic Sciences, San Francisco, CA, [3]University of Washington, Brotman Baty Institute for Precision Medicine, Seattle, WA, [4]University of Washington, Howard Hughes Medical Institute, Seattle, WA

Expression quantitative trait locus (eQTL) and genome-wide association studies (GWAS) are powerful paradigms for mapping the determinants of gene expression and organismal phenotypes, respectively. However, eQTL mapping and GWAS are limited in scope (to naturally occurring, common genetic variants) and resolution (by linkage disequilibrium). Here, we present CRISPR-QTL mapping, a framework in which large numbers of CRISPR/Cas9 perturbations are introduced to each cell on an isogenic background, followed by single-cell RNA-seq (scRNA-seq). CRISPR-QTL mapping is analogous to conventional human eQTL studies, but with individual humans replaced by individual cells; variants replaced by 'unlinked' combinations of guide RNA-programmed perturbations per cell; and tissue-level RNA-seq of many individuals replaced by scRNA-seq of many cells. As a proof-of-concept, we applied CRISPR-QTL mapping to evaluate 1,119 candidate enhancers with no strong *a priori* hypothesis as to their target gene(s). Perturbations were made by a nuclease-dead Cas9 tethered to KRAB, and introduced at a mean 'allele frequency' of 1.1% into a population of 47,650 profiled human K562 cells. We tested for differential expression of all genes within 1 megabase of each candidate enhancer, effectively evaluating 17,584 potential enhancer-target gene relationships within a single experiment. At an empirical FDR of 10%, we identify 128 *cis* CRISPR-QTLs (11%) whose targeting resulted in downregulation of 105 nearby genes. CRISPR-QTLs were strongly enriched for proximity to their target genes (median 44.3 kilobases) and the strength of H3K27ac, p300, and lineage-specific transcription factor ChIP-seq peaks. Our results establish the power of the eQTL mapping paradigm as applied to programmed variation in populations of cells, rather than natural variation in populations of individuals. We anticipate that CRISPR-QTL mapping will facilitate the comprehensive elucidation of the *cis*-regulatory architecture of the human genome.

functional genetic variants revealed by massively parallel precise genome editing

Eilon Sharon*[1,2], Shi-An A Chen*[1], Neil M Khosla*[1], Justin D Smith[2], Jonathan K Pritchard[1,2,3], Hunter B Fraser[1]

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

A major challenge in genetics is to determine which sequence differences between species, populations, and individuals drive phenotypic and fitness differences. However, current methods of genetic mapping generally have limited resolution. To address this gap, we developed a CRISPR-Cas9-based high-throughput genome editing approach that can introduce thousands of specific genetic variants in a single experiment. This approach enabled us to study the fitness consequences of 16,006 SNPs and indels that differ between the yeast strains RM and BY. We identified 572 variants with significant fitness differences in glucose media; many of these have large effect sizes, with 171 estimated to affect fitness by >1%. Most of the significant variants are regulatory: they are highly enriched in promoters, particularly in transcription factor binding sites, while only 19.2% affect amino acid sequences. Strikingly, nearby variants nearly always favor the same strain, suggesting that lineage-specific selection on particular genes is often driven by multiple variants. Finally, ribosomal gene promoters are highly enriched for variants whose RM alleles increase fitness, suggesting polygenic adaptation of the ribosome. In sum, our genome editing approach reveals the genetic architecture of fitness variation at single-base resolution, and could be adapted to measure the effects of genome-wide genetic variation in any screen for cell survival or cell-sortable markers.

# CRISPR-SURF: EXPLORATORY AND INTERACTIVE SOFTWARE FOR ANALYZING CRISPR-BASED TILING SCREENS

Jonathan Y Hsu[1,2,3], Charles P Fulco[3], Mitchel Cole[3,4], Matthew C Canver[2,3], Danilo Pellin[3,4], Falak Sher[3,4], Rick Farouni[2,3], Kendell Clement[2,3], Luca Biasco[3,4], Jesse M Engreitz[3], Eric S Lander[3], J. Keith Joung[2,3], Daniel E Bauer[3,4], Luca Pinello[2,3]

[1]MIT, Biological Engineering, Boston, MA, [2]MGH, Molecular Pathology, Boston, MA, [3]Broad Institute, Boston, MA, [4]Boston Children's Hospital, Hematology/Oncology, Boston, MA

The advent of programmable genome editing using CRISPR-based technologies has allowed for high-throughput functional interrogation of non-coding elements in the genome. Fine-mapping can be achieved by densely tiling single guide RNAs (sgRNAs) across a non-coding region of interest, where each sgRNA enables linkage of a unique genomic location to an observable phenotype. These tiled sgRNAs can be used with CRISPR nucleases, CRISPRi, or CRISPRa to introduce mutations, repress a target gene, or activate a target gene, respectively. The ability to systematically and quantitatively assess causal links between non-coding regulatory elements and their respective target genes with CRISPR-based tiling screens will have substantial implications for elucidating the complex regulatory architecture underlying gene expression.

Although it is now possible for investigators to experimentally tile a genomic region of interest with genetic and epigenetic perturbations, no publicly-available, open-source software solution currently exists to model the spatial interdependencies between proximal sgRNAs, which is inherent and fundamental for the interpretation of CRISPR-based tiling screens. This lack of access to analysis software is a bottleneck for wider access to CRISPR screening technology, making it more difficult for investigators to analyze and interpret data from their experiments. To address this issue, we developed CRISPR-SURF (Screening Uncharacterized Region Function), an intuitive web-based and desktop application that will enable any user to analyze data from any type of CRISPR-based tiling screen.

We performed two CRISPR-based tiling screens using CRISPRi (dCas9-KRAB) and guide saturating mutagenesis (SpCas9) to highlight the generalizability and robustness of the CRISPR-SURF pipeline, and also provide a general strategy to perform hierarchal enhancer mapping and dissection. Using CRISPR-SURF, we identified non-coding regions regulating the expression of transcription factor BCL11A, a known potent repressor of fetal hemoglobin (HbF) levels. In addition, sgRNA and sequencing read count down-sampling were performed to provide guidelines for efficient large-scale tiling screens.

CRISPR-SURF offers an exploratory and interactive environment that will better enable biologists and bioinformaticians to explore and intuitively analyze their CRISPR-based tiling screen data.

# SEQUENCE-BASED APPROACHES UTILIZING COMPLETE MODERN AND ANCIENT GENOMES TO INVESTIGATE EARLY HUMAN HISTORY

Mattias Jakobsson

Uppsala University, Department of Organismal Biology, Uppsala, Sweden

Complete human genomes are becoming available from a large number of individuals and populations across the world, including from archaic and prehistoric individuals. Genetic studies consistently show that southern African Khoe-San populations carry more unique variants and more divergent lineages than any other living groups, and that they encompassing the deepest divergence among modern-day humans. I will discuss traditional and new population genomics approaches that benefit from mutation models, molecular clocks, and full genome sequence data to infer demographic parameters. I will exemplify these approaches by investigating the deep history of modern humans and discuss deep population structure, archaic admixture and links to population sizes. I will finally discuss recent genome sequencing of ancient individuals from southern Africa that show several recent migrations and admixture events into all modern-day Khoe-San groups. Based on the ancient Stone-Age genome of the southern African Ballito Bay boy, that was not affected by these admixture events, the deepest human population divergence time is estimated to between 350,000 and 260,000 years ago. This estimate increases the deepest divergence amongst modern humans, coinciding with anatomical developments of archaic humans into modern humans as represented in the local fossil record, and suggest that modern humans emerged around 300,000 years ago.

# WHOLE GENOME SEQUENCE REVEALS SELECTION FOR MUSCLE, CARDIOVASCULAR AND NEURONAL GENES IN SPORT HUNTING DOGS

Jaemin Kim, Falina J Williams, Dayna L Dreger, Jocelyn Plassais, Brian W Davis, Elaine A Ostrander

Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

Modern dogs are distinguished among domesticated species by the vast breadth of phenotypic variation produced by strong and consistent human-driven selective pressure. The resulting breeds reflect the development of closed populations with well-defined physical and behavioral attributes. The sport hunting dog group has long been employed in assistance to hunters, reflecting strong behavioral pressures to locate and pursue quarry over great distances and variable terrain. Comparison of whole genome sequences data between sport hunting and terrier breeds, groups at the ends of continuum in both form and function, reveals that genes underlying cardiovascular, muscular and neuronal functions are under greatest selection in sport hunting breeds, including *ADRB1*, *TRPM3*, *RYR3*, *UTRN*, *ASIC3*, and *ROBO1*. We also identified an allele of *TRPM3* that was significantly associated with increased racing speed in whippets, accounting for 11.6% of total variance in racing performance. Finally, we observed a significant association of *ROBO1* with breed-specific accomplishments in competitive obstacle course events. These results provide strong evidence that sport hunting breeds have adapted to their occupations by improving endurance, cardiac function, blood flow, and cognitive performance, demonstrating how strong behavioral selection alters physiology to create breeds with distinct capabilities.

# WIDESPREAD DIFFERENCES IN THE MUTATION SPECTRUM OF X AND AUTOSOMES: CAUSES AND CONSEQUENCES

Ipsita Agarwal[1], Molly Przeworski[1,2]

[1]Columbia University, Department of Biological Sciences, New York, NY,
[2]Columbia University, Department of Systems Biology, New York, NY

The number of de novo mutations carried by egg or sperm depends on the balance of error and repair rates across stages of development. Because human male and female germ cells differ in many of these regards, sex differences in both the number and type of mutations are expected. Yet differences between paternal and maternal mutation spectra have proven difficult to identify, even in a recent data set of 100,000 human de novo mutations. We instead use a large number of genome sequences (~14,000 chromosomes), in which the tens of millions of rare polymorphisms should faithfully reflect de novo mutations. We compare patterns on the X chromosome, which disproportionately reflects the female germline, to the autosomes and the Y chromosome. To disentangle the effects of exposure to the male and female germlines from those of X-inactivation and other idiosyncratic features of the X chromosome, we distinguish pseudo-autosomal regions and regions on the X that do or do not escape inactivation. Considering the 96 mutation types defined by the 5' and 3' context of a single nucleotide change, we find ubiquitous differences in the mutation spectrum across regions of the X chromosome and between X and autosome. Contrasting the different compartments of the X, we show that X-autosome differences in the mutation spectrum reflect a combination of sex differences in the germline, methylation and recombination rates. These findings imply that sex-specific life history traits interact with the biochemical processes of mutation to generate the mutation spectrum seen in variation data. As a consequence, changes in life history traits of males and females will not only shape X-autosome differences, but will lead to the evolution of the mutation spectrum of autosomes.

# THE GENETIC ARCHITECTURE OF HUMAN DNA REPLICATION ORIGIN ACTIVITY

Qiliang Ding[1], Xiang Zhu[2], Joyce Hsiao[3], Florian T Merkle[4], Robert E Handsaker[5,6], Sulagna Ghosh[5,6,7], Kevin Eggan[5,7], Steven A McCarroll[5,6], Matthew Stephens[3], Yoav Gilad[3], Andrew G Clark[1], Amnon Koren[1]

[1]Cornell University, Department of Molecular Biology and Genetics, Ithaca, NY, [2]Stanford University, Department of Statistics, Palo Alto, CA, [3]University of Chicago, Department of Human Genetics, Chicago, IL, [4]University of Cambridge, Medical Research Council Institute of Metabolic Science, Cambridge, United Kingdom, [5]Broad Institute of MIT and Harvard, Stanley Center for Psychiatric Research, Cambridge, MA, [6]Harvard Medical School, Department of Genetics, Boston, MA, [7]Harvard University, Department of Stem Cell and Regenerative Biology, Cambridge, MA

Despite having detailed maps of functional elements in the human genome, one component is still critically missing: the locations of DNA replication origins. Furthermore, the activity of DNA replication origins interfaces with several genetic and epigenetic aspects of chromosome biology, yet we know very little about the determinants of origin activation timing and its molecular and phenotypic consequences. Here, we take a population genetics approach for mapping the locations and activity of replication origins in the human genome and for elucidating their genetic basis. We deep-sequenced the genomes of 108 proliferating human embryonic stem cell lines of European ancestry. Analysis of DNA copy number fluctuations along chromosomes resulted in high-resolution replication timing profiles that were highly consistent across individuals yet revealed clear variation at hundreds of genomic sites. By comparing DNA replication timing to sequence variation in these cell lines, we mapped 757 replication timing quantitative trait loci (rtQTLs) at a 10% false discovery rate. The majority of rtQTLs mapped to replication timing peaks, providing a highly accurate prediction of the locations of human replication origins and indicating that human replication origin activity is strongly influenced by sequence determinants in cis. Many replication origins were associated with several nearby but independent SNP haplotypes, suggesting that replication timing may be subject to complex, combinatorial regulation at the DNA sequence level. rtQTLs were enriched for DNase I hypersensitivity sites, active chromatin states including active transcription start sites and enhancers, and active chromatin marks (e.g. H4K12ac and H3K4me2). rtQTLs were further enriched for several transcription factor binding sites and motifs, notably POU5F1 (Oct4) and NANOG, two central pluripotency factors required for stem cell self-renewal. SNP alleles that stabilize the binding of these TFs were associated with earlier origin activity, pointing to a mechanism for human origin activation. Finally, we found several SNP alleles that have specifically evolved in Europeans and gave rise to new or more active replication origins, suggesting that DNA replication timing is an evolving trait with potentially important implications to human biology. Comprehensive mapping of genetic determinants of DNA replication timing in human populations holds the promise of uncovering molecular mechanisms of DNA replication regulation and their potential influence on the human genome and on human biology.

# NOVEL LOCI ASSOCIATED WITH SKIN PIGMENTATION IDENTIFIED IN AFRICAN POPULATIONS

Nicholas G Crawford[1], Derek Kelly[1], Matthew E Hansen[1], Marcia Holsbach Beltrame[1], Shaohua Fan[1], Shanna L Bowman[2], Ethan Jewett[3,4], Alessia Ranciaro[1], Michael Campbell[1,5], Yancy Lo[1], Yun S Song[3,4], Kevin M Brown[6], Michael S Marks[2], Stacie K Loftus[7], William J Pavan[7], Meredith Yeager[8], Stephen Chanock[8], Sarah A Tishkoff[1,9]

[1]Univ of Pennsylvania, Genetics, Philadelphia, PA, [2]CHOP, Pathology and Laboratory Medicine, Philadelphia, PA, [3]UC, Berkeley, EECS, Berkeley, CA, [4]UC Berkeley, Statistics, Berkeley, CA, [5]Howard University, Biology, Washington, DC, [6]NCI, NIH, Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, Bethesda, MD, [7]NHGRI, NIH, Genetic Disease Research Branch, Bethesda, MD, [8]NCI, NIH, Division of Cancer Epidemiology and Genetics, Bethesda, MD, [9]Unive of Pennsylvania, Biology, Philadelphia, PA

Despite the wide range of variation in skin pigmentation in Africans, little is known about its genetic basis. To investigate this question we performed a GWAS on pigmentation in 1,593 Africans from populations in Ethiopia, Tanzania, and Botswana. We identify significantly associated loci in or near SLC24A5, MFSD12, DDB1, TMEM138, and OCA2 and HERC2. Allele frequencies at these loci in global populations are strongly correlated with UV exposure. At SLC24A5 we find that a non-synonymous mutation associated with depigmentation in non-Africans was introduced into East Africa by gene flow, and subsequently rose to high frequency. At MFSD12, we identify novel variants that are strongly correlated with dark pigmentation in populations with Nilo-Saharan ancestry. Functional assays reveal that MFSD12 codes for a lysosomal protein that influences pigmentation in cultured melanocytes, zebrafish and mice. CRISPR knockouts of murine Mfsd12 display reduced pheomelanin pigmentation similar to the grizzled mouse mutant (gr/gr). Exome sequencing of gr/gr mice identified a 9 bp in-frame deletion in exon two of Mfsd12. Thus, using human GWAS data we were able to map a classic mouse pigmentation mutant. At DDB1/TMEM138, we identify mutations in melanocyte-specific regulatory regions associated with expression of UV response genes. Variants associated with light pigmentation at this locus show evidence of a selective sweep in Eurasians. At OCA2 and HERC2 we identify novel variants associated with pigmentation and at OCA2, the oculocutaneous albinism II gene, we find evidence for balancing selection maintaining alleles associated with both light and dark skin pigmentation. We observe at all loci that variants associated with dark pigmentation in African populations are identical by descent in southern Asian and Australo-Melanesian populations and did not arise due to convergent evolution. Further, the alleles associated with skin pigmentation at all loci but SLC24A5 are ancient, predating the origin of modern humans. The ancestral alleles at the majority of predicted causal SNPs are associated with light skin, raising the possibility that the ancestors of modern humans could have had relatively light skin color, as is observed in the San population today. This study sheds new light on the evolutionary history of pigmentation in humans.

# NONPARAMETRIC ESTIMATION OF ALLELE AGE FOR RARE VARIANTS IN POPULATION-SCALE SEQUENCING DATA

Patrick K Albers, Gil McVean

University of Oxford, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Oxford, United Kingdom

The age of rare genetic variants is potentially informative for dating connections between individuals and populations, detecting and inferring the effects of natural selection, and reconstructing demographic history. However, to date, estimates of allele age have been based on highly-parameterized models and typically use only a fraction of the information available in population-scale genomic data. Here, we introduce a non-parametric method for estimating allele age from whole genome-sequencing data that exploits haplotype structure around variants and combines information from both the mutation and recombination clocks across many pairs of individuals to obtain a composite posterior estimate. Through simulation, we show that our method produces accurate estimates of allele age (rank correlation between true and inferred age, $\rho = 0.89$) in scenarios of constant or exponentially growing populations with diverse patterns of migration. Our method is robust to realistic models of genotype error ($\rho = 0.85$) and scales to data sets of arbitrary size. We apply the method to variants below 5% frequency from the 1000 Genomes Project (1KGP) and estimated a median allele age of 500 generations (c. 15,000 years) for variants found exclusively in Europeans; 710 generations for those found exclusively in Americans; 830 and 1,100 generations respectively for variants exclusive to East and South Asian populations; and 2,600 generations for those exclusive to Africans. In contrast, cosmopolitan variants of the same frequency were estimated to have a median allele age of 3,100 generations. We applied our method to all SNPs in the 1KGP dataset, release a genome-wide map of the estimated age of alleles in the human genome, and demonstrate how age-stratified allele sharing can be used to provide a rich description of an individual's genealogical history.

# POLYGENIC ADAPTATION IN RESPONSE TO A SUDDEN CHANGE IN THE ENVIRONMENT

Laura K Hayward[1], Guy Sella[2,3,4]

[1]Columbia University, Mathematics, New York, NY, [2]Columbia University, Biological Sciences, New York, NY, [3]Columbia University, Center for Computational Biology and Bioinformatics, New York, NY, [4]Columbia University, Program for Mathematical Genomics, New York, NY

Multiple lines of evidence suggest that adaptive changes to the genome often involve traits that are highly polygenic. Yet the response to selection on polygenic traits is poorly understood, limiting our ability to look for and interpret signatures of polygenic adaptation in population genetics data. To address this question, we study the phenotypic and genetic response to selection on a quantitative trait under stabilizing selection, after a change in environment induces a change in the trait's optimal value. We find that the phenotypic response displays one of two qualitative behaviors, depending on population genetic parameters. When mutations affecting the trait have small effect sizes and arise at high rates, the genetic variance in the population remains approximately constant and the mean phenotype approaches the new optimum at an exponential rate that depends on this variance (Lande, 1976). But when mutations have large effect sizes or arise at lower rates, the phenotypic response is more complex. In turn, changes in allele frequencies underlying the phenotypic response exhibit vastly different short and long-term behaviors. The rapid approach to the new optimum is mostly driven by alleles with large effect sizes. However, over longer time scales, the contribution of these alleles to the change in mean phenotype almost entirely disappears, and the shift in phenotype is taken over by alleles with relatively small effect sizes, some of which eventually fix. One implication of these dynamics is that fixations resulting from polygenic adaptation should have negligible effects on neutral diversity levels at linked sites.

# THE IMPACT OF NEANDERTHAL ANCESTRY ON HUMAN PHENOTYPES

Christopher Robles*[1], Andrea Ganna*[5,6,7,4], Alexander Gusev*[9,10], Arun Durvasula[1], Steven Gazal[3,4], David Reich[8,4,9], Sriram Sankararaman[1,2]

[1]UCLA, Department of Human Genetics, Los Angeles, CA, [2]UCLA, Computer Science Department, Los Angeles, CA, [3]Harvard T.H. Chan School of Public Health, Department of Epidemiology, Boston, MA, [4]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA, [5]Massachusetts General Hospital, Department of Genomic Medicine, Boston, MA, [6]Broad Institute, Stanley Center for Psychiatric Research, Cambridge, MA, [7]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Stockholm, Sweden, [8]Harvard Medical School, Department of Genetics, Boston, MA, [9]Harvard Medical School, Dana Farber Cancer Institute, Boston, MA, [10]Brigham and Women's Hospital, Division of Genetics, Boston, MA

While multiple instances of admixture between archaic and modern humans have now been documented, the biological consequences of these admixture events on modern humans are not fully understood. We analyzed about 80,000 introgressed Neanderthal alleles across 107 distinct phenotypes measured in up to 495,000 people of European ancestry in the UK Biobank dataset, taking advantage of more than 6,000 single nucleotide polymorphisms that we specifically added to the UK Biobank array prior to genotyping to study the impact of these Neanderthal alleles. We discovered 1158 independent associations of introgressed Neanderthal alleles with 75 phenotypes. We developed rigorous methodology to assess whether Neanderthal ancestry is over- or underrepresented in modulating phenotypes compared to random genetic variation, appropriately controlling for the frequencies and allelic age distribution of such variants. We find that the contribution of Neanderthal alleles to phenotypic variation is significantly depleted in the great majority of the phenotypes examined which we show is consistent with the observation that in general, Neanderthal mutations are on average older and that natural selection has acted to remove Neanderthal mutations since introgression. However, Neanderthal alleles were significantly overrepresented in contribution to variation in a handful of traits including male balding, chronotype, wheat intake, lung capacity, and eye-related phenotypes. Most notably, we document directional effects for Neanderthal-derived mutations that affect a number of traits including propensity for baldness, later age of female puberty, high lung capacity, number of children, alertness in the morning, and neuroticism. These analyses highlight aspects of modern human biology that have been influenced by Neanderthal introgression during the last fifty thousand years of non-African history when evidence for symbolic behavior and innovation become evident in the archaeological record.

*-contributed equally

# SENSITIVE DETECTION OF LOW FREQUENCY SINGLE NUCLEOTIDE VARIANTS FROM AMPLICON AND CAPTURE SEQUENCING DATA WITH LEUCIPPUS

Nikolaos Vasmatzis[1], Jamie N Bakkum-Gamez[2], Flora M Vaccarino[3,4], Alexej Abyzov[1]

[1]Mayo Clinic, Health Sciences Research, Rochester, MN, [2]Mayo Clinic, Department of Obstetrics and Gynecology, Rochester, MN, [3]Yale University, Child Study Center, New Haven, CT, [4]Yale University, Department of Neuroscience, New Haven, CT

Based on existing evidence, the leading paradigm in the field of medical genomics is that the primary cause of cancer is genomic alteration in somatic cells. Such alterations include Copy Number Alterations (CNA), single nucleotide variants (SNVs), small insertions and deletions (indels), and chromosome fusions or translocations. SNV is the most common and best understood alteration leading to cancer. Sensitive detection (i.e., when a variant is present in a minority of analyzed cells) of cancer-associated SNVs can be leveraged in cancer screening, diagnostics, and therapeutics. However, existing approaches based on Next Generation Sequencing (NGS) data focus on detecting variants that are present in all or most cells sequenced per assay. We therefore developed an analytical approach to sensitive detection of somatic variants from deep sequencing of captured or amplified genomic regions. This novel approach is based on reducing multiple hypotheses testing by interrogating only the sites of interest, which are a small fraction of all sites. Experiments are designed so that paired reads overlap, which allows estimating sequencing error for each pair of reads and discard pairs with excessing mismatches. Furthermore, read sequencing errors are corrected at the overlapping 3'-ends, and the overall sequencing error rate is empirically estimated. We implemented this approach in Leucippus software (freely available at GitHub) and validated it with an orthogonal technique, digital droplet PCR, demonstrating that SNVs with allele frequency as low as 0.1% can be detected with Leucippus. We then applied the approach to DNA from endometrial brushings collected from endometrial cancer and benign endometrium and detected 44 protein-truncating SNVs in genes known to be associated with this cancer. In conclusion, Leucippus can be used in both research and clinical settings for sensitive detection of somatic SNVs.

# DEFINING THE REGULATORY GRAMMAR OF HUMAN DENDRITIC CELLS ACTIVATION

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

[1]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, [2]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, [3]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA, [4]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA

Changes in gene expression are mediated by a complex regulatory network, which 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.

We profiled the chromatin landscape of hDCs at six time-points following LPS stimulation with both ATAC-seq and ChIP-seq of few selected histone marks. We identified 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. We take a supervised learning approach to detect the short DNA sequences and specific TF binding events which regulate gene induction following LPS stimulation. By applying our model on data from several time points and several donors, we will be able to detect significant changes to the regulatory landscape both across time and across individuals and allow us to uncover the grammar of transcriptional regulation in hDCs.

The large expression changes as well as the various temporal transcriptional responses makes activation of hDCs an ideal system to understand general mechanisms of gene regulation and gain a better grasp of the human immune system. Moreover, our computational method is easily generalizable and, combined with the experimental platform, could be applied to study many other biological systems.

# OPEN READING FRAME FILTERING OF HELICOBACTER PYLORI GENOME

Nayra Al-Thani, Simeon Andrews, Joel Malek

Weill Cornell Medicine Qatar, Research, Doha, Qatar

Studying protein-protein interactions (PPIs) is essential for understanding normal and pathological physiology within a cell. Understanding PPIs helps us to understand disease processes such as cancer and their mechanisms. Ideally, we would like to understand and study the entire network of protein interactions, which is referred to as the "interactome". An interactome defines the full network of PPIs that take place within a cell. DNA-sequencing technologies have long been robust, and particularly in the past decade "next-generation" sequencing has revolutionized our ability to quickly and accurately sequence vast amounts of genetic material.

Initial experiments with the two-hybrid systems generally employed full-length genes. The use of fragments rather than full-length libraries is perhaps counterintuitive, as we know that many protein structures and interactions are impossible with fragments, and might be expected to result in a high rate of false negatives. The use of fragments allows localization of interactions to specific regions of proteins. Rather than knowing only that two proteins interact, we can define their interacting regions as well.

If the sequence is random, and with fragments of 300bp, there is a 99% chance that random sequence will have a stop codon. By contrast, a coding sequence of DNA will of course avoid stop codons until the end of the sequence has been reached. If we take fragments of DNA just 300bp long, but in random frames, then 5/6 (83.3%) will be in the wrong frame. Yet 99% of those will have a stop codon; if we can selectively eliminate fragments with stop codons, the in-frame percentage of the library will go from 16.7% up to 96%.

For a fragment to potentially demonstrate a physiological interaction, it must be in an open reading frame (ORF). An ORF is an in-frame sequence that lacks stop codon and has the potential to encode proteins. The process of removing non-ORF (or out-of frame) sequences from a total population is what is termed "ORF filtering". In order to filter ORFs, genomic DNA of Helicobacter pylori was randomly sheared into fragments (100-500bp). The fragments were then inserted upstream of the Ampicillin resistance (AmpR) gene and selected for resistance. Colonies which grew or survived in the presence of Ampicillin were considered to have in-frame ORFs.

Using this method the in-frame ORFs derived from Helicobacter pylori were enriched 70% with 92% coverage of the total coding sequence. Also, filtering ORFs to enrich for genic fragments results in a physiological protein and this offers a robust path for drug targets, treatments towards cancer and the arising resistance to antibiotics.

# RECALL BY GENOTYPE AND CASCADE SCREENING FOR FAMILIAL HYPERCHOLESTEROLEMIA IN A NATIONAL BIOBANK FROM ESTONIA

Maris Alver[1,2], Marili Palover[1,2], Aet Saar[3,4], Kristi Läll[1,5], Seydeh M Zekavat[6,7], Liis Leitsalu[1], Anu Reigo[1], Tiit Nikopensius[1], Tiia Ainla[3,4], Mart Kals[1,5], Reedik Magi[1], Alar Irs[8], Toomas Marandi[3,4], Neeme Tonisson[1,9], Pradeep Natarajan[6,10,11], Andres Metspalu[1,2], Sekar Kathiresan *[6,10], Tonu Esko *[1,6]

[1]Institute of Genomics, University of Tartu, Estonian Genome Center, Tartu, Estonia, [2]Institute of Molecular and Cell Biology, University of Tartu, Department of Biotechnology, Tartu, Estonia, [3]Institute of Clinical Medicine, University of Tartu, Department of Cardiology, Tartu, Estonia, [4]North Estonia Medical Centre, Centre of Cardiology, Tallinn, Estonia, [5]University of Tartu, Institute of Mathematics and Statistics, Tartu, Estonia, [6]Broad Institute of Harvard and MIT, Cambridge, MA, [7]Yale School of Medicine, New Haven, CT, [8]Tartu University Hospital, Heart Clinic, Tartu, Estonia, [9]Tartu University Hospital, Department of Clinical Genetics in TallinnTartu, Estonia, [10]Massachusetts General Hospital, Cardiovascular Research Center and Center for Genomic Medicine, Boston, MA, [11]Harvard Medical School, Department of Medicine, Boston, MA

**AIM**

Large-scale, national biobanks integrating health records and genomic profiles may provide a powerful platform to identify individuals with disease-predisposing genetic variants. Here, we recall such probands carrying familial hypercholesterolemia (FH)-associated variants, perform cascade screening of family members, and describe health outcomes affected by such a strategy.

**METHODS AND RESULTS**

The Estonian Biobank of Estonian Genome Center comprises 52,274 individuals. Among 4,776 participants with exome or whole genome sequences, we identified 27 individuals who carried FH-associated variants in the LDLR, APOB or PCSK9 genes. Cascade screening of 64 family members identified and additional 20 participants carrying FH-associated variants. Via genetic counselling and clinical management of carriers, we were able to reclassify 51% of the study participants from having previously established non-specific hypercholesterolemia to having FH, and identify 32% who were completely unaware of harbouring a high-risk disease-associated genetic variant. Imaging-based risk stratification of study participants led to the initiation or up-titration of statin therapy for 68% of the carriers.

**CONCLUSION**

Genotype-guided recall of probands and subsequent cascade screening for familial hypercholesterolemia is feasible within a national biobank and may facilitate more appropriate clinical management.

# UNCOVERING THE HIERARCHICAL CONFORMATION OF TOPOLOGICALLY ASSOCIATING DOMAINS FROM HI-C DATA

Lin An[1], Tao Yang[1], Johannes Nuebler[2], Jiahao Yang[3], Qunhua Li[1], Yu Zhang[1]

[1]The Pennsylvania State University, Bioinformatics and Genomics, State College, PA, [2]Massachusetts Institute of Technology, Cambridge, MA, [3]TsingHua University, Statistics, Beijing, China

**Motivation**: Mammalian genomes are organized into different levels. Observations from chromosome conformation capture methods (Hi-C) suggests chromatin forms frequent local interactions in certain regions, which is called as Topologically Associating Domains (TADs). While TADs are often used as the smallest structure units to study regulatory mechanism, previous observation shows that hierarchy is present in TADs, with smaller TADs nested within larger ones. Though several different TAD calling algorithms have been developed, limited research has been done to reliably identify hierarchical TAD structures and understand their roles in gene regulation.

**Result**: We developed a new method to call TADs in hierarchy. Our systematic validation based on epigenomics information and gene expression information shows that our method greatly outperforms existing TAD callers in both terms of accuracy and reproducibility. Using this method, we uncovered novel biological insights related to TAD hierarchies. For example, we found the cohensin density has significant positive association with interaction strength within TADs. We found that hierarchical TADs and TADs without hierarchy have many distinct features. Notably, we found that active epigenetic states (promoter associated and enhancer associated) are more enriched in nested TADs than larger TADs. Moreover, our results suggest boundaries in nested TADs tend to be more CTCF-enriched than those in TADs without hierarchical structures. Finally, our results also are in good agreement with the loop excursion model.

**Conclusion**: The new hierarchical TAD caller is able to identify different levels of TADs . Combined with epigenetics and transcriptome information, our results generate new insights towards understanding the complex system of gene regulation.

# RARE MITOCHONDRIAL DNA VARIANT ANALYSIS IN SINGLE OOCYTES USING DUPLEX SEQUENCING

Barbara Arbeithuber[1], Nicholas Stoler[2], Bonnie Higgins[1], James Hester[3], Francisco J Diaz[3], Anton Nekrutenko[2], Kateryna D Makova[1]

[1]Pennsylvania State University, Department of Biology, University Park, PA, [2]Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA, [3]Pennsylvania State University, Department of Animal Science, University Park, PA

Mutations in mitochondrial DNA (mtDNA) contribute to a variety of diseases such as type 2 diabetes mellitus, Leber hereditary optic neuropathy, mitochondrial myopathy, cancer, but also aging. Most of the disease-causing mtDNA mutations are associated with heteroplasmic sites, i.e. mtDNA sites for which multiple variants coexist within a cell or an individual. When the frequency of a disease-associated variant exceeds a threshold, symptoms occur. Considering the maternal transmission of mtDNA, a detailed study of (low-level) mtDNA heteroplasmies and de novo mutations in single oocytes would aid in a better understanding of heteroplasmy inheritance, as well as provide important information on mtDNA germline mutations.

Despite great advances in single-cell sequencing, the analysis of rare variants in mtDNA is hindered by the high error rates associated with the sequencing methods, as well as false positive mutation calls resulting from amplification and DNA lesions. Duplex sequencing, a specialized sequencing method that has the power to greatly reduce such errors (down to the levels below $10^{-7}$) by independently tagging each of the two strands of a DNA duplex, followed by the amplification and sequencing of both strands separately. While true mutations are found at both DNA strands, false positive mutations can be identified and discarded since they are only present at one of the strands or in several reads of one strand.

We were able to improve the library preparation efficiency and successfully performed duplex sequencing on single oocytes originating from different species (mouse, rhesus macaque, and human), despite the high amount of DNA required in the original duplex sequencing protocol. We established a library preparation workflow involving single oocyte lysis, enzymatic enrichment of the circular mitochondrial DNA, and duplex library preparation, followed by paired-end sequencing. In addition to improvements in library preparation, the implementation of a more efficient bioinformatic pipeline for single-strand and duplex consensus formation – *Du Novo* – further improved the power of this method. mtDNA in multiple single oocytes could be sequenced at a duplex consensus sequence depth ranging from about 200x to 2200x yielding high quality sequencing data, and germline mutation frequencies ranging from $2.1 \times 10^{-7}$ to $5.7 \times 10^{-6}$ substitutions per nucleotide were measured in individual oocytes.

# IMPROVED HIGH THROUGHPUT SEQUENCING METHOD TO MAP PROVIRAL INTEGRATION SITES AND MEASURE CLONAL ABUNDANCE.

Maria Artesi[1], Keith Durkin[1], Vincent Hahaut[1], Michel Georges[1], Anne Van den Broeke[1,2]

[1]Unit of Animal Genomics, GIGA, Université de Liège, Liege, Belgium, [2]Laboratory of Experimental Hematology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

Bovine Leukemia Virus (BLV) and Human T-cell leukemia virus-1 (HTLV-1) are closely related deltaretroviruses provoking a polyclonal expansion of infected B- and T- cells respectively, with monoclonal leukemia/lymphoma developing in about ~5% of infected individuals. To date, the molecular mechanisms leading to cellular transformation remain unclear. Both proviruses are largely transcriptionally silent in tumors and their integration sites into the host genome appear highly variable. Identifying proviral insertion sites in the host genome using high throughput sequencing techniques has provided insights into the evolution of BLV/HTLV-1 infections and the expansion of transformed clones in deltaretrovirus induced leukemia/lymphoma.
The methods currently used have a number of limitations such as the utilisation of custom sequencing primers, relatively high sequencing costs, no examination of the 5'LTR host flanking region and limited dynamic range for determining clone abundance. We have developed an alternative high throughput sequencing protocol for tracking proviral integration sites and measuring clonal abundance in BLV and HTLV-1 infected individuals. Our method reduces the amount of sequencing of PCR duplicates by reducing the number of PCR cycles via an enrichment of BLV- and HTLV-1 carrying DNA fragments. This is achieved with a selective end-repair primer extension that incorporates biotinylated dUTPs. In addition to the proviral 3'LTR of the provirus, our approach also assays the 5'LTR, giving additional information on the frequency of 5'-end deletions in proviruses and increasing the dynamic range of the assay. Moreover, we implemented the use of off-the-shelf Illumina primers for the addition of adapters and indexes, which facilitates library multiplexing and avoids the need for custom sequencing primers. We have tested the approach on over 100 BLV and HTLV-1 samples, representing both tumors and preleukemic stages. Our approach allowed for a more accurate genome wide mapping of proviral insertion sites and determination of clone abundance. By assaying the provirus 5' end we identified clones overlooked with previously published methods. Finally, by facilitating greater multiplexing of libraries we have reduced the cost to a level where the technique may be attractive in a clinical setting and adapted for mapping the insertion of other retroviruses, retro-elements or vectors integrated into the genome.

# DISCOVERY OF THE FIRST GERMLINE-RESTRICTED GENE BY SUBTRACTIVE TRANSCRIPTOMIC ANALYSIS IN THE ZEBRA FINCH TAENIOPYGIA GUTTATA

Michelle K Biedermann, Megan M Nelson, Kathryn C Asalone, Alyssa L Pederson, Colin J Saldanha, John R Bracht

American University, Biology, Washington, DC

Developmentally programmed genome rearrangements are rare in vertebrates but have been reported in scattered lineages including the bandicoot, hagfish, lamprey, and zebra finch (Taeniopygia guttata). In the finch, a well-studied animal model for neuroendocrinology and vocal learning, one such programmed genome rearrangement involves a Germline-Restricted Chromosome, or GRC, which is found in germlines of both sexes but eliminated from mature sperm. Transmitted only through the oocyte, it displays uniparental female-driven inheritance, and early in embryonic development it is apparently eliminated from all somatic tissue in both sexes. The GRC comprises the longest finch chromosome at over 120 million basepairs and previously, the only known GRC-derived sequence was repetitive and non-coding. Because the zebra finch genome project was sourced from male muscle (somatic) tissue the remaining genomic sequence and protein-coding content of the GRC remain unknown. Here we report the first protein-coding gene from the GRC: a member of the α-Soluble NSF Attachment Protein (α-SNAP) family hitherto missing from zebra finch gene annotations. In addition to the GRC-encoded α-SNAP, we find an additional paralogous α-SNAP residing in the somatic genome (a somatolog)-making zebra finch the first example in which α-SNAP is not a single-copy gene. We show divergent, sex-biased expression for the paralogs and also that positive selection is detectable in the bird α-SNAP lineage, including the two zebra finch genes. This study presents the identification and evolutionary characterization of the first protein-coding GRC gene in any organism.

# A POSITIVELY SELECTED COMMON MISSENSE VARIANT IN FBN1 CONFERS A 2.2CENTIMETER REDUCTION OF HEIGHT IN THE PERUVIAN POPULATION

S Asgari[1], Y Luo[1], E Bartell[1], R Calderon[2], L Lecca[2], C Contreras[1], R Yataco[2], J T Galea[1], S R Leon[2], J Jimenez[2], J Hirschhorn[1], M Murray[1,2], S Raychaudhuri[1]
[1]Harvard Medical School, Boston, MA, [2]Socios En Salud Sucursal Peru, Lima, Peru

Height is a highly heritable, polygenic trait that shows extensive signals of polygenic selection. Previous height genetic studies, done predominantly in European populations, have identified ~700 independent height-associated variants. These variants might differ between populations due to selection. Peruvians, whose genetic makeup is shaped by admixture between Native Americans (NAT), Europeans (EUR), and Africans (AFR), are among the shortest people in the world. Here we present the first large-scale genetic study of height in the Peruvian population.
We genotyped 4002 individuals from 1,769 households in Lima, Peru, using a customized 720K array, designed to optimize for population-specific rare and coding variants. We first measured the correlation between global ancestry proportions and height using a linear mixed model, accounting for the household as a proxy for environmental and socioeconomic factors. The average proportion of NAT, EUR, and AFR ancestries per individual was 0.81, 0.16, and 0.03 respectively. Increased EUR or AFR ancestry was associated with increased height (p=3.6e-6, 14.5 (0.89) and p=8.3e-5, 16 (1.7) respectively). On the contrary, increased NAT ancestry was associated with reduced height (p=2.3e-6, -12.5 (0.72)).
To identify specific variants driving this association, we performed a GWAS. We identified a new height-associated variant (p=1.5e-9, -2.2 (0.36)). This variant is a missense SNP in FBN1. The minor allele (MAF~5%) is associated with a 2.2cm reduction of height in Peruvians. Most of the known common height-associated variants have effects less than 2mm/allele. We observed that 650 previously identified height-associated variants explained ~7% of height heritability in our cohort whereas, the FBN1 coding variant alone explained ~1%. This variant is hence the single largest effect common variant reported to our knowledge. Local ancestry inference showed the minor allele is carried by NAT haplotypes only. Haplotypes carrying the minor allele were longer (iHS=1.6) and had significantly less nucleotide diversity ($p<0.03$) compared to those with the major allele, suggesting positive selection for this variant.
FBN1 encodes Fibrillin-1, an extracellular matrix glycoprotein that provides structural support in connective tissues. While the clinical significance of this variant is unknown, other missense mutations in FBN1 cause skeletal anomalies, including the monogenic diseases Marfan syndrome and Acromicric dysplasia. Our findings highlight that studying the genetics of polygenic traits in new populations can uncover novel trait-associated variants of large effects in functionally relevant genes. These results suggest that either small stature or other FBN1-related traits like skin thickness have offered evolutionary advantage in Native American ancestors of Peruvians. Both may have offered advantage at high altitude. Functional studies in appropriate cell lines and tissues are nowneeded to better understand the relationship between this variant and stature.

# MPRANALYZE: A STATISTICAL FRAMEWORK FOR MASSIVELY PARALLEL REPORTER ASSAY (MPRA) DATA

Tal Ashuach[1], David S Fischer[1,3], Anat Kreimer[1,2], Fabian Theis[3], Nadav Ahituv[2], Nir Yosef[1,2]

[1]University of California, Berkeley, Department of Electrical Engineering and Computer Science, Berkeley, CA, [2]University of California, San Francisco, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA, [3]Helmholtz Centre, Institute of Computational Biology, Munich, Germany

Massively parallel reporter assays (MPRAs) is a technique that enables testing thousands of regulatory DNA sequences and their variants in a single, quantitative experiment. Since MPRA is still a nascent technology, there's no set of computational methods dedicated to effectively leverage their promise. Development of such methods could help improve future MPRA candidate sequence selection, enhance our ability to predict functional regulatory sequences and increase our understanding of the regulatory code and how its alteration can lead to a phenotypic consequence.

Here we present MPRAnalyze: a statistical framework dedicated to analyzing MPRA count data. MPRAnalyze addresses the major questions that are posed in the context of MPRA experiments: estimating the magnitude of the effect of a regulatory sequence in a single condition setting, and comparing differential activity of regulatory sequences across multiple conditions. The framework allows for various distributional assumptions and uses generalized linear models to account for uncertainty in both DNA and RNA observations, control for various sources of unwanted variation, and incorporate negative controls for robust hypothesis testing, thereby providing clear quantitative answers in complex experimental settings.pr

We demonstrate the robustness, accuracy and applicability of MPRAnalyze on simulated data and published data sets. MPRAnalyze is implemented as a publicly available R package.

# KIPOI: SHARING AND RE-USE OF PREDICTIVE MODELS FOR REGULATORY GENOMICS

Ziga Avsec[1], Roman Kreuzhuber[2,3], Johnny Israeli[4], Jun Cheng[1], Lara Urban[3,5], Avanti Shrikumar[4], Anshul Kundaje[4], Oliver Stegle[3,5], Julien Gagneur[1]
[1]Technical University of Munich, Bioinformatics, Munich, Germany, [2]University of Cambridge, Dept. of Haematology, Cambridge, United Kingdom, [3]EMBL, EBI, Hinxton, United Kingdom, [4]Stanford University, Dept. of Genetics, Stanford, CA, [5]EMBL, Genome Biology Unit, Heidelberg, Germany

The recent explosion of genomics technologies is gradually unraveling the relationships between every step of gene expression, from genome to chromatin, transcription, splicing, post-transcription, and translation. Along with the experimental progresses, advances in machine learning led to a wave of models predicting these molecular phenotypes from genetic sequence alone. These advances, experimental and computational, hold the promise to reach a global understanding of gene expression and to interpret personal genomes.
However, regulatory genomics models are implemented in various programming languages and frameworks, and are made available through different channels, such as code repositories or supplementary material of articles. This heterogeneity hinders their application to new data, including personal genomes. It also makes it difficult to compare existing models and to combine them into composite models.
Here, we introduce Kipoi, a programmatic infrastructure and a repository of ready-to-use trained models for regulatory genomics. Kipoi provides unified means to access models trained in different machine learning frameworks. This allowed us to wrap models such as DeepSEA, DeepBind, BASSET, CpGenie, Basenji, Labranchor, deepTarget and MaxEntScan, thereby covering canonical predictive tasks in transcriptional and post-transcriptional regulation. Kipoi furthermore specifies unified means for data reading and preprocessing, which greatly facilitates direct application of existing models to new data coming in standard formats like fasta or bed. As a showcase, we demonstrate how benchmarks of existing models for predicting transcription-factor affinity across multiple datasets are easily programmed in few lines of code.
Through standardization, Kipoi enables building generic tools for downstream analyses, which are uniformly applicable to Kipoi models. We implemented a single module that performs variant effect prediction uniformly for any wrapped model based on genomic sequences. Also, standardization allows composing models into larger ones. We demonstrate both of these functionalities by composing models scoring 3' and 5' splice sites, splice branchpoints and RNA-binding of proteins to predict pathogenic splice site variants. Composing multiple models increased the auROC performance in distinguishing pathogenic from benign splice site variants of the ClinVar database from 0.69 to 0.88.
Altogether, by providing a central place for sharing, benchmarking and efficiently re-using predictive models in regulatory genomics, we foresee Kipoi as a catalyst for genome research and personalized medicine.
Kipoi is implemented as a python package (github.com/kipoi/kipoi) and it is also accessible from the command line or R. The model repository is located at github.com/kipoi/models.

# DISCOVERING SOMATIC MOSAIC VARIANTS WITH HIGH ALLELE FREQUENCY USING EXHAUSTIVE PAIRWISE COMPARISON OF SINGLE-CELL GENOMES

Taejeong Bae[1], Vaccarino Flora[2], Alexej Abyzov[1]

[1]Mayo Clinic, Department of Health Sciences Research, Rochester, MN,
[2]Yale University, Child Study Center, New Haven, CT

Somatic mosaicism refers to the genetic phenomenon in which cells within an organism have distinct genomes derived from accumulating post-zygotic mutations. Mosaic mutations may have different phenotypic consequences in normal and pathophysiological conditions depending on their genomic location, the developmental stage at which they occur, and the proportion of cells that they affect. Generally, most mosaic mutations go unnoticed due to their existence within only a few or even just one cell. To detect mutations that occur at such low frequencies, a single-cell sequencing approach is crucial. Single-cell analysis finds variants that uniquely exist in a single cell genome by typically comparing variants in each single cell data with germline variants in bulk tissue data. However, this approach is likely to miss mosaic mutations with high allele frequency that arise at a very early time (i.e., between the first and fifth division after fertilization), manifesting at the level of roughly 1–25% of the allele frequency within the human body. In contrast to the variants with low allele frequencies, those mutations may potentially affect organismal health because they persist in more cells the earlier they arise. Here, we developed a new strategy for variant discovery using exhaustive pairwise comparison of single-cell genomes to detect mosaic mutations with high allele frequency. In exhaustive pairwise comparison, only comparisons between single cells that do and do not carry a somatic variant result in a call for the variant. Based on such comparisons and conformity of call recurrence to single cells that are expected and are not expected to carry the same mosaic variants, we calculated an "explanation score" quantifying the degree of match between actual calls and the expectation. Using this score, we filtered out germline and false-positive calls. Applying this method to real data, we discovered mosaic variants over 98% concordant with calls from comparing single cells with bulk tissue. We also discovered 68 additional mosaic variants. These variants were enriched in the variants with high allele frequencies as determined by the capture-seq experiments. This method, therefore, can be used for precisely detecting mosaic mutations, enabling studies of mutation rates, mutational processes and developmental outcomes of cell dynamics during very early embryogenesis.

# AQUATIC PLANTS HAVE LOST A KEY IMMUNE SIGNALLING PATHWAY AND REVEAL PREVIOUSLY UNKNOWN COMPONENTS OF IMMUNITY.

Erin L Baggs[1,2], Wilfried Haerty[1], Ksenia V Krasileva[1,2]

[1]Earlham Institute, Plant genomics, Norwich, United Kingdom, [2]The Sainsbury Laboratory, Krasileva group, Norwich, United Kingdom

A key paradigm in our understanding of plant disease is that the transition of plants from sea to land, sparked an arms race with pathogens. The increased susceptibility of plants on land is largely due to their dependence on micro-organisms for nutrients. The ensuing co-evolution between plants and their pathogens has shaped the plant immune system into the one we see today.

By profiling the immune gene family of NLRs across over 50 plant species we identified species with by far the fewest NLRs (10s compared to 100s-1000s). Interestingly, these species descended from multiple lineages of monocots and dicots sharing only an aquatic lifestyle.

Using clustering methods to group all gene families across 10 species of monocots of which two were aquatic and between nine land and two aquatic dicot species, we observed that in both the monocot and dicot lineage aquatic species lost the same characterized downstream immune signalling components.

To identify novel components in the NLR immune signalling pathway we surveyed both lineages for gene families with similar patterns of gene family loss. We identified genes previously implicated in disease resistance as well as novel candidates. We are now carrying out functional assays to validate our predictions of other novel immune signalling components.

Our analysis revealed the loss of a key immune signalling pathway in independent lineages of plants that moved back into an aquatic environment. By looking at other genes with a similar loss pattern we were able to predict new putative components of plant immunity. The retained immune genes shed light on a minimal plant immune system required for life under water, while the lost genes highlight the additional components required for the life of land plants.

# LONGITUDINAL STUDY OF GENE EXPRESSION AND REGULATION DURING A CRITICAL PERIOD OF THE AGING PROCESS

Brunilda Balliu[1], Matt Durrant[1], Olivia M de Goede[1], Nathan S Abell[1], Bosh Liu[1], Kevin S Smith[1], Lars Lind[2], Erik Ingelsson[3], Stephen B Montgomery[1]

[1]Stanford University School of Medicine, Pathology, Stanford, CA, [2]Stanford University School of Medicine, Medicine, Stanford, CA, [3]Uppsala University, Medical Sciences, Uppsala, Sweden

Molecular and cellular changes are intrinsic to aging and age-related diseases. Existing studies have investigated the combined effects of age and genetics on gene expression; however, there has been no large long-term, longitudinal characterization of these molecular changes within individuals. We performed RNA sequencing in whole-blood from participants of PIVUS, a population-based longitudinal cohort study, at ages 70 and 80. We examined how gene expression and genetic regulation of transcription are altered over this 10-year period. We observed that gene expression is strongly correlated between the two ages, and that, for the majority of individuals, measurements at the two ages cluster together. Moreover, we discovered 1,291 genes (8% of tested genes) whose expression level was significantly associated with age (FDR < 5%), with a larger proportion of differentially expressed genes showing down-regulation with age ($\pi$ = 54.29%, $\chi^2$ test; p-value = $2.2 \times 10^{-3}$). Application of enrichment analysis to the differentially expressed genes showed significant enrichment (q-value < 1%) in pathways related to metabolism of proteins and RNA, regulation of DNA replication, adaptive immune system, and regulation of apoptosis. Importantly, we observed a depletion in the number of regulated genes at age 80, relative to age 70, both in terms of eQTL as well as sQTL effects. Moreover, while the replication rate of eQTLs from age 70 (or 80) to age 80 (or 70) was high (>90%); replication of eQTLs from age 70 to age 80 was lower (Binomial proportion test; p-value = $3.3 \times 10^{-3}$), with impacted genes enriched in DNA repair pathways. This difference remained significant for a range of discovery and replication FDR thresholds, as well as minor allele frequency thresholds for the eQTLs. Last, we observed that overall allelic imbalance increased slightly within individuals with age (Wilcoxon signed rank test; p-value = $1.5 \times 10_{-2}$). Together, these findings suggest that, although genetic effects on expression and splicing are highly stable as individuals age, the genetic architecture of a subset of genes is unstable and is characterized by a loss of genetic control at the population level which occurs primarily through increases in environmental or stochastic effects. This is the first study to investigate changes in genetic regulation of total gene expression and alternative splicing in this critical period of the aging process.

# CONSERVATION OF TRANSCRIPTIONAL VARIATION ACROSS HUMAN, MOUSE AND … ARMADILLO?!

Sara Ballouz, Jesse Gillis

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

Controlling for sources of variability within a targeted experiment is standard practice in gene expression studies, where transcriptional noise is inherent and not fully understood. A common control is to condition on genetic background, usually through inbred model organisms, strains, clones or cell lines. Although this decreases variability within the experimental system, generalizing to another genetic background or strain is perilous. This is rarely more true than in translational or comparative biology, where genetic background changes profoundly (i.e., across species).

In order to assess the impact of genetic background on mammalian transcriptomic findings, we exploited the polyembrony of the wild nine-banded armadillo (Dasypus novemcinctus). It is an ideal model system for our purposes since it is outbred, yet produces monozygotic quadruplets in every litter, serving as natural biological replicates. Surprisingly, this unique reproductive mode has yet to be exploited in transcriptional studies, even though its discovery well over a hundred years ago was critical to the field of developmental biology.

First, we sequenced the blood transcriptomes of five litters of armadillo quadruplets to generally assess transcriptional variation. We found 2982 genes with human and mouse homologs exhibiting statistically significant differences (FDR<0.001) between quadruplet sets, indicating variability sensitive to genetic background. Immune function and cell cycle homologs were particularly prominent. To determine if these genes, sensitive to genetic background, were generating spurious results within mouse and human experiments, we performed a meta-analysis across 3275 pre-existing gene expression studies. We find highly variable genes are often called differentially expressed in mouse ($rs=0.28$, $p<2E-16$) and human ($rs=0.27$, $p<2E-16$).

Our findings suggest that genes sensitive to genetic background can be easily identified and are a potentially useful probe for results that will not generalize across species, helping to address the replicability “crisis” in transcriptomic, functional genomics and beyond.

# DANIO-CODE: A CENTRAL RESOURCE FOR STANDARDISED ANNOTATION AND RE-ANNOTATION OF WHOLE-GENOME DATA FOR THE MODEL VERTEBRATE ZEBRAFISH

Damir Baranasic[1], Abdul K Mukarram[2], Matthias Hörtenhuber[2], Michaël Dong[2], Piotr J Balwierz[1], Dunja Vucenovic[1], Fiona Wardle[3], Carsten Daub[2], Ferenc Mueller[4], Boris Lenhard[1]

[1]Imperial College London, Computational Regulatory Genomics group, MRC London Institute of Medical Sciences, London, United Kingdom, [2]Karolinska Institute, Department of Biosciences and Nutrition, Huddinge, Sweden, [3]King's College London, Randall Division of Cell and Molecular Biophysics, London, United Kingdom, [4]University of Birmingham, School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, Birmingham, United Kingdom

Because it is easy to grow and manipulate, Zebrafish is a prominent model organism in the fields of developmental biology, vertebrate genomics, epigenetics, and disease modeling, among others. Although diverse projects generate and publish new Zebrafish datasets on a daily basis, no platform exists which would uniformly process and annotate them, as ENCODE or modENCODE do for human and mouse or for worm and fly datasets, respectively. Therefore, available Zebrafish datasets are unlinked and non-uniformly processed which complicates their comparison and integrative analysis. Moreover, many high-quality datasets remain underexplored. The DANIO-CODE consortium encompasses several research groups which use Zebrafish as a model organism. It aims to gather all available Zebrafish datasets, annotate and process them uniformly, and share results with the community. So far, 4 TB of data were collected, mostly RNA-seq, ChIP-seq, and BS-seq. Adapted ENCODE or custom pipelines were used to uniformly process RNA-seq, ChIP-seq, BS-seq, ATAC-seq, MNase-seq, and CAGE-seq data. As a result, we constructed gene expression and epigenetic profiles, detecting novel transcripts and functional elements, and modeling dynamic changes in gene expression and epigenetic landmarks during Zebrafish embryonic development. Produced annotations are available at the DANIO-CODE Data Coordination Centre (http://danio-code.zfin.org).

# 28,113 Haploid Sperm Genomes From 18 Individuals Ascertained By A Droplet-Based Single-Sperm Sequencing Technology

Avery Davis Bell[1,2], Curtis J Mello[1,2], Steven A McCarroll[1,2]

[1]Harvard Medical School, Department of Genetics, Boston, MA, [2]Broad Institute of MIT and Harvard, Stanley Center and Medical and Population Genetics, Cambridge, MA

Meiosis and recombination generate genetic diversity and many mutations. Earlier work has inferred recombination events from genetic variation in pedigrees and populations and has investigated a broader set of meiotic phenotypes in studies of 96-122 individual sperm cells. These studies demonstrate that recombination varies among individuals, but cost and scalability have limited inferences about variation and co-variation of meiotic outcomes within individuals and across individual cells.

To greatly expand the ability to measure and learn from meiotic outcomes, we developed Sperm-seq, a droplet-based single-cell DNA sequencing technology for human sperm cells. We now routinely make sequencing libraries for 1000-2000 individual sperm cells, sequence them to low coverage (generally 0.01-0.02x, though up to 0.1x is permitted by the complexity of Sperm-seq libraries), and infer recombination patterns and chromosomal ploidy for each cell. We completely phase human genomes using Sperm-seq data, generating whole genome sequence data in which all heterozygous SNPs are phased into chromosome-length haplotypes. We identify crossovers in each gamete using a Hidden Markov Model to find transitions between haplotypes. We recognize aneuploidy events from deviations in sequence coverage and the presence of multiple haplotypes.

We have so far analyzed 28,113 sperm from 18 individuals of presumed normal fertility, yielding sequence coverage of 0.8-4% of the genome in 1,000-2,200 cells per individual. We identified >700,000 crossovers (median 26 per cell, range 10-40), creating recombination maps for 18 people. We find that these recombination maps correlate broadly with pedigree- and population-based male recombination maps, including concentration of crossovers near the telomeres, yet they exhibit substantial inter-individual differences.

We are now analyzing these datasets to characterize meiotic outcomes, such as frequency and meiotic division of origin of aneuploidy and spatial distribution of crossovers, and are investigating the relationships among them, their inter-individual variation, and their co-variation across cells. We uncovered inter-individual variation in the extent and pattern of crossover interference, the tendency of crossovers on the same chromosome to occur farther apart from one another than would be expected by chance. We observe crossover interference in each of the 18 individuals (median distance between adjacent crossovers 66-84Mb; $p <<< 10^{-4}$ in a permutation test). The magnitude of crossover interference differs among individuals and negatively correlates with recombination rate.

High-throughput single-sperm sequencing makes newly visible the variability of meiotic processes within and among individuals and will allow new approaches to genetic mapping and genome assembly.

# GENOMIC AND PHARMACO-GENOMIC CHARACTERISTICS OF PANCREATIC CANCER ORGANOIDS

Pascal Belleau[1], Astrid Deschênes[2], Hervé Tiriac[2], Lindsey A Baker[2], Timothy Somerville[2], Alexander Krasnitz[1], David A Tuveson[2]

[1]Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY, [2]Cold Spring Harbor Laboratory, Cancer Center, Cold Spring Harbor, NY

The Tuveson laboratory has developed a novel methodology to culture organoids from human healthy pancreatic ductal epithelial tissues and from pancreatic ductal adenocarcinomas (PDAC). These patient-derived, 3-dimensional models of PDAC open new opportunities for deep genomic and transcriptomic studies of the disease, and for massively parallel individualized drug screens. These studies are facilitated by nearly 100% neoplastic purity of the organoid cultures, as compared to very low cellularity of primary tumors.

Our first objective was to determine the extent to which the organoids preserve the genomic characteristics of the primary tumors of origin. To this end we compared the organoid whole-genome sequences with those of the corresponding primary tumors and found a high degree of similarity in the high-frequency part of somatic mutation spectrum. Importantly, additional low-frequency variants could be observed in the organoids due to their higher purity.

In parallel, gene expression of 77 organoids (14 normal and 63 PDAC) was evaluated by RNA-seq. Using non-negative matrix factorization (NMF), we identified three distinct classes among these transcriptional profiles, each with its own expression signature.

For 55 of these organoids, we also determined their sensitivity to five commonly used chemotherapies. Combining these data with organoid gene expression profiles in a pharmaco-transcriptomic analysis, we identified pathways associated with specific drug responses. Our data suggest that organoid profiling may help direct patient treatment to enhance response and provide the best possible outcome while potentially reducing toxicity.

# PROTEOMIC PROFILING OF CD4+ AND CD8+ T CELLS FROM MULTIPLE SCLEROSIS PATIENTS AND HEALTHY CONTROLS

Tone Berge[1,2], Anna Eriksson[2,3], Ina S Brorson[2,3], Einar A Høgestøl[2,3], Steffan D Bos[2,3], Hanne F Harbo[2,3], Frode S Berven[2,3]

[1]OsloMet – Oslo Metropolitan University, Department of Biotechnology and Chemistry, Oslo, Norway, [2]Oslo University Hospital, Department of Neurology, Oslo, Norway, [3]University of Oslo, Department of Neurology, Oslo, Norway, [4]University of Bergen, Department of Biomedicine, Bergen, Norway

Multiple sclerosis (MS) is an autoimmune disorder affecting the central nervous system. It is one of the most common neurological conditions among young adults, leading to both physical and cognitive impairments. The disease develops in genetic susceptible individuals, triggered by common environmental factors such as reduced serum levels of vitamin D, virus infection and smoking. The exact underlying pathogenesis remains unclear, but T lymphocytes, both CD4+ and CD8+ T cells, have long been considered to play pivotal roles in MS. Also, the genetic architecture of MS susceptibility emerged from genome-wide association studies, indicates an important role for the adaptive immune system, in particular T cells, for MS-disease.

To get an insight into immune-cell processes in MS, we evaluated immune dysregulation at the protein level using liquid chromatography combined with mass spectrometry analyses. We have analyzed the proteomic profile of purified immune-cell subsets, i.e. CD4+ and CD8+ T cells, which allows us to disentangle potential cell-subtype specific differences that could not be detected in a heterogeneous cell material, allowing a comprehensive understanding of disease mechanisms.

CD4+ and CD8+ T cells were purified from blood by magnetic separation from 13 treatment-naïve female patients with relapsing-remitting MS and 14 age and gender-matched healthy controls. More than 2000 proteins were identified, of which 195 and 230 proteins were differentially expressed in CD4+ and CD8+ T cells respectively ($p<0.05$). In order to extract the true candidate proteins differentially expressed between MS cases and healthy controls, we used a more stringent filter for selection (i.e. fulfilling two of the following criteria; fold change > 2, area under the curve > 0.8, $p<0.01$). We ended up analyzing 91 and 62 proteins from CD4+ and CD8+ T cells, respectively, using the Ingenuity Pathway Analysis software. The results from these analyses will be presented. It should be highlighted that a larger verification study is needed to replicate the current findings.

# QUANTITATIVE ANALYSIS OF DOSAGE COMPENSATION IN FLY BRAINS USING TRANSCRIPT 5' PROFILING

Vivek Bhardwaj*[1,2,3], Giuseppe Semplicio*[1], Thomas Manke[2], Asifa Akhtar[1]

[1]Max Planck Institute of Immunobiology and Epigenetics, Department of Chromatin Regulation, Freiburg, Germany, [2]Max Planck Institute of Immunobiology and Epigenetics, Bioinformatics Unit, Freiburg, Germany, [3]University of Freiburg, Faculty of Biology, Freiburg, Germany

* authors contributed equally

Promoter architecture, shape and position of transcription start sites (TSS) play an important role in the regulation of eukaryotic gene expression. Cap Analysis of Gene Expression (CAGE) has been the widely used method to detect transcription start sites, perform promoter profiling and detect promoter shifts between tissues or during development. RNA-Sequencing on the other hand, has been the method of choice for gene expression profiling and differential expression analysis. In this study, we describe an approach to perform high resolution detection of TSS as well as quantitative TSS expression analysis across samples with a single experiment. We developed a fast and simple protocol called MAPCap (Multiplexed Affinity Purification of Capped RNA) which allows multiplexed processing of low input samples for the analysis of TSS. Using our newly developed tool ICETEA (Integrating Cap Enrichment with Transcript Expression Analysis), we detect TSS at base-pair resolution, quantify TSS expression after UMI-based removal of PCR duplicates, and perform differential expression analysis using both internal and external normalization controls. We analyzed TSS expression in the brains of *Drosophila melanogaster* larvae, using each Fly larva as a biological replicate and observed the effects of knockout of MLE (male-less) helicase on X chromosome dosage compensation at genes and their transcript isoforms. Our results expand the scope of TSS profiling methods to differential transcript expression analysis.

# SOMATIC VARIANT CALLING PIPELINE FOR DETECTION OF TUMOR-DERIVED VERY LOW FREQUENCY VARIANTS IN CELL-FREE DNA: SOMATICTRICALLER

Preetida J Bhetariya[1], David A Nix[2], Sabine Hellwig[3], Gabor Marth[1], Mary P Bronner[1], Hunter R Underhill[1]

[1]University of Utah, Salt Lake City, UT, [2]Huntsman Cancer Institute, Salt Lake City, UT, [3]ARUP Laboratories, Salt Lake City, UT

Circulating tumor DNA (ctDNA) is DNA released by tumor cells into the blood carrying the mutations of the original tumor. The detection of ctDNA in blood samples via sequencing is a promising approach for cancer diagnosis and disease monitoring. Due to the dilutional effects of naturally occurring cell-free DNA (cfDNA) derived from healthy cells, ctDNA associated with early-stage and premetastatic solid tumor malignancies are commonly present at a very low frequency (<1%). In addition, variant calling for ctDNA may be confounded by noise due to sequencing errors, PCR artifacts, and alignment errors. Therefore, a highly sensitive and specific analysis method is critical to distinguish true mutations from noise-driven, low frequency false positives. We have developed and optimized an in-house variant calling pipeline, SomaticTriCaller by combining three variant callers MuTect2, LoFreq and Strelka1.0. We used the cfDNA from 11 healthy controls to generate a local error rate corrected, variant quality score. To evaluate the performance of SomaticTriCaller, a simulated dataset was created. Buffy coat DNA was sheared (fragments ~ 170bp) to simulate the normal size distribution of cfDNA and only 10 ng of sheared DNA was used for library preparation. Eight libraries were individually prepared, enriched using a custom NGS panel and then sequenced (HiSeq 2500) with read depth of ~3,500X. Four samples were merged and spiked with specified variant allele frequency (VAF; 10% to 0.1%) to create tumor database. Variant calling was performed with SomaticTriCaller, Freebayes, and VarScan2. SomaticTriCaller was able to detect VAF down to 0.4% with 98.6% sensitivity with false discovery rate <4%. In comparison, Freebayes and VarScan2 was 95% and 98% sensitive with false positive rate at <20% and <19%, respectively at 0.4% VAF. To validate the pipeline, we sequenced 10 cfDNA samples at a similar read depth from patients with colorectal, pancreatic and melanoma cancer, harboring a BRAF or KRAS ctDNA variant ranging from 0.38% to 2.5% verified by droplet digital PCR (ddPCR). Variant calling by SomaticTriCaller successfully detected 100% of the known variants. There was a strong correlation with VAF by ddPCR (Pearson's r = 0.98; P<0.001). Collectively, these results indicate that SomaticTriCaller is an accurate tool for detecting very low frequency ctDNA variants. The coupling of SomaticTriCaller with improved NGS protocols (high read depth, improved error correction) will substantially advance the use of cell-free DNA as a non-invasive diagnostic test.

# RODENTS ANNOTATION IN ENSEMBL

Konstantinos Billis*[1], Osagie Izuogu*[1], Carlos García Girón[1], Thibaut Thibaut Hourlier[1], Leanne Haggerty[1], Denye Ogeh[1], Daniel N Murphy[1], Rishi Nag[1], Daniel Barrell[2], Fergal J Martin*[1], Paul Flicek[1], Bronwen Aken[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [2]Eagle Genomics Ltd, Cambridge, United Kingdom

The Ensembl gene annotation system has recently been redesigned to allow the rapid inclusion of new species in Ensembl. Processing time for generating a gene set has been reduced to approximately two weeks, while still providing the high quality annotation and large number of secondary analysis tracks Ensembl is known for. Our approach is based on a combination of RNA-seq alignments, annotation projection via whole genome alignments and protein-to-genome alignments using selected UniProt proteins.
Using the new pipeline, we recently released annotation for 13 rodent assemblies including prairie vole, male and female naked mole rat and two different Chinese hamster ovary cell line assemblies. This represents part of a broader approach to greatly increase the amount of rodent data available in Ensembl. This includes our continually updated, high quality GENCODE mus musculus gene set and imported annotation generated by UCSC for 15 mouse strains and 3 other mus species (mus caroli, mus pahari and mus spretus). When combined with our comparative genomics data and APIs, users have an unprecedented opportunity for researching rodent evolution through Ensembl. All of these data are available in our current release (92). Where available, annotated genomes will include RNA-seq data, including tissue-specific gene tracks, which can be viewed on the Ensembl genome browser. Data and tools to facilitate research on rodents will be accessible through our website (www.ensembl.org and particularly http://www.ensembl.org/Mus_musculus/Info/Strains ), REST API (http://rest.ensembl.org), Variant Effect Predictor (www.ensembl.org/Tools/VEP), BioMart (http://www.ensembl.org/biomart) and our public MySQL server (ensembldb.ensembl.org).

# A GLOBAL TRANSCRIPTIONAL NETWORK CONNECTING NONCODING MUTATIONS TO CHANGES IN TUMOR GENE EXPRESSION

Wei Zhang*[1], Ana Bojorquez-Gomez*[1], Daniel Ortiz Velez[2], Guorong Xu[3], Kyle S Sanchez[1], John P Shen[1], Kevin Chen[2], Katherine Licon[1], Collin Melton[4], Katrina M Olson[1,5], Michael K Yu[1], Justin K Huang[1,6], Hannah Carter[1], Emma K Farley[1,5], Michael Snyder[4], Stephanie I Fraley[2], Jason F Kreisberg[1], Trey Ideker[1,2,6,7]

[1]University of California San Diego, Department of Medicine, La Jolla, CA, [2]University of California San Diego, Department of Bioengineering, La Jolla, CA, [3]University of California San Diego, Center for Computational Biology and Bioinformatics, La Jolla, CA, [4]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [5]University of California San Diego, Division of Biological Sciences, La Jolla, CA, [6]University of California San Diego, Bionformatics and Systems Biology Program, La Jolla, CA, [7]University of California, Cancer Cell Map Initiative (CCMI), San Diego and San Francisco, CA

Although cancer genomes are replete with noncoding mutations, the effects of these mutations remain poorly characterized. Here we perform an integrative analysis of 930 tumor whole genomes and matched transcriptomes, identifying a network of 193 noncoding loci in which mutations disrupt target gene expression. These 'somatic eQTLs' (expression quantitative trait loci) are frequently mutated in specific cancer tissues, and the majority can be validated in an independent cohort of 3,382 tumors. Among these, we find that the effects of noncoding mutations on *DAAM1*, *MTG2* and *HYI* transcription recapitulate in multiple cancer cell lines, and that increasing DAAM1 expression leads to invasive cell migration, *in vitro*. The somatic eQTL network is disrupted in 88% of tumors, suggesting widespread impact of noncoding mutations in cancer.

# WHOLE-EXOME SEQUENCING DISCOVERIES OF RARE GENETIC VARIANTS ASSOCIATED WITH HUMAN BLOOD METABOLITES

Lorenzo Bomba[1], Klaudia Walter[1], Adam Butterworth[2], Ian Dunham[3], Nicole Soranzo[1,4,5,6]

[1]Wellcome Sanger Institute, Open Targets and Department of Human Genetics, Hinxton, Cambridge, United Kingdom, [2]MRC/BHF Cardiovascular Epidemiology Unit, University of Cambridge, Strangeways Research Laboratory, Department of Public Health and Primary Care, Cambridge, United Kingdom, [3]European Bioinformatics Institute (EMBL-EBI), Open Targets and European Molecular Biology Laboratory, Hinxton, Cambridge, United Kingdom, [4]University of Cambridge, Department of Haematology, Cambridge, United Kingdom, [5]University of Cambridge, The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, Cambridge, United Kingdom, [6]British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Cambridge, United Kingdom

Inborn errors of metabolism are extremes of a much wider spectrum of genetic variation in human metabolism. High-throughput technologies (i.e. NMR spectroscopy, ultra-high-performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) and massively-parallel genome sequencing) allow to investigate this assumption through statistically powered genetic association scans in representative human population samples. Many common genetic variants in genes encoding enzymes and transporters have already been discovered to be associated with metabolite levels (Shin *et al.* 2014; Long *et al.* 2017). Here we extend previous discoveries, focusing on rare genetic variants characterised through whole exome sequencing (WES) and whole-genome sequencing (WGS) in a cohort of 6,729 healthy individuals (INTERVAL study) for whom ~ 1000 metabolite concentrations were profiled in blood using the Metabolon platform. We apply four rare variants tests (RVTs, including burden, Madsen and Browning, variable threshold, sequence kernel association tests) to explore different allelic architectures of combinations of rare variants (MAF < 0.1%). We describe 40 gene-metabolite associations implicating novel rare variants. Of the 27 unique genes implicated, 17 (63%) had no previous evidence for genetic associations with metabolites, and 100% had associations driven by variants newly reported here. The majority of these associations identify genes of potential pharmacological importance, suggesting that genetically-controlled alteration in biochemical functions represented by the metabolites may modulate the activity of the corresponding gene targets.

# EXPLORING THE RELATIONSHIP BETWEEN EXPRESSION AND CHROMATIN DYNAMICS FROM A TEMPORAL PERSPECTIVE

Beatrice Borsari[1,2], Cecilia Klein[1,2], Ramil Nurtdinov[1,2], Emilio Palumbo[1,2], Bruna R Correa[1,2], Amaya Abad[1,2], Alexandre Esteban[1,2], Roderic Guigó[1,2]

[1]Centre for Genomic Regulation, Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat Pompeu Fabra, Experimental and Health Sciences, Barcelona, Spain

Combinations of various histone modifications occurring at genomic regulatory elements are key in determining the location of transcription factors, and contribute to a fine temporal and spatial modulation of transcription. Methods to predict gene expression, which rely on patterns of histone modifications, have been implemented with great accuracy. Nevertheless, evidence of expression for genes that are selectively transcribed during fly development but lack active histone marks has been reported recently. Similarly, studies of diurnal rhythms in mouse liver suggest that, in a situation of fast changing transcription, methylation marks do not provide linear measures of transcription. These results have contradicted the widely accepted belief of association of histone modifications with transcriptional activation or repression. To decode the temporal dynamics of histone modifications and their contribution to changes in gene expression, we rely on a seven-days model of trans-differentiation of human proB cells to macrophages. For each time point, we have analyzed the transcriptome and the profile of nine histone marks. With a dynamic programming technique we are able to reconstruct, for a given gene, pairwise alignments between its expression and chromatin time-series profiles. This methodology provides a computational framework to unravel genome-wide temporal interplays between expression and histone marks. In our model, the epigenome appears to be a more stable system when compared to the dynamic behavior of the transcriptome, which responds fast to external stimuli. Expression and signatures of active chromatin marks correlate within single time points; however, we identify groups of genes lacking temporal consistency between these two components. Genes characterized by constant expression profiles typically show increased signals of active chromatin marking, as if the presence of these marks was more related to gene transcriptional stability through time, rather than being required for the transcription initiation process. Moreover, chromatin signals show either constant or delayed profiles, which do not promptly reflect the temporal changes in expression observed for a fraction of genes. In this context, we propose a model in which active histone post-translational modifications may not be responsible for changes in gene expression, but rather cooperate for the maintenance of specific transcriptional programs throughout cell divisions.

# REFINED MAP OF GENE EXPRESSION REGULATION IN HUMAN CD4 REGULATORY T CELLS GUIDES FUNCTIONAL FINE-MAPPING OF IMMUNE DISEASE ASSOCIATED VARIANTS

Lara Bossini-Castillo[1], Dafni A Glinos[1], Natalia Kunowska[1], Gosia Golda[1], Abigail Lamikanra[2], David Roberts[2], Gosia Trynka[1]

[1]Wellcome Sanger Institute, Cellular Genetics, Cambridge, United Kingdom, [2]John Radcliffe Hospital, Medical Sciences Division, Oxford, United Kingdom

Regulatory T cells (Tregs) suppress the immune response and help prevent autoimmune diseases (ADs). Variants associated to ADs are enriched within promoters and enhancers specifically active in Tregs. Despite their importance to human diseases most studies focused on mouse models and no large scale resources from human Tregs are available. Therefore, to address this gap and provide a resource for functional fine-mapping of disease alleles to effector function, we isolated Tregs from 100 healthy blood donors. We quantified gene expression using RNA-seq, and profiled chromatin activity by ATAC-seq and levels of histone modifications corresponding to promoters and enhancers through ChIPmentation assay for H3K4me3 and H3K27ac.

We defined 40,257 high confidence chromatin accessible regions and captured the expression of 13,275 genes. Histone modification analysis is ongoing. We detected genetic effects on expression of 5,460 genes (eQTL) (at FDR < 0.05). For over 35% of the Treg eQTLs we did not observe an effect in naïve CD4 T cells from the BLUEPRINT project. Furthermore, when considering all the blood cell types assayed by the BLUEPRINT, 15.9% of eQTLs were detected only in our Treg dataset, suggesting that we captured many cell type specific effects. We found 1,253 chromatin accessible regions under genetic control (caQTL) (at FDR < 0.05). Sixty three percent (503 loci) of the caQTL variants also overlapped with eQTLs at a nearby gene (LD r2>0.8 between eSNP and caSNP). We observed that 72 of the Treg eQTLs colocalized (coloc score > 0.9) with GWAS SNPs associated to different immune-related diseases, especially for inflammatory bowel disease (IBD), rheumatoid arthritis and type-1 diabetes. Many of the disease variants also colocalized with Treg chromatin accessible loci. Particularly, we observed 18 variants colocalizing with IBD SNPs, an example including the BACH2 locus where the IBD protective allele of the rs62408233 variant was located under the ATAC peak in the intron 1 of the BACH2 gene and resulted in reduced chromatin accessibility and decreased BACH2 expression. The caQTL peak regulated by the rs62408233 also overlapped with H3K27ac implicating that the SNP may be disrupting an active enhancer.

By mapping genetic effects regulating chromatin activity and gene expression in Tregs we elucidate the function of genetic variants associated to ADs and further build support for the role of Tregs in the development of immune-mediated diseases.

**BAM.IOBIO** - A VISUAL, REAL-TIME WEB-BASED ALIGNMENT FILE INSPECTOR ADAPTED FOR CLINICIAN USE

Megan Bowler[1,2], Chase Miller[1,2,3], Tonya DiSera[1,2], Alistair Ward[1,2,3], Gabor Marth[1,2,3]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]Universty of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]Frameshift Genomics, Boston, MA

***Iobio*** (http://iobio.io) provides real-time genomic analysis using immediate visual feedback to make understanding complex genomic datasets more intuitive and analysis more interactive. The first ***iobio*** web application developed, **bam.iobio** (http://bam.iobio.io), is an open-source dashboard application which samples the contents of sequent alignment (BAM) files and displays a series of metrics in real time, providing users with rapid feedback on the quality their alignments. This application is already immensely useful and heavily utilized by thousands of users. We recently introduced this tool into research-grade, rare-disease diagnostic studies by research clinicians and diagnostic molecular pathologists. Through extensive and critical use of these clinicians, we identified a number of areas where improvements were needed, in order to deploy this tool for the widest possible range of users, which include clinicians often without extensive training in genetics/genomics, users who have the most to gain from these tools but whose clinical duties often leave them very little time for research.

The most important feedback was the need tool features that allow the user not only to easily visualize key metrics of the BAM files (e.g. mean sequence coverage), but to present these metrics in the context of acceptable ranges to allow rapid critical evaluation of the data. We addressed this need with custom modals with each metric, allowing the user to interpret measured metrics in the appropriate context. We improved the ability of the user to meaningfully evaluate exome sequencing dataset. We improved the scaling of the charts to enhance visual identification of genome features (e.g. large genome deletions/amplifications). We improved auto-zooming functionality to focus the users of the most important regions of genome distributions (e.g. sequencing fragment length, sequencing depth). Finally, we have entirely re-implemented the ***bam.iobio*** code using the Vue.js JavaScript framework, to facilitate easier code maintenance and code reuse.

These improvements, along with some other minor enhancements, all lend to greater user understanding and application utility. The revamped ***bam.iobio*** app is currently being incorporated into a comprehensive ***iobio***-based clinical diagnostic tool suite.

# AN ANALYTICAL FRAMEWORK FOR WHOLE GENOME SEQUENCE DATA AND ITS IMPLICATIONS FOR AUTISM SPECTRUM DISORDER

Harrison Brand*, Donna M Werling*, Joon-Yong An*, Matthew R Stone*, Lingxue Zhu*, Joseph T Glessner, Ryan L Collins, Shan Dong, Ryan M Layer, Joseph D Buxbaum, Mark J Daly, Matthew W State, Aaron Quinlan, Gabor T Marth, Kathryn Roeder, Bernie Devlin†, Stephan J Sanders†, Michael E Talkowski†

SSC-ASC Genomics Consortium, MGH, Boston, MA

Genomic studies in autism spectrum disorder (ASD) have demonstrated a substantial contribution of *de novo* variation from large copy number variation (CNV) and protein-disrupting coding mutations from exome sequencing (WES). With the emerging availability of whole-genome sequencing (WGS), we can begin to explore the contribution of structural variation (SV) beyond large CNVs, as well as noncoding regulatory variation, which has been predicted to harbor much of the unexplained heritability in ASD. However, there are significant challenges with the evaluation and interpretation of *de novo* noncoding mutations, which represents two orders of magnitude greater mutational burden than the coding sequence, but without the triplet code to predict mutational impact. Moreover, no standardized approaches for analysis comparable to WES have been established. We analyzed WGS from 519 quartet families (n=2,076 genomes) with an ASD proband and unaffected sibling. We developed a novel variant discovery pipeline optimized for the detection of *de novo* SNV, InDels, and the entire spectrum of SV. Application of this pipeline observed 69 *de novo* SNVs/InDels per child, and a total of 171 *de novo* SVs. The pipeline also yielded 5,843 SVs per child. Variants were annotated using an extensive series of noncoding functional annotations and gene sets of plausible relevance to ASD, resulting in 51,801 combinations of annotation categories. ASD association within each category was assessed using a binomial test to compare variant counts in cases and controls in a Category-Wide Association Study (CWAS). To account for the multiplicity of hypotheses tested in such studies, correlations of p-values were assessed between the 51,801 categories from 20,000 sets of simulated variants. Eigenvalue decomposition estimated that 4,123 effective tests explained 99% of the variation. After appropriate correction for the number of hypothesis tests performed, no class of noncoding variation was within two orders of magnitude of the empirical significance threshold. By contrast, without appropriate correction, an equal number of biologically plausible associations were observed in both cases and controls. Thus, the relative risk of *de novo* noncoding variation is probably modest in ASD compared to *de novo* coding variants. Robust results from future WGS studies will require much larger cohorts and analytical strategies that consider the substantial multiple testing burden.

# HIGH-THROUGHPUT SEQUENCING DATA: A PROPOSAL OF PROCESSING PIPELINE FOR HUMAN POPULATION AND EVOLUTIONARY GENOMICS STUDIES

Gwenna Breton, Carina Schlebusch, Mattias Jakobsson

Human Evolution, Department of Organismal Biology, Evolutionary Biology Center, Uppsala, Sweden

High-throughput sequencing has rapidly become omnipresent in human genomics. Notably, it alleviates the problem of ascertainment bias and opens new possibilities for understanding human history: we are not longer limited to short sequences or to ascertained markers. It enables much better description of both single nucleotide polymorphisms and genome rearrangements. We are able to come closer to the true number of variants and heterozygosity estimates. Most SNP arrays are ascertained towards individuals of European descent and give a biased picture of the diversity in divergent populations, for example from Sub-Saharan Africa, making high-throughput sequencing data particularly promising for studying these populations.
In terms of evolutionary and population genetics studies, one drawback of high-throughput sequencing is the small sample sizes for high-coverage genomes. Thus, one might wish to combine new data with available comparative data – but combining data can be cumbersome. The data is often combined at the VCFs (variants files) stage, which are obtained after numerous processing steps starting at the sequencing machine output. This can introduce a number of biases due to: 1-the sequencing plate-form (e.g. Illumina, Complete Genomics); 2-the bioinformatics processing pipeline; 3-different coverages; 4-the filtering of variants. This makes it difficult to distinguish true biological signals from dataset biases.
We investigated genomes from African populations and analyzed them together with publicly available samples (Mallick et al 2015, The 1000 Genomes Project 2015, Meyer et al 2012, Choudhury et al 2017). In order to minimize dataset bias, we selected genomes based on strict criteria, including: Illumina paired-end data, minimum coverage of 20X, and access to the raw reads. We processed all samples with the same pipeline, which is based on the "GATK Best Practices for Germline SNPs + Indels" (DePristo et al 2011) and using the hg38 reference genome, but with two main modifications: 1-a two-steps BaseQualityScoreRecalibration procedure, involving variant calling on individual intermediate files to avoid penalizing variation not present in dbsnp and: 2-modification of the variant calling step to obtain an all sites VCF after joint genotyping.
First results show that our estimates of the number of variants fall within reported observations (The 1000 Genomes Project 2015, Choudhury et al 2017). The number of singletons in Bantu-speaking populations is inferior to the only comparative values (Choudhury et al 2017). Multi-dimensional scaling plot of the allele distance matrix show the expected distribution of worldwide populations. Future analyses include deeper description of genetic diversity and evaluation of the bias of some SNP arrays on basic diversity analyses.

FINE-MAPPING REGULATORY VARIANTS ACROSS 49 TISSUES.

A A Brown[1], F Hormozdiari[2], F Aguet[3], GTEx Consortium[3], E T Dermitzakis[1], X Wen[4]

[1]University of Geneva, Genetic Medicine & Development, Geneva, Switzerland, [2]Harvard, Epidemiology, Cambridge, MA, [3]Broad Institute, Boston, MA, [4]University of Michigan, Biostatistics, Ann Arbor, MI

Fine-mapping of cis-regulatory variants has enhanced our understanding of mechanisms of gene expression and the links between expression and disease. The latest GTEx data provided RNA-seq and WGS from 838 individuals across 49 tissues. We applied 3 fine-mapping methods to these data: CaVEMaN, CAVIAR and dap-g, finding 68,948 high confidence causal variants (P>0.8). The median no. of variants in a 95% credible set is 6. Fine-mapping approaches can also find evidence for independent eQTLs. We found 7.2% (kidney-cortex) to 45% (nerve-tibial) of eGenes have >1 eQTL, with differences mainly due to sample size ($\rho$=0.95).

Fine-mapping methods can be used to derive properties of regulatory variants; these properties could be used, e.g., for developing personalised genome interpretation methods to predict deleterious variants solely from genomic context. We annotated eQTLs to 715 ChIP-seq peaks and motifs for 54 transcription factors, finding greatest enrichment in peaks marking active promoters (H3K4me3, 16.5% of regulatory variants are in these regions, 372 fold enrichment (FE)). Motifs showed more enrichment (up to 1,100FE), but explained a far smaller proportion of regulatory variation (0.3%).

Moving beyond classifications from single annotations, we used the large catalogue of regulatory variants to predict consequences of combined annotations; producing tailored inference on specific variants. Assigning priors to variants (from 1-668 FE, IQR = 27.2), we find 32.2% of variants have unique prior weight due to unique genomic context. Using these priors significantly increased the causal probability of the lead variant (p<1e-316, median P=0.41 vs 0.27). In 58% of cases we found a different variant to be causal compared to only considering statistical information; these also had a higher replication rate in a new dataset (88.1% vs 87.6%).

We used the whole blood results to look for colocalization with GWAS hits for 4 lipid traits. We find 36 of 148 GWAS hits colocalized with eQTLs (P>50%), explaining 21 new GWAS hits compared to the previous GTEx release. This demonstrates the large increase in power due to the 332 increase in sample size and WGS genotypes.

For a majority of eQTLs there remains considerable uncertainty in the precise causal variant. Here we show how leveraging molecular phenotype associations, annotations and posterior probabilities can bring benefits to a diverse set of analyses, from studying properties of regulatory variants, to assessing allelic heterogeneity, to exploring the relationship with GWAS hits.

# IDENTIFYING CORE BIOLOGICAL PROCESSES DISTINGUISHING EYE TISSUES WITH SYSTEMS-LEVEL GENE EXPRESSION ANALYSES, WEIGHTED CORRELATION NETWORKS, AND SINGLE CELL RNA-SEQ

John M Bryan, Robert B Hufnagel, Brian P Brooks, David M McGaughey

National Eye Institute, Ophthalmic Genetics and Visual Function Branch, Bethesda, MD

The human eye has several specialized tissues which direct, capture, and pre-process information to enable vision. RNA-seq gene expression analyses have been used extensively, for example, to profile specific eye tissues and in large consortium studies, like the GTEx project, to study tissue-specific gene expression patterning. However, there has not been an integrated study of multiple eye tissues patterning with other human body tissues. We have collated current publicly available healthy human RNA-seq datasets and a substantial subset of the GTEx project RNA-seq datasets and processed all in a consistent bioinformatic workflow. We use this fully integrated dataset to probe the relatedness and biological processes between the cornea, retina, RPE-choroid complex, and the rest of the human tissues with differential expression, clustering, and GO term enrichment tools. We also leverage our large collection of retina and RPE-choroid tissues to build the first retina and RPE-choroid human weighted gene correlation networks and use them to highlight known biological pathways and eye gene disease enrichment. Additionally, we utilize a previously published single cell RNA-seq dataset of over 10,000 mouse retinal cells to perform further clustering analyses and assess the relationship between gene-level expression, network status, and cell type. These data, analyses, and visualization are available via a powerful web application (http://eyeIntegration.nei.nih.gov). We also explore the addition of non-human eye RNA-seq data, facilitated by machine learning classification of important tissue phenotypes like age and tissue type. We anticipate the incorporation of these analyses into the eyeIntegration web application will further support the ultimate aim of eyeIntegration serving as a catalyst for user-led gene discovery within the field of vision research.

# DISSECTING TRANSCRIPTOMIC SIGNATURES OF NEURONAL DIFFERENTIATION AND MATURATION USING iPSCs

EE Burke*[1], JG Chenoweth*[1], JH Shin[1], L Collado-Torres[1], S Kim[1], N Micali[1], Y Wang[1], RE Straub[1], DJ Hoeppner[1], D Hiler[1], KF Berman[2], JA Apud[2], AJ Cross[3], NJ Brandon[3], DR Weinberger[1], BJ Maher[1], RDG McKay+[1], AE Jaffe+[1]

[1]Lieber Institute for Brain Development, Baltimore, MD, [2]NIMH Intramural Research Program, Bethesda, MD, [3]AstraZeneca Neuroscience, IMED, Cambridge, MA

*Contributed equally; +Co-corresponding

**Background:** Human induced pluripotent stem cells (hiPSCs) are a unique model of neural differentiation and maturation, and potentially of neurodevelopmental disorders among patients ascertained long after etiological deficits occur. Many large-scale stem cell transcriptomics experiments surrounding corticogenesis have focused on rodent embryonic stem cells (ESCs), single hESC lines in bulk differentiating cells, single cells, or on the identity and quality of the initial iPSCs without data across differentiation.

**Methods:** After generating hiPSC lines and inducing neuronal differentiation, we performed RNA-seq on 165 samples and quantified expression of genes and their transcript features. We performed differential expression analyses between conditions, WGCNA to find networks of genes, and PCA to identify global transcriptional patterns. We lastly leveraged public bulk and single-cell human data across the lifespan with regression calibration models to directly compare these cellular systems to the human brain cell types and stages.

**Results:** We present a hiPSC transcriptomics resource on corticogenesis from 5 iPSC donor and 14 subclone lines across 9 timepoints over 5 broad conditions: self-renewal, early neuronal differentiation, neural precursor cells (NPCs), assembled rosettes, and differentiated neurons that were validated using electrophysiology. We identified widespread and robust changes in the expression of individual transcript features and their splice variants, gene networks, and global patterns of transcription. We demonstrated that co-culturing human NPCs with rodent astrocytes resulted in mutually synergistic maturation, and cell type-specific expression data can be extracted using only sequencing read alignments without potentially disruptive cell sorting. We lastly demonstrated that almost half (48%) of total gene expression in our neuronal cultures after 8 weeks of differentiation reflected signatures of neurons from the adult cortex, suggesting that a subpopulation of our neuronal cultures represent mature adult cortical neurons.

**Discussion:** These data suggest iPSC-derived neuronal cultures represent mixed maturationally heterogeneous populations of neurons that can be quantified using expression profiles. Expression trajectory visualization databases and statistical tools for decomposition expression profiles from RNA-seq data are available at: http://research.libd.org/libd_stem_timecourse.

# CHARACTERIZING HUMAN IMMUNE RESPONSE WITH IMPLICATIONS FOR UNDERSTANDING AUTOIMMUNE TRAIT ARCHITECTURE

Diego Calderon*[1], Michelle L Nguyen*[2], Anja Mezger*[3,4], Jessica V Ribado*[3], Arwa Kathira[3], Beijing Wu[3], Lindsey A Criswell[5], William J Greenleaf[3,6,7], Alex Marson[2,7], Jonathan K Pritchard[3,8,9]

[1]Stanford University, Program in Biomedical Informatics, Stanford, CA, [2]UCSF, Department of Microbiology and Immunology, San Francisco, CA, [3]Stanford University, Department of Genetics, Stanford, CA, [4]Karolinska Institutet, Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Stockholm, Sweden, [5]UCSF, Rosalind Russell/Ephraim P. Engleman Rheumatology Research Center, San Francisco, CA, [6]Stanford University, Department of Applied Physics, Stanford, CA, [7]Biohub, Chan Zuckerberg Initiative, San Francisco, CA, [8]Stanford University, Department of Biology, Stanford, CA, [9]Stanford University, Howard Hughes Medical Institute, Stanford, CA

A better understanding of complex patterns of cell and context-specific gene regulation within the immune system would lead to insights into the genetic architecture of autoimmune disorders, and potentially, new disease treatments. Thus, we collected ATAC and RNA-seq data from 25 differentiated cell types under resting and stimulated conditions from blood of 4 healthy individuals. We observed strong chromatin accessibility and transcriptional changes attributable to cell and stimulation-specific effects. Interestingly, stimulation-specific chromatin responses tended to be broadly shared within cell subsets and even across B and T cell lineages. To characterize disease-relevant regions of chromatin accessibility, we performed a GWAS enrichment analysis on 39 traits. Generally, GWAS from autoimmune disorders were enriched within regions more accessible in stimulated cells compared to open chromatin regions from resting counterparts, progenitor blood cells, and control tissues. Signal from certain immune traits was enriched in broadly shared stimulation-response peaks, such as rheumatoid arthritis, whereas other traits were associated with cell-specific chromatin peaks, e.g., systemic lupus erythematosus (SLE) and B cell accessible regions. In addition to general enrichments, we used our data to prioritize specific disease-relevant genetic variants. For example, with many cell types from each individual, we leveraged lineage-specific allelic imbalance ATAC-seq to fine-map the causal variant connecting the expression of *POLB* in B cells to SLE. Finally, we demonstrated that one would likely overlook such an association if only relying on commonly available functional information such as blood eQTLs. Overall, our results highlight the importance of cell and condition-specific functional information for understanding patterns of gene regulation and linking genetic variation to complex autoimmune disorders.

# A COMPREHENSIVE BENCHMARKING TOOLKIT FOR SEQUENCING DATA AND ANALYTICAL TOOLS

Andrew Carroll

DNAnexus, Science, Mountain View, CA

Understanding the quality and reproducibility of both sequencing technology and computational applications is essential to correct research and clinical use.
The recent generation of gold-standard genome datasets – such as the Genome in a Bottle (GIAB) and Platinum Genomes – has given greater understanding about the quality of sequencing and led to the development of new computational methods.
However, these benchmarks are based on specially contributed datasets of high-quality sequencing runs. We compare the quality of typical sequencing runs to this benchmark data and show that relying solely on these datasets to measure accuracy gives an incomplete picture that may give false confidence.
We present Readshift, a method capable of biased resampling of high-coverage Genome in a Bottle datasets that can be used to generated standard-coverage datasets at a desired quality level mirroring a given research project to allow determination of precision and recall for that set.
We use this method to profile multiple sequencing platforms (HiSeq2500, HiSeqX, NovaSeq) and bioinformatics tools (GATK, Sentieon, ISAAC, DeepVariant, and Freebayes) highlighting unforeseen behaviors in low quality data from certain platforms and tools.
We demonstrate that for certain applications – especially somatic variant calling and the detection of low frequency events the differences in quality between runs has a disproportionate impact. This is highly relevant to use cases which must consider candidate low-frequency events. We identify low quality runs that will consistently cause somatic callers to fail.
We detail the impact of PCR-free versus PCR+ samples across these methods, identifying certain approaches which appear more robust to PCR errors.
Finally, we present cases detailing how developers have used this data to improve their methods to handle these scenarios and discuss both pitfalls and caveats in current approaches as well as opportunities for improvement.
This analysis will help the community to understand how accurate the NGS results in their study are, to quantify the importance of high quality data, and to improve the accuracy and robustness of both sequencing and analytical methods.

# ABUNDANT GENOME STRUCTURAL VARIATION SHAPES HERITABLE PHENOTYPIC VARIATION IN *DROSOPHILA*

Mahul Chakraborty[1], J. J Emerson[1], Stuart J Macdonald[2], Anthony D Long[1]

[1]University of California Irvine, Department of Ecology and Evolutionary Biology, Irvine, CA, [2]The University of Kansas, Molecular Biosciences, Lawrence, KS

Large scale structural mutations (e.g. duplications, deletions, insertions, etc.) play pivotal roles in genome evolution and the genetic basis of diseases. Aligning high throughput short reads to a reference genome is the conventional approach to find structural variants (SV). Recent results from human and *D. melanogaster* show that such approaches miss 40-80% of SVs. To determine the functional and evolutionary significance of SVs, we resequenced the founder strains of the Drosophila Synthetic Population Resources (www.flyrils.org) using long read sequencing technology and constructed *de novo* assemblies for each strain. The completeness and contiguity of the assemblies are comparable to or better than the current release of the standard reference strain, with most the genome represented by contiguous sequences (contigs) measuring 20Mb or longer. Collectively, we discovered thousands of structural variants, including duplicates, transposon insertions, and inversions, many of which are evolving under natural selection. Additionally, comparison of our comprehensive SV map with candidate genes obtained from published QTL mapping studies employing the DSPR unveil segregating SVs at majority of the candidate genes. One such candidate gene for nicotine resistance consists of five SV alleles at a Cytochrome P450 gene (*Cyp28d1*), comprising tandem gene duplications and TE insertions. These results suggest that sizable proportion of the phenotypic variation of complex traits in Drosophila may be due to complex genome structural changes which is shaped by natural selection.

# SVCURATOR: AN APP TO VISUALIZE STRUCTURAL VARIANTS FOR CROWDSOURCING MACHINE LEARNING LABELED DATA

Lesley M Chapman[1], Noah Spies[1,2], Nancy F Hansen[3], Fritz Sedlazeck[4], Marc Salit[1,2]

[1]National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD, [2]National Institute of Standards and Technology, Material Measurement Laboratory, Palo Alto, CA, [3]National Human Genome Research Institute, Cancer Genetics and Comparative Genomics Branch, Rockville, MD, [4]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

Next generation sequencing (NGS) technologies are rapidly evolving. Yet, discordance exists amongst structural variant (SV) calls as a result of variance between NGS sequencing and analysis pipelines. Improvements in the accuracy of calling these difficult SVs is needed to enable confidence in clinical decision making. The central aim of the current study is to crowdsource labeled data in order to train machine learning classifiers to generate a high confidence list of large indels and SVs.

SVcurator is a Python flask based app that will allow users to manually curate insertions and deletions from the Ashkenazim Jewish Trio son (NIST RM 8391). The sites included in the app are a result of data collection and integration steps. Data was generated for an Ashkenazim Jewish mother-father-son trio (NIST RM 8392) from short, long, and linked read whole genome sequencing platforms. Members of the Genome in a Bottle (GIAB) Consortium generated over 1 million candidate large indels and SVs (>= 20bp) within this trio from 30+ informatics pipelines and 5 sequencing technologies. Of these there were over 500000 unique sequence resolved calls. These calls were merged using svanalyzer which compares sequence resolved calls and merges calls less than 20% different using 3 distance measures. After the calls were merged 36600 insertions and 37600 deletions were shown to be supported by either 2 or more technologies, 4 or more callers, or had BioNano or Nabsys support.

In a preliminary analysis, genotypes were determined heuristically based on read support for reference and alternate alleles using svviz - a structural variant realignment and visualization tool, and, after removing variants without read support, 34500 insertions and 32600 deletions remained in our v0.5.0 draft "straw man" benchmark set. To develop more robust, data-driven classifications of genotypes, machine learning will be used to further characterize these variants by predicting genotype class labels for a subset of these calls. SVcurator will be used to generate labeled data for the machine learning classifiers. Over 1000 insertions and deletions (>=50bp) were randomly sampled from the v0.5.0 benchmark set based on size, and will be included in the app. For each SV, an IGV image as well as svviz dotplot and read aligned images are included in the app to describe each variant. In addition, 4 questions regarding the genotype and accuracy of the variant are also included for each event. In future studies, we will use the crowdsourced labeled data for insertions and deletions to train machine learning models to generate a high confidence list of large indels and SVs.

# AN INTEGRATIVE MAP OF HUNDREDS OF DNA BINDING PROFILES AND DNA METHYLATION LANDSCAPE IN A SINGLE CELL TYPE

Surya B Chhetri[1,2], Christopher Partridge[1], Jeremy W Prokop[1], Ryne C Ramaker[1,3], Mark Mackiewicz[1], Barbara J Wold[5], Ali Mortazavi[4], Richard M Myers[1], Eric M Mendenhall[1,2]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, [2]The University of Alabama in Huntsville, Department of Biological Sciences, Huntsville, AL, [3]University of Alabama at Birmingham, Department of Genetics, Birmingham, AL, [4]University of California Irvine, Department of Developmental and Cell Biology, Irvine, CA, [5]California Institute of Technology, Division of Biology, Pasadena, CA

Genome wide identification and characterization of TF (Transcription Factor) binding sites at large scale is critical in understanding transcriptional regulatory networks. Thousands of factors bind to DNA, yet only a minority of them have been assayed by ChIP-seq, underscoring the gap in our understanding of regulatory networks. Here we present the analysis of 208 DNA binding profiles produced by the ENCODE Consortium in HepG2 cell type - representing 22% of all factors expressed within this cell type – which allow us to uncover novel gene regulatory insights, not possible from few factor binding maps. We identify novel DNA sequence-motifs, as well as discrepencies from previously described motifs. We used genome segmentation based on chromatin modifications to classify TFs based on their localization, and find 42.3% of the factors to be predominantly promoter-associated whereas 30.3% as predominantly enhancer-associated. Assessing cis-regulatory landscape of HepG2, we observe 12,928 putative cis regulatory regions to have 50-125 TFs bound. These highly bound regions have been previously termed High Occupancy Target (HOT) regions. We confirm that HOT regions are more evolutionarily constrained. Of note, using genome segmentation, 51% of HOT regions were identified as promoters whereas 47% as strong enhancers and only 1% as weak enhancers. Using co-binding analysis, we identified 3 novel TFs (FOXA3, SOX13, ARID5B) co-localizing with GATAD2A, which we propose for NuRD complex recruitment to active enhancer loci. Lastly, using whole-genome bisulfite data in HepG2 cells, we performed an integrative analysis to determine TF binding association of 208 factors with DNA methylation, and we show that TFs associated to methylated loci is largely limited to “weak or poised” enhancer states. In addition, we find a subset of TFs that can recognize both methylated and unmethylated DNA including TFs that show different DNA sequence motifs at bound loci with high vs. low DNA methylation levels. Altogether, here we present novel gene regulatory insights using 208 ChIP-seq maps in single cell type, HepG2, building a deeper and more complete picture of transcriptional gene regulation.

# EVOLUTION OF AN INTRATUMORAL ECOLOGY SUSCEPTIBLE TO SUCCESSIVE TREATMENT IN BREAST CANCER XENOGRAFTS

Hyunsoo Kim[1], Pooja Kumar[1], Francesca Menghi[1], Javad Noorbakhsh[1], Eliza Cerveira[1], Mallory Ryan[1], Qihui Zhu[1], Guruprasad Ananda[1], Joshy George[1], Henry Chen[2], Susan Mockus[1], Chengsheng Zhang[1], James Keck[2], R. Krishna Murthy Karuturi[1], Carol J Bult[3], Charles Lee[1], Edison T Liu[3], Jeffrey H Chuang[1]

[1]The Jackson Laboratory, Genomic Medicine, Farmington, CT, [2]The Jackson Laboratory, In Vivo Services, Sacramento, CA, [3]The Jackson Laboratory, Mammalian Genetics, Bar Harbor, ME, [4]UConn Health, Genetics and Genome Sciences, Farmington, CT

The processes by which tumors evolve are essential to the efficacy of treatment, but quantitative understanding of intratumoral dynamics has been limited. Although intratumoral heterogeneity is common, quantification of evolution is difficult from clinical samples because treatment replicates cannot be performed and because matched serial samples are infrequently available. To circumvent these problems we derived and assayed large sets of human triple-negative breast cancer xenografts and cell cultures from two patients, including 86 xenografts from cyclophosphamide, doxorubicin, cisplatin, docetaxel, or vehicle treatment cohorts as well as 45 related cell cultures. We assayed these samples via exome-seq and/or high-resolution droplet digital PCR, allowing us to distinguish complex therapy-induced selection and drift processes among endogenous cancer subclones with cellularity uncertainty <3%. For one patient, we discovered two predominant subclones that were granularly intermixed in all 48 co-derived xenograft samples. These two subclones exhibited differential chemotherapy sensitivity -- when xenografts were treated with cisplatin for 3 weeks, the post-treatment volume change was proportional to the post-treatment ratio of subclones on a xenograft-to-xenograft basis. A subsequent cohort in which xenografts were treated with cisplatin, allowed a drug holiday, then treated a second time continued to exhibit this proportionality. In contrast, xenografts from other treatment cohorts, spatially dissected xenograft fragments, and cell cultures evolved unsystematically but with substantial population bottlenecks. These results show that ecologies susceptible to successive retreatment can arise spontaneously in breast cancer in spite of a background of irregular subclonal bottlenecks, and our work provides to our knowledge the first quantification of the population genetics of such a system. Intriguingly, in such an ecology the ratio of common subclones is predictive of the state of treatment susceptibility, suggesting that this ratio can be measured to optimize dynamic treatment protocols in patients.

# GENETIC MAPPING OF UBIQUITIN-PROTEASOME SYSTEM ACTIVITY IN LARGE YEAST POPULATIONS.

Mahlon A Collins, Frank W Albert

University of Minnesota, Genetics, Cell Biology, and Development, Minneapolis, MN

The ubiquitin-proteasome system (UPS) is the primary pathway for cellular protein degradation. UPS protein degradation is essential for many biological processes, including gene expression regulation, removal of aged and damaged proteins, and the cellular response to stress. Cellular viability is thus critically dependent on the UPS and a variety of diseases are associated with impaired UPS function. UPS activity is influenced by the concerted action of many regulatory mechanisms, including synthesis, assembly, and post-translational modification of UPS components, as well as their interaction with other activity-modifying factors.
While many UPS regulatory mechanisms have been described, how they are shaped by genetic variation is largely unknown. In particular, there have been no systematic studies of how natural genetic variation affects UPS activity. This constrains our understanding of the complex genetic basis of biological processes and diseases influenced by the UPS.
To address these limitations, we are performing quantitative trait locus (QTL) mapping of UPS activity in a cross of two genetically divergent strains of the yeast Saccharomyces cerevisiae. We developed, built, and characterized a series of genetically encoded reporters of UPS activity consisting of the superfolder green fluorescent protein (sfGFP) attached to individual peptide tags that assay various forms of UPS activity, including ubiquitin-independent and -dependent UPS protein degradation. To control for genetic effects on reporter transcription and translation, we added the red fluorescent protein mCherry to each reporter and separated it from sfGFP with a viral 2A sequence. Reporter expression results in the stoichiometric production of a stable mCherry molecule and a separate, rapidly-degraded sfGFP molecule from the same mRNA transcript. With this reporter system, low sfGFP levels indicate high UPS activity and vice versa. For each reporter, we mate a laboratory yeast strain with the reporter to a vineyard strain and use a modified synthetic genetic array method to create millions of recombinant haploid progeny (segregants). UPS activity is measured in the segregants by fluorescence-activated cell sorting (FACS). We collected pools of 10,000 cells each from the 1% tails of the sfGFP distribution at constant mCherry levels by FACS, extracted DNA from each pool, and are currently performing pooled whole-genome sequencing. Allele frequency differences between the pools will indicate QTL that shape UPS activity. Our results will provide new mechanistic insights into the complex genetic basis of UPS function, as well as biological processes and diseases influenced by the UPS.

# GENETIC DIVERSITY AND POSITIVE SELECTION FOR FRUIT SHAPE IN A WIDE COLLECTION OF *CAPSICUM* SPECIES.

Vincenza Colonna[1], Nunzio D'Agostino[2], Erik Garrison[3], Roberto Sirica[1], Teodoro Cardi[2], Pasquale Tripodi[2]

[1]National Research Council, Institute of Genetics and Biophysics, Naples, Italy, [2]Research Centre for Vegetable and Ornamental Crops, Pontecagnano Faiano, Italy, [3]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Peppers of genus *Capsicum* are among the most important vegetable crops in terms of economic and nutritional importance, cultivated worldwide as a food and spice. The *Capsicum* genus originated in tropical and temperate areas in South America and has since spread around the world, giving rise to numerous cultivars of diverse phenotypes. Here we present the broadest and deepest study of *Capsicum*, using genotyping by sequencing to obtain 1.5M single nucleotide polymorphisms across 373 accessions of 11 species from 51 countries. We use this resource to study population structure and natural selection with a precision that gives new insight into the species complexes within the genus.

Population structure analysis recovers known species subdivisions and indicates very little genetic exchange among species. The only exception is *C. annuum*, which presents two distinct components that correlate with fruit size. We find that the group with largest fruit has lower effective population size than one with smaller fruit. Following this finding, we investigate positive selection between *C. annuum* peppers with small and large fruits on a genome-wide scale, uncovering putative loci associated with this trait.

We discovered that *C. pubescens* and *C. chacoense* are the species with the lowest genetic variation. Among all species, *C. pubescens* and *C. chacoense* have the lowest effective population size, the highest kinship and the longest regions in linkage disequlibrium. These results match observations from breeding experiments which demonstrate that these species are more difficult to cross with other clades and have consequently contributed relatively little to common cultivars of the genus. In contrast, *C. annum* is the species with the highest genetic diversity as indicated by the large use of the species in breeding schemes.

# TRANSLATING CANCER GENOMICS TO THE CLINIC, FOR ADVANCED CHILDHOOD AND RARE ADULT CANCERS

John Grady[1], Marie Wong[1], Marcel Dinger[1], Michelle Haber[2], David M Thomas[3], Mark J Cowley[1,2,3]

[1]Garvan Institute of Medical Research, Kinghorn Centre for Clinical Genomics, Sydney, Australia, [2]Children's Cancer Institute, Precision Medicine Program, Sydney, Australia, [3]Garvan Institute of Medical Research, Cancer, Sydney, Australia

Patients with rare cancers (less than 5 in 100,000), including all paediatric cancers account for 30% of cancer deaths. Genome screening of rare tumours offers the opportunity to substantially improve patient care, though identifying targetable mutations linked to targeted therapies tested in other cancers, assessing inherited cancer predisposition variants. In Australia, we are leading two ambitious national precision cancer genomics trials for adults with rare tumours, or kids with relapsed or poor-outcome tumours. Each trial uses genomic screening in real time to provide targeted treatment recommendations for patients. We have established molecular tumour boards, and the analytical (*refynr*, *Seave*), interpretational, and reporting frameworks (*gentian*) to make these trials feasible.

For adults with rare cancers, we have established the Molecular Screening and Therapeutics (MoST) program, where patients that have exhausted all treatment options are recruited, and DNA + RNA is screened with the Illumina TST170 panel. For each patient tumour, we determine tumour purity, ploidy, SNV, INDEL, CNV, fusions, tumour mutation burden, and microsatellite instability scores, which are presented at our molecular tumour board. To date we have screened >450 rare cancer patients, and enrolled over 100 onto either targeted treatment arms (eg Palbociclib or Olaparib) or dual immunotherapy, with additional treatment arms opening. We have identified at least one treatment recommendation in 70% of patients.

For children with relapsed, or high-risk cancer, we have established the Lions Kids Cancer Genome Project (LKCGP), which is using deep whole genome sequencing (WGS) to comprehensively characterise each tumour (SNV, INDEL, CNV, SV). Tumours from children enrolled on the Zero Childhood Cancer (ZCC) trial, are subject to additional molecular screening (targeted and RNA-Seq) in vitro drug screenspatient-derivedatient derived xenografts. In the pilot feasibility study we enrolled 59 patients, half of which had relapsed disease, and mostly CNS tumours (47%). We identified reportable SNVs, fusions and CNVs in 56%, 24% and 40% of patients, respectively. 5 patients had a reportable germline cancer predisposition variant and in 3 patients, the genomic findings changed the primary diagnosis. Overall, 57% of patients received a personalised medicine recommendation, and 36% of patients with a therapy recommendation have currently received that therapy. In at least two cases, the tumours have completely resolved using Larotrectinib, resulting from unexpected NTRK fusions. In Sept 2017 we launched a national multicentre trial, and have recruited all 40 patients since then.

Collectively the high rates of actionable variants identified in these rare cancers suggests an enormous potential for improving outcomes using existing targeted therapies developed for other cancers.

# HARNESSING LONGITUDINAL DATA TO DERIVE A NEW GENETIC RISK SCORE FOR CHILDHOOD OBESITY

Sarah J Craig[1,2], Junli Lin[3], Ana Kenney[3], Ian M Paul[2,4], Leann L Birch[5], Jennifer S Savage[6,7], Michele M Marini[7], Francesca Chairomonte[2,3], Matthew Reimherr[2,3], Kateryna D Makova[1,2]

[1]Penn State University, Biology Department, University Park, PA, [2]Penn State University, Center for Medical Genomics, University Park, PA, [3]Penn State University, Department of Statistics, University Park, PA, [4]Penn State College of Medicine, Department of Pediatrics, Hershey, PA, [5]University of Georgia, Department of Foods and Nutrition, Athens, GA, [6]Penn State University, Department of Nutritional Sciences, University Park, PA, [7]Penn State University, Center for Childhood Obesity Research, University Park, PA

The childhood obesity epidemic affects one in three children in the United States. Obesity has a complex etiology with 40 to 70% of the interindividual variability attributed to genetics. However, only 5% is explained by known genetic variants. We genotyped 225 first-born children from the INSIGHT study on the Affymetrix Precision Medicine Research Array. Anthropometric measurements were collected at five time points between birth and one year, and at two and three years. Genetic risk scores (GRS) were calculated as the sum of the SNPs weighted by their effect sizes on BMI. We found that GRS defined by SNPs identified in GWAS with adult obesity is not associated with growth patterns in the first three years after birth. Using functional data analysis techniques, we identified 24 SNPs associated with growth curves over the first three years after birth. Using these variants we calculated a novel GRS using allele effect sizes that exploits the longitudinal structure of the data. This new GRS is most predictive in the first one to two years after birth, after which the GRS based on adult obesity has been shown to be significant.

# ASSESSING BEHAVIOR AND ANXIETY IN THE *DHCR7*$^{T93M/\Delta3\text{-}5}$ MOUSE MODEL OF SMITH-LEMLI-OPITZ SYNDROME

Joanna Cross[1], Margaret Keil[2], Forbes Porter[2], Frances Platt[3]
[1]National Institutes of Health, National Institute of Mental Health, Bethesda, MD, [2]National Institutes of Health, National Institute of Child Health and Human Development, Bethesda, MD, [3]University of Oxford, Department of Pharmacology, Oxford, United Kingdom

Smith-Lemli-Opitz Syndrome (SLOS) is an autosomal recessive inborn error of cholesterol synthesis caused by mutation of the 7-dehydrocholesterol reductase (*DHCR7*) gene. This results in abnormal sterol levels; increased 7-dehydrocholesterol and typically decreased cholesterol. Although SLOS has a characteristic physical phenotype, with the most common finding being 2-3 toe syndactyly, there are also multiple behavioral abnormalities. These include cognitive deficits, anxiety, hyper-activity, sleep cycle disturbance, language impairment and autism spectrum behaviors. There are currently two main mouse models of SLOS; a homozygous null, *Dhcr7*$^{\Delta3\text{-}5/\Delta3\text{-}5}$, and a model combining the null mutation and the common p.T93M missense mutation, *Dhcr7*$^{T93M/\Delta3\text{-}5}$. Unlike the null model, the hypomorphic *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice can live to adulthood and are therefore suitable for behavioral studies. Anxiety can be measured via multiple commonly used protocols; elevated plus maze, open field test and assessing burrowing and nesting behavior.

Although there was variation in all genders and genotypes, overall, the *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice were marginally more likely to have increased burrowing compared to controls. However, this was only significant at 4 months in the overnight study. This could be showing the increased anxiety phenotype or hyperactivity, which is also noted in the increased gait speed of the *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice in the open field test. In contrast, the Nestlet test and elevated plus maze are both suggestive of decreased anxiety in the *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice. As this is opposite to what is typically observed in patients, these findings are either the result of a secondary defect or suggest that the *Dhcr7*$^{T93M/\Delta3\text{-}5}$ model does not replicate the behavioral traits of patients.

Histological analysis showed that *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice tended to have a smaller hippocampus than age-matched controls. The hippocampus is the part of the brain responsible for learning, memory and spatial awareness. As such, the changes seen in the hippocampus could explain the decreased shredding observed in the Nestlet test and the potential delayed learning of the *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice, observed by the normalisation of the behaviours in the elevated plus maze and open field tasks. This could also provide an explanation for the decreased anxiety phenotype observed in the elevated plus maze test as the *Dhcr7*$^{T93M/\Delta3\text{-}5}$ mice may have altered special awareness and may not be fully aware of the elevation or lack of walls on the open arms. Additionally, the anterior commissure was smaller or absent in a pilot study of female *Dhcr7*$^{T93M/\Delta3\text{-}5}$ compared to control mice. These results are suggestive of subtle structural defects and justify further analysis, particularly using mice with an isogenic background where the differences may be more distinct. As SLOS does not have an optimal treatment pathway, if the results suggested by these tests are accurate, it could help aid the development of therapeutic interventions.

# CHARACTERIZING THE REPLICABILITY OF CELL TYPES DEFINED BY SINGLE CELL RNA-SEQUENCING DATA USING METANEIGHBOR

Megan Crow, Anirban Paul, Sara Ballouz, Z. Josh Huang, Jesse Gillis

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

Single cell RNA-sequencing technology (scRNA-seq) provides a new avenue to discover and characterize cell types, but the experiment-specific technical biases and analytic variability inherent to current pipelines may undermine its replicability. Cross-dataset comparison is further hampered by the use of *ad hoc* naming conventions. To address this we developed MetaNeighbor, a tool that quantifies the degree of cell type replicability across datasets and enables rapid identification of clusters with high similarity. We benchmark MetaNeighbor by assessing the replicability of neuronal identity, comparing results across eight technically and biologically diverse datasets to define best practices for more complex assessments. We then use MetaNeighbor to evaluate novel interneuron subtypes, finding that 24/45 subtypes have evidence of replication. Finally, we demonstrate its wide applicability across tissues, technologies and species, including use cases from mouse bipolar neurons and human pancreas. Across all tasks we find that large sets of variably expressed genes can identify replicable cell types with high accuracy, suggesting a general route forward for large-scale evaluation of scRNA-seq data.

# UNDERSTANDING VARIATION IN HUMAN RETINAL MORPHOLOGY USING UK BIOBANK DATA

Hannah Currant[1], Tomas Fitzgerald[1], Anthony P Khawaja[2], Pearse A Keane[2], Charles A Reisman[3], Qi Yang[3], Peng T Khaw[2], Paul J Foster[2], Praveen J Patel[2], Ewan Birney[1], the UK Biobank Eye and Vision Consortium[4]

[1]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute, Cambridge, United Kingdom, [2]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, University College London, London, United Kingdom, [3]Topcon Healthcare Solutions, Research and Development, Oakland, NJ, [4]1 Members of the UK Biobank Eye and Vision Consortium, are listed before References, http://www.ukbiobankeyeconsortium.org.uk, United Kingdom

The macula is the layered area of the human retina responsible for detailed vision, with the foveal pit being a central sub-region responsible for highest clarity sight. It has a distinct valley-like morphology, and damage to it will severely affect central vision, with changes to foveal shape seen in several conditions including albinism and diabetes. The macula is routinely imaged in the clinic using Optical Coherence Tomography (OCT) imaging producing a high resolution, 3-D representation of the central macula. The UK Biobank has performed OCT scans for over 60,000 individuals, and contains a wide range of anthropometric, lifestyle and clinical measures as well as high quality genotyping data, leading to a rich opportunity to understand the variation and impact of retinal morphology.

The OCT data are inherently multi-dimensional, with even the crudest of quantitative transformations leading to 9 dimensions, and conceptually into the hundreds of thousands of dimensions. We are exploring different ways to capture information from OCT scans followed by different styles of dimensionality reduction, using a variety of methods, some of which utilise genetics in the dimensionality reduction and others which are free of genetic priors.

We have performed initial analysis and quality control of the OCT data. The OCT data are transformed to a series of quantitative metrics using industry-standard software. However there can be both technical measurement issues and extreme retinal phenotypes which prevent the sensible comparison of these metrics. We have created an initial filtering scheme to remove outlier data, and are exploring initial proof-of-concept genetic associations with dimensionality reduction.

# QUANTITATIVE TRAIT META-ANALYSIS IDENTIFIES RARE NONCODING VARIANTS IN *DENND1A* ASSOCIATED WITH ALTERED HORMONE LEVELS IN PCOS

Matthew Dapas[1], Ryan Sisk[1], Richard S Legro[2], Margrit Urbanek[1], Andrea Dunaif[3], M Geoffrey Hayes[1,4]

[1]Northwestern University Feinberg School of Medicine, Department of Medicine, Chicago, IL, [2]Penn State College of Medicine, Department of Obstetrics and Gynecology, Hershey, PA, [3]Icahn School of Medicine at Mount Sinai, Department of Medicine, New York, NY, [4]Northwestern University, Department of Anthropology, Evanston, IL

Polycystic ovary syndrome (PCOS) is a complex genetic disorder characterized by a combination of hyperandrogenism, chronic anovulation, and polycystic ovarian morphology. It affects up to 15% of premenopausal women worldwide. PCOS is the leading cause of anovulatory infertility and a major risk factor to type 2 diabetes. A number of susceptibility loci have been reproducibly mapped for PCOS in genome-wide association studies (GWAS). The risk alleles identified to date confer only modest increases in disease risk and can account for only a small proportion of the estimated genetic heritability of PCOS. To test whether rare genetic variants can account for this missing heritability, we performed whole genome sequencing on DNA from 76 families with one or more daughters affected with PCOS.

Variants were filtered for allele frequency (minor allele frequency ≤2%), call quality, consistency with Mendelian inheritance, and predicted deleteriousness according to evolutionary conservation and functional genomic data. Associations between sets of rare variants and PCOS and its quantitative hormonal traits were assessed using sequence kernel association tests, grouping variants at the gene-level (including 3' UTR, introns, and 7.5kb upstream of 5' TSS), accounting for relatedness, and adjusting for age and BMI. Quantitative trait associations were combined into a single test statistic using a modified Fisher's method for correlated traits.

After correcting for multiple testing ($P_c$), an association with altered reproductive and metabolic trait levels was found for 30 rare, non-coding variants in *DENND1A* ($P=2.71\times10\text{-}5$, $P_c=0.015$). DENND1A is involved in clathrin-mediated endocytosis. It is highly expressed in androgen-producing tissues. Multiple genome-wide association studies and meta-analyses have previously found associations between common variants in *DENND1A* and PCOS. Overexpression of the alternatively spliced, truncated *DENND1A* isoform DENND1A.V2 produces a PCOS phenotype in theca cells. However, the common GWAS variants in *DENND1A* do not directly affect splicing.

Our results indicate that rare noncoding variants in *DENND1A* contribute to elevated androgen levels in PCOS. These findings support the hypothesis that *DENND1A* plays a key role in the development of PCOS. This study also demonstrates how quantitative trait meta-analysis can be a powerful approach in rare variant association testing. Future studies will include prioritizing variants based on *in silico* prediction of functionality and assessment of pathogenicity in appropriate cell systems.

# INTERACTIONS BETWEEN THE GUT MICROBIOME AND HOST GENE REGULATION SHED LIGHT ON THE PATHOGENESIS OF COLORECTAL CANCER IN CYSTIC FIBROSIS PATIENTS

Gargi Dayama*[1], Sambhawa Priya*[1], Alexander Khoruts[2], Ran Blekhman[1]

[1]University of MN, Genetics, cell biology and development, Minneapolis, MN, [2]University of MN, Divison of Gastroenterology Hepatology and Nutrition, Minneapolis, MN

*authors contributed equally

Cystic Fibrosis (CF) is the most common autosomal recessive genetic disease in Caucasians. It is caused by mutations in the CFTR gene, leading to poor hydration of mucus and impairment of the respiratory, digestive, and reproductive organ functions. Advancements in medical care have lead to markedly increased longevity of patients with CF, but new complications have emerged, such as early onset of colorectal cancer (CRC). Although the pathogenesis of CRC in CF remains unclear, altered host-microbe interactions might play a critical role. We find that 1544 host genes, including CFTR, show differential expression in CF patients relative to healthy controls. Interestingly, we find that these genes are enriched with functions related to gastrointestinal cancer, such as metastasis of CRC, tumor suppression, cell growth, cell proliferation and apoptosis. Here, we characterize the changes in the gut microbiome and host gene expression in colonic mucosa of CF patients relative to healthy controls. We find that CF patients show decreased microbial diversity, decreased abundance of taxa such as Bilophila and Bacteroides, and increased abundance of other taxa, such as Blautia and Dorea. Lastly, we identify functional host-microbe interactions by modeling associations between gut microbiota abundances and host gene expression, revealing cancer-related associations between specific microbes and host genes in the gut. Our results provide new targets for potential treatment and therapeutic research for improving patient outcomes in CF.

# INTRA- AND INTER-CHROMOSOMAL CHROMATIN INTERACTIONS MEDIATE GENETIC EFFECTS ON GENE EXPRESSION

O Delaneau[1], M Zazhytska[2], C Borel[1], D Marbach[4], S Bergmann[4], P Bucher[3], S Antonarakis[1], A Reymond[2], E Dermitzakis[1]

[1]Univ. of Geneva, Dpt of Genetic Medicine and Development, Geneva, Switzerland, [2]Univ. of Lausanne, CIG, Lausanne, Switzerland, [3]EPFL, SIECR, Lausanne, Switzerland, [4]Univ. of Lausanne, Dpt of Computational Biology, Lausanne, Switzerland

Genome-wide studies on the genetic basis of gene expression and the structural properties of chromatin have considerably advanced our understanding of the function of the human genome. However, it remains unclear how structure relates to function and, in this work, we aim at bridging both by assembling a dataset that combines the activity of regulatory elements (measured by ChIP-seq for H3K4me1, H3K4me3 and H3K27ac), expression of genes (RNA-seq) and genetic variations of 317 European individuals across two cell types (Lymphoblastoid Cell Lines and Firboblasts).
First, we show that the regulatory activity is structured in 12,583 Cis Regulatory Domains (CRDs) that are reflective of the local chromatin organization into Topologically Associating Domains (TADs). Our work suggests that TADs mix together active (realized) and inactive (potential) subdomains with the most active ones corresponding to CRDs. In addition, we also find 25,315 significant associations between CRDs located on distinct chromosomes that form 200 Trans Regulatory Hubs (TRHs). These TRHs reflect the global chromatin organization into A/B nuclear compartments.
Second, we show that CRDs and TRHs essentially delimit the sets of active regulatory elements controlling expression for most genes (=82.4%), vary substantially across cell types and are key factors involved in the cis and trans co-expression of genes.
Third, we show that CRDs are under strong genetic control. We discovered 58,968 chromatin peaks affected by nearby genetic variants (cQTLs) which is, to our knowledge, the largest collection of cQTLs assembled so far. At the CRD level, this converts into genetic control of the activity of 6,157 CRDs and the structure of 110 CRDs.
Finally, we show that CRDs and TRHs capture complex regulatory networks along which the effects of eQTLs are propagated and combined to affect gene expression. In practice, we estimated that 75% of the eQTLs also affect the activity of CRDs and described three types of genetic effects that can be mediated by CRDs: long range eQTLs, rare eQTLs and trans eQTLs. Specifically, we find 33 genes with expression being perturbed by the accumulation of rare variants within CRDs and 9 well-replicated trans eQTLs acting on gene expression through direct inter-chromosomal chromatin interactions.
Overall, our study reveals the complexity and specificity of cis and trans regulatory networks and their perturbation by genetic variations.

# MOTIF ELUCIDATION IN CHIP-SEQ DATASETS WITH A KNOCKOUT CONTROL

Danielle Denisko[1,2], Coby Viner[2,3], Michael M Hoffman[1,2,3]

[1]University of Toronto, Medical Biophysics, Toronto, Canada, [2]Princess Margaret Cancer Centre, Research, Toronto, Canada, [3]University of Toronto, Computer Science, Toronto, Canada

**Introduction.** Many transcription factors show affinities for particular DNA sequences, enabling sequence specificity of transcriptional control. Chromatin immunoprecipitation sequencing (ChIP-seq) is a technique used to elucidate transcription factor binding sites. Multiple steps of the protocol can introduce noise and mask true binding signals, such as antibody cross-reactivity with off-target transcription factors *in vivo*. To mitigate noise, we can perform an additional control ChIP-seq experiment in which the transcription factor of interest has been knocked out. The knockout dataset can be used as a negative set to complement the wild-type set in downstream analyses, since sequences found in this experiment are not specific to the transcription factor of interest. Distinct approaches in processing these paired datasets have been proposed, but they have not yet been thoroughly compared.

**Methods.** We tested two pipelines of differential analyses and developed a method to combine them. The first pipeline, knockout implemented normalization (KOIN), incorporates differential peak calling through MACS; the knockout set is used to parametrize a background distribution model. The second pipeline, differential MEME-ChIP, uses the knockout set for differential *de novo* motif elucidation. Both pipelines use CentriMo to infer direct DNA binding via motif centrality and rank motifs according to their p-values. We combine these pipelines by extracting significant peaks from CentriMo and calculate a new score to rank motifs. The score reports the proportion of peaks in the wild-type set that overlap peaks in the KOIN set after subtracting peak regions that are common to both wild-type and knockout sets.

**Results.** We used four previously published wild-type and knockout ChIP-seq datasets of factors GATA3, SRF, OCT4, and KLF4. In general, both pipelines ranked motifs with variable success, in relation to parameters such as the relative number of peaks in knockout datasets. Our intersection method improved motif ranking of the transcription factor of interest in GATA3 and SRF datasets and maintained an optimal ranking in the OCT4 dataset relative to each pipeline alone. We are currently investigating complementary approaches to improving motif rankings by developing new statistical metrics that incorporate effect size. Overall, our work will provide a standardized method for wild-type/knockout paired motif elucidation in ChIP-seq datasets, improving our ability to detect and exclude spurious motifs that might otherwise lead to costly experimental follow-ups.

# PRIORITIZATION OF DELETERIOUS VARIANTS IN THE REGULATORY GENOME BY MODELING 3D CHROMATIN STRUCTURE, GENOME ESSENTIALITY AND ALLELIC EXPRESSION

Alex Wells[1,2], Pejman Mohammadi[3,4], David Heckerman[2], Tuuli Lappalainen[4], Amalio Telenti*[3], Julia di Iulio*[2,3]

[1]Stanford University, 94305, Stanford, CA, [2]Human Longevity Inc., 92121, San Diego, CA, [3]The Scripps Research Institute, 92037, La Jolla, CA, [4]New York Genome Center, 10013, New York, NY

*Correspondence: jdiiulio@scripps.edu, atelenti@scripps.edu

Understanding the consequences of variation in the non-coding genome is crucial. A number of non-coding scoring methods have been recently developed on the basis of biochemical, evolutionary and population genetics information. Here, we aim at providing additional inroads to the identification of functional domains in the non-coding genome, by incorporating three additional sources of information that have not been taken into consideration in the past: 3D chromatin structure, genome essentiality and gene regulatory constraint.

We trained a Random Forest model to differentiate between pathogenic and control genomic positions using 123 features, including (i) existing non-coding deleteriousness scores (e.g.: ncEigen, FATHMM, FunSeq2, GERP, LINSIGHT and ReMM), (ii) genome essentiality (e.g.: CDTS, pLI, and gene regulatory constraint from ANEVA), and (iii) biochemically detected genomic interactions (e.g.: 3D chromatin structure data). The model was trained on 1225 non-coding pathogenic variants from ClinVar and HGMD, and 8093 control variants. Controls were common variants matched for the distance to nearest splice site and genomic elements. Model parameters were tuned using 5-fold cross validation, and the performance was assessed on three independent test sets of non-coding pathogenic variants (N=367 manually curated set, N=156 from HGMD outside of non-coding RNA genes and N=183 from HGMD inside non-coding RNA genes).

The final model achieved 88% ROC-AUC and 71% PR curve in cross-validation and outperforms other metrics by at least 11% ROC-AUC and 29% in PR curve. The independent test sets yielded a ROC-AUC ranging from 82% to 95% and a PR curve ranging from 57% to 87%. The top 20 most important features in the model included seven deleteriousness metrics, five gene essentiality features and two 3D chromatin structure features. On the test sets, the ROC-AUC goes down to 78% when removing gene essentiality features and expression data, and 80% when removing 3D structure features emphasizing the importance of those contributions.

The use of new information sources on genome structure, human constraint and functional essentiality increases the precision in predicting the impact of genetic variations at nucleotide resolution. This new metric – ncE score – can guide functional experimentation and prioritization of variants for clinical diagnosis in the non-coding genome.

# THE EVOLUTIONARY DYNAMICS OF microRNAs IN DOMESTIC MAMMALS

Luca Penso-Dolfin, Simon Moxon[2], Wilfried * Haerty, Federica Di Palma

Earlham Institute, Norwich, United Kingdom, [2]University of East Anglia, Norwich, United Kingdom

MicroRNAs are crucial regulators of gene expression found across both the plant and animal kingdoms. While the number of annotated microRNAs deposited in miRBase has greatly increased in recent years, few studies provided comparative analyses across sets of related species, or investigated the role of microRNAs in the evolution of gene regulation.
We generated small RNA libraries across 5 mammalian species (cow, dog, horse, pig and rabbit) from 4 different tissues (brain, heart, kidney and testis). We identified 1676 miRBase and 413 novel microRNAs by manually curating the set of computational predictions obtained from *miRCat* and *miRDeep2*.

Our dataset spanning five species has enabled us to investigate the molecular mechanisms and selective pressures driving the evolution of microRNAs in mammals. We highlight the important contributions of intronic sequences (366 orthogroups), duplication events (135 orthogroups) and repetitive elements (37 orthogroups) in the emergence of new microRNA loci.
We use this framework to estimate the patterns of gains and losses across the phylogeny, and observe high levels of microRNA turnover. Additionally, the identification of lineage-specific losses enables the characterisation of the selective constraints acting on the associated target sites.

Compared to the miRBase subset, novel microRNAs tend to be more tissue specific. 20 percent of novel orthogroups are restricted to the brain, and their target repertoires appear to be enriched for neuron activity and differentiation processes. These findings may reflect an important role for young microRNAs in the evolution of brain expression plasticity.

Many seed sequences appear to be specific to either the cow or the dog. Analyses on the associated targets highlight the presence of several genes under artificial positive selection, suggesting an involvement of these microRNAs in the domestication process.

Altogether, we provide an overview on the evolutionary mechanisms responsible for microRNA turnover in 5 domestic species, and their possible contribution to the evolution of gene regulation.

# GENE.IOBIO - AN INTERACTIVE TOOL FOR REAL-TIME VARIANT INTERROGATION AND DISCOVERY

Tonya L Di Sera[1,2], Chase A Miller[1,2,3], Alistair Ward[1,2,3], Matt Velinder[1,2], Yi Qiao[1,2], Gabor Marth[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]Frameshift Genomics, Boston, MA

Identifying causative variants in genetic diseases relies on the expertise of clinicians and diagnostic analysts who possess detailed knowledge of disease presentation, clinical phenotype, and family history. However, current state-of-the-art bioinformatics tools are complex UNIX command line tools that often need to be run on large compute clusters and produce text-based output files with complex formats. To address these challenges we have developed **gene.iobio** (http://gene.iobio.io). Within this web based application we provide variant prioritization and interrogation for research investigation and discovery in real-time using informative and engaging visualizations.

To start work, the analyst specifies the sequence alignment and variant files for the proband or trio, including affected or unaffected siblings when available. The analyst then compiles a comprehensive list of genes associated with the phenotype or disorder, generated in gene.iobio using Phenolyzer, a phenotype-to-gene search tool, or in our new **genepanel.iobio** tool (http://genepanel.iobio.io), a companion app for gene list generation using the Genetic Testing Registry (GTR). Next, **gene.iobio** analyzes the variants in each defined gene, assessing functional impact and predicted protein impact (VEP), clinical significance (ClinVar), allele frequency (gnomAD), and inheritance mode, not only for the canonical transcript, but for all catalogued transcripts of each gene.

These analyses are performed in real-time, returning a list of candidate variants in a matter of minutes. By clicking on a gene, the user sees the variants in the same coordinate space as the sequence coverage area chart and the transcript model which highlights any exons showing insufficient coverage. Variants can also be called on-demand from alignment files with our Freebayes variant caller tool integrated into **gene.iobio** to see if less restrictive filtering or a priori consideration of ClinVar variants detects additional variants. Reported ClinVar variants in the gene under examination can be shown to survey for similar phenotypes in close proximity to a variant of interest, or to understand if particular exons/domains are enriched with pathogenic variants. Advanced, ad hoc filtering can be performed on any variant annotation and the gene list can be dynamically expanded to a larger search space. Recent performance improvements permit large gene lists (over 300 genes) to be analyzed in minutes. These intuitive visualizations coupled with real-time, iterative analysis places in-depth, nuanced variant interrogation directly in the hands of the clinician.

# IDENTIFICATION AND EXCLUSION OF PROBLEMATIC REGIONS IN THE GENOME ASSEMBLY OF THE LEEDS MELANOMA COHORT CNV DATA

Joey Mark S Diaz[1], Alastair Droop[2], Julia Newton-Bishop[1], David Timothy Bishop[1]

[1]Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom, [2]Leeds MRC Medical Bioinformatics Centre, Leeds, United Kingdom

There are regions in the human genome which are difficult to characterize because of incompleteness and errors in the human reference genome, multicopy or repeated sequence, polymorphisms, differences in sample quality, and deviations in the sequencing procedures[1]. Quantification of these regions tend to produce unreliable information by introducing noise to the data. Some studies addressed this problem and demonstrated how these regions were identified in the human genome assembly GRCh37 or earlier. The usual process is to exclude these regions from follow up analysis by creating a "blacklist". Our copy number data were derived from NGS output aligned against GRCh38 human reference. This requires identification of blacklist using this build. Lifting over of blacklisted region from earlier build to a newer one was explored though this does not include updated information from the newer genome build.
There were 33,096 10k windows initially identified as blacklisted in the whole genome based upon an earlier method (Pickrell al., 2001) [2]. A list of modelled centromeres and heterochromatins obtained from https://www.ncbi.nlm.nih.gov/grc/human and a list of gaps in the human genome taken from http://genome.ucsc.edu/cgi-bin/hgTables were considered as an initial part of the blacklisted regions denoted as Centrogaps. Residual filter based QDNAseq pipeline was used to identify windows that are highly variable in the genome based on the 312 control samples that met our criteria from the 1000 Genomes Project [1]. Considering only the autosomal genomes, Centrogaps identified 17,904 10k windows while residual filter identified 35, 856 10k windows to be included in the blacklisted areas. All windows in the Centrogaps list were also found in the list based on residual filter. A total of 38, 215 unique 10k windows from our dataset, including the 2, 359 unique windows from the earlier method were considered as the final blacklist accounting for 13 % of the autosomal genome.

1. Scheinin, I., et al., DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res, 2014. 24(12): p. 2022-32.
2. Pickrell, J.K., et al., False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics, 2011. 27(15): p. 2144-6.

# PROFILING THE LANDSCAPE OF TRANSCRIPTION, CHROMATIN ACCESSIBILITY AND CHROMOSOME CONFORMATION OF CATTLE, PIG, CHICKEN AND GOAT GENOMES [FAANG PILOT PROJECT "FR-AgENCODE]

Sylvain Foissac[1], Sarah Djebali[1], Andrea Rau[2], Sandrine Lagarrigue[3,4], Hervé Acloque[1], the FR-AgENCODE group[1], Elisabetta Giuffra[2]

[1]GenPhySE, INPT, ENVT, INRA, Université de Toulouse, Animal Genetics, Castanet-Tolosan, France, [2]GABI, AgroParisTech, INRA, Université Paris Saclay, Animal Genetics, Jouy-en-Josas, France, [3]PEGASE, INRA, Animal Genetics, Rennes, France, [4]PEGASE, AGROCAMPUS OUEST, Animal Genetics, Rennes, France

Functional annotation of livestock genomes is a critical and obvious next step to derive maximum benefit for agriculture, animal science, animal welfare and human health. From each of 2 males and 2 females per species (pig, cattle, goat, chicken), chromatin accessibility ATAC-seq assays and strand-oriented RNA-seq were performed on liver tissue and on 2 T-cell types (CD4 and CD8). Chromosome Conformation Capture (*in situ* Hi-C) was also carried out on liver. Sequencing reads from the 3 techniques were processed using standard pipelines and differential analyses of chromatin accessibility and gene expression were performed.

Although many ATAC-seq peaks were located in intronic and intergenic regions, most ubiquitous and open ones were located close to transcription start sites (TSS), thus supporting TSS prediction. By performing a differential analysis between tissues and using predicted transcription factor binding sites (TFBS), we found a higher TFBS density in differential ATAC-seq peaks compared to non differential ones, thus supporting a stronger likelihood of their being regulatory. Despite the small number of assayed tissues, 60% of annotated genes were found to be expressed, and tens of thousands of novel transcripts were detected, most forming new isoforms of annotated genes, and some forming completely new genes.

Based on the expression of the 9461 genes found to be ortholog among the 4 species, hierarchical clustering of samples indicated greater similarity in hepatic samples across species than between liver and T-cell samples within species, suggesting that the regulations underlying tissue differentiation (T-cell versus liver) are much stronger than those responsible for speciation. By requiring an RNA-seq transcript model to have an ATAC-seq peak close to its TSS and a polyA cluster close to its transcription termination site, we were able to define ~15,000 "golden" transcripts per species, the vast majority of which were coding (80%) and formed novel isoforms of annotated genes (60%). Interestingly, correlations between gene expression and promoter accessibility across samples were skewed towards both positive and negative values, suggesting distinct regulatory mechanisms of gene expression.

Using 40kb-resolution interaction maps generated with the Hi-C data, we identified topologically-associating domains and active "A" versus inactive "B" compartments, which were characterized by significantly different gene density and chromatin accessibility, therefore showing a high degree of consistency among our 3 kinds of data.
In summary, we present here an overview of the first multi-species and -tissue annotations of chromatin accessibility and genome architecture related to gene expression for farm animals.

UNDERSTANDING GENE REGULATION VIA INTEGRATION OF MULTI-OMICS DATA IN HUMAN TISSUES.

Alexander Dobin, Thomas R Gingeras, ENCODE/EN-TEx Consortium

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Identifying regulatory regions and quantifying their effect on gene expression is one of the major challenges for Genomics, imperative for a comprehensive understanding of gene regulation mechanisms. Functional annotation of the non-coding genome is also crucial for the interpretation of disease/trait associated variants identified in Genome-Wide Association Studies, since the majority of the GWAS variants are found in the non-coding regions, and thus affect the gene expression rather than protein function.

In this presentation, we describe the integration of ChIP-seq and RNA-seq data from the EN-TEx project, a collaboration between ENCODE and GTEx consortia, which produced multiple *omics datasets for 20 tissues from 4 donors. First, the ChIP-seq signals from several activation (H3K4me1, H3K4me3, H3K27ac, H3K36me3) and repression (H3K9me3, H3K27me3) chromatin marks, Pol2, EP300, CTCF transcription factors and DNase hypersensitivity are used for unsupervised segmentation of the genome into multiple chromatin states, creating spatially resolved combinatorial regulatory maps in a tissue-specific manner. The states are annotated as various types (active, inactive, bivalent) of promoters, enhancers, and transcripts, using prior knowledge as well as overlap with known genomic features. Next, gene expression is quantified in multiple tissues using RAMPAGE (promoter-specific RNA sequencing) and correlated with the ChIP-seq signal on the chromatin segments to further refine the regulatory maps. Unlike the standard RNA-seq assay, which quantifies the expression of entire transcripts/genes, RAMPAGE detects and quantifies transcription start sites, thus allowing for direct comparison between ChIP-seq and RNA signals on promoters. The active promoter regions defined in the unsupervised segmentation of ChIP-seq data are strongly enriched for RAMPAGE (promoter) peaks, attesting to the high fidelity of chromatin segmentation. We build a neural network model for predicting the promoter expression from the ChIP-seq signals and find that ~65% of the promoter expression variation can be explained by the ChIP-seq signal variation, with the main predictors being H3K4me3, H3K4me1, and Pol2. Finally, tissue-specific functional genome segmentations are used for annotating the variants contributing to variation in gene expression (eQTLs identified by the GTEx consortium), and for interpreting the disease/trait associated variants detected in GWAS studies.

DEEP LEARNING COUPLED WITH ABC FOR THE INFERENCE OF NATIVE AMERICAN EVOLUTIONARY HISTORY

Olga Dolgova, Iago Maceda, Oscar Lao

Population Genomics Group, Centre Nacional d'Anàlisi Genòmica (CRG-CNAG), Barcelona, Spain

Quantifying the amount of admixture and the relationship of (sometimes) unknown ancestral populations is a complex task in human population genomics. In the present ongoing project we have been developing a novel approach based on coupling of Deep Learning with Approximate Bayesian Computation for alleviating these problems with the aim to detect the demographic processes along human history and to infer the more reliable *Hominin* phylogeny, including multiple introgression and admixture events among ancient and even unknown archaic populations.

The consensus view on the peopling of the Americas is that ancestors of modern Native Americans entered the Americas from Siberia via the Bering Land Bridge and this occurred at least ~14.6 thousand years ago (ka). However the number and timing of migrations into the Americas remain controversial, with conflicting interpretations based on anatomical and genetic evidence (Raghavan et al. 2015; Soglund et al. 2015).

Here we present the preliminary results on population history of present-day Native Americans being a hot debated topic in last decades. Positive correlations between replication and training datasets as an evidence of robustness of Neural Network results were strong and significant in most of the model/population combinations. Estimation of posterior probabilities from six models elucidated the footprints of archaic introgression at least in four Native American populations (Pima, Karitiana, Maya and Mixtec tribes) of Neanderthal-Denisovan ancestor nature.

**References**

Raghavan M. et al. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349(6250), aab3884 (2015)

Soglund P. et al. Genetic evidence for two founding populations of the Americas. Nature 525, 104-108 (2015).

# DETECTION OF PATHOGENIC REPEAT EXPANSIONS FROM HIGH-THROUGHPUT WHOLE-GENOME SEQUENCE DATA

Egor Dolzhenko[1], Kristina Ibanez[2], Arianna Tucci[2], Joke J van Vugt[3], Giuseppe Narzisi[4], Katherine R Smith[2], Richard Scott[2], Augusto Rendon[2], Jan H Veldink[3], Mark J Caufield[2], David R Bentley[5], Michael A Eberle[1]

[1]Illumina Inc, Clinical Genomics Research, San Diego, CA, [2]Genomics England, Queen Mary University London, London, United Kingdom, [3]University Medical Center Utrecht, Neurology, Utrecht, Netherlands, [4]New York Genome Center, New York, NY, [5]Illumina Cambridge Ltd, Clinical Genomics Research, Little Chesterford, United Kingdom

Repeat expansions cause a variety of disorders including Fragile X syndrome, ALS, and Huntington's disease. Targeted assays that are typically used to test for the presence of these expansions are laborious and place constraints on the number of repeat regions that can be tested. In addition, test results can be compromised for technical reasons, such as degradation of large expansions in tests that use PCR amplification. With whole-genome sequencing (WGS) becoming the basis of many precision medicine initiatives, there is an opportunity to replace many targeted tests with a single WGS-based test. Repeat expansions previously have been considered undetectable using short-read sequence data. We have now shown that we are able to detect pathogenic repeat expansions accurately, by combining PCR-free sequencing technology with novel analytical methods. Our software tool, ExpansionHunter, has been incorporated into the ongoing discovery efforts by 100,000 Genomes Project Rare Disease Programme. Currently, this work has identified ten pathogenic repeat expansions consistent with the patient phenotype. Six of these expansions have been validated and the remaining are awaiting validation.

We anticipate that there are further repeat expansion disorders to be discovered. Thus, in addition to identifying known repeat expansions, it is also important to be able to identify new pathogenic repeat expansions without prior knowledge of their location or the repeat motif. To this end, we have developed a new method that enables us to perform a genome-wide search for novel repeat expansions. To demonstrate the power of this method we have analyzed 4 groups of samples each containing between 10 and 20 individuals harboring a known repeat expansion (associated with fragile X syndrome, Friedreich Ataxia, Myotonic Dystrophy type 1, and Huntington's disease) against 150 control samples. In this proof-of-concept analysis we were able to rediscover both the location and nucleotide composition of each underlying repeat region. Our findings add to the mounting evidence of the utility of whole-genome sequencing in precision medicine and rare disease diagnostics.

# TRANSCRIPTIONAL FATES OF HUMAN-SPECIFIC DUPLICATE GENES

Max L Dougherty[1], Jason G Underwood[1,2], Bradley J Nelson[1], Katherine M Munson[1], Alex A Pollen[3], Evan E Eichler[1,4]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]Pacific Biosciences, Menlo Park, CA, [3]UCSF, Neurology, San Francisco, CA, [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA

Structurally complex regions of the human genome are reservoirs for unexplored variation and include new genes generated within the human lineage (~80 genes from 33 gene families), some of which have been associated with human-specific aspects of neurodevelopment. However, our ability to understand their function and interpret variation has been confounded by misassembly of genomic sequence and the difficulty of distinguishing isoforms and paralogs with short reads. We developed a method that combines long-read full-length RNA-sequencing with target enrichment, enabling isoform-level resolution from unambiguously mapped reads, to identify gene innovations in these regions.

We targeted 53 genes/families with biotinylated probes in human brain cDNA and performed single-molecule long-read (PacBio) RNA sequencing, generating a total of 1.4 million sequences. We reconstructed long-read-based gene models, characterized newly identified features, and examined the expression of the new transcripts using short-read data.

We find that while the majority of human-specific gene duplications are incomplete gene copies (63%), nearly all are transcribed, including 89% of partially duplicated genes, and that fusion transcription across duplication boundaries is common, connecting these partial gene duplicates to new upstream and downstream exons. We find that the resemblance between expression patterns of paralogs depends on the extent of the duplication, and identify splicing differences between paralogs that may indicate changes in selective pressures. Finally, we identify new, previously unannotated features including conserved coding sequence.

Correcting and completing the annotation of highly identical duplicate genes will improve our understanding of the basis of human-specific traits and the role of these new genes in human evolution and disease. More broadly, the methods we developed can be applied to other genomes to characterize duplicated genes important for adaptation.

ACCESSING ENCODE PROJECT DATA USING A REST API AND JSON OBJECTS.

Idan Gabdank, Esther T Chan, Jason A Hilton, Jean M Davidson, Seth Strattan, Aditi K Narayanan, Kathrina C Onate, Marcus C Ho, Timothy R Dreszer, Ulubek K Bayadov, Laurence D Rowe, Stuart R Miyasato, Forrest Y Tanaka, Matt Simison, Benjamin C Hitz, Cricket A Sloan, Michael Cherry

Stanford University School of Medicine, Department of Genetics, Palo Alto, CA

The Encyclopedia of DNA Elements project (ENCODE) has been producing data for over a decade to investigate DNA and RNA binding proteins, chromatin structure, transcriptional activity and DNA methylation on a variety of human and model organism tissues and cell lines. As the complexity and diversity of the data grows, the tools required to organize, search and access the data in meaningful ways need to be more sophisticated. The ENCODE Data Coordination Center (DCC) has incorporated a representational state transfer application programming interface (REST API) with JSON (Javascript Object Notation) objects to facilitate the access of ENCODE experimental metadata using a web portal. Metadata can be accessed and data can be searched for at http://www.encodedcc.org/ using the HTTP request from a script or the curl command. We further expand on the access capability by allowing filtering of the metadata with the use of search urls. This system allows external researchers to write their own interfaces to access, analyze and visualize the ENCODE data. It also facilitates the integration of other large-scale datasets such a REMC, modENCODE, modERN and GGR with the ENCODE data. Here we will present our JSON schemas, examples of the REST API and use-cases for the search functions. Our goal is for the scientific community to use the ENCODE data through these methods for data mining and integration.

# CHARACTERIZING LINEAGE SPECIFIC CIS-REGULATORY EVOLUTION

Noah Dukler[1,2], Yi-Fei Huang[1], Adam Siepel[1]

[1]Weill Cornell Medical College, PBSB, NYC, NY, [2]Cold Spring Harbor Laboratory, SCQB, Cold Spring Harbor, NY

With the proliferation of high-throughput functional genomics assays (e.g. ChIP-seq and ATAC-seq), there has been an explosion in the amount of epigenetic data available within and between species. Comparative analysis of epigenomic data provides the opportunity to improve our understanding of the evolution of regulatory elements. A wide variety of phylogenetic tools (e.g phastCons, GERP++, and DLESS) have been developed to interrogate the forces that shape the evolution of genomic sequences, and to identify loci associated with disease, development, and molecular phenotypes. However, no similar methods exist for analyzing comparative epigenomic data in a rigorous phylogenetic setting.

While the evolution of primary sequence has been well characterized, how higher order regulatory elements evolve with and without selective pressure remains an open question. A better understanding of how selective pressure shapes the turnover of enhancers and promoters would provide insight into the mechanisms by which cis-regulatory circuits are evolved. Here we propose a new phylogenetic hidden Markov model, epiPhylo, for reconstructing the evolutionary histories of enhancers and promoters from comparative ChIP-seq data. EpiPhylo combines a phylo-HMM with a negative binomial error model to jointly infer the location of regulatory elements and their evolutionary trajectories from noisy data. We apply epi-Phylo to existing mammalian datasets (Villar et al. 2015 & Berthelot et al. 2017) for histone marks canonically associated with enhancers and promoters, and show that sequence and epigenetic conservation are positively correlated. We also investigate whether specific regulatory pathways undergo accelerated turnover of regulatory elements in a lineage specific fashion and estimate the prevalence of compensatory turnover of regulatory sequences. More broadly, we show the power of epiPhylo to analyze comparative data for any binary trait (e.g. TF binding, DNase hypersensitivity) at multiple genomic scales. In summary, epiPhylo is a rigorous statistical phylogenetic model of epigenomic evolution, enabling the inference of explicit evolutionary histories and selective pressure for a variety of regulatory elements.

# THE LANDSCAPE AND EVOLUTION OF SOMATIC MUTATIONS IN BOVINE LEUKEMIA VIRUS INDUCED TUMORS

Keith Durkin[1], Maria Artesi[1], Vincent Hahaut[1], Philip Griebel[3], Natasa Arsic[3], Arsène Burny[2], Michel Georges[1], Anne Van den Broeke[1,2]

[1]Unit of Animal Genomics, GIGA-R, University of Liège, Liège, Belgium, [2]Laboratory of Experimental Hematology, Institut Jules Bordet, Université Libre de Bruxelles (ULB), Brussels, Belgium, [3]VIDO, University of Saskatchewan, Saskatoon, Canada

Bovine Leukemia Virus (BLV) is a deltaretrovirus that integrates into B-cells producing a lifelong infection in cattle. Like its close relative Human T-cell leukemia virus-1 (HTLV-1), BLV induces an aggressive leukemia/lymphoma in about ~5% of infected individuals. While not a natural host it is possible to infect sheep with BLV and in contrast to cattle, all infected sheep develop tumors at an accelerated rate (~24 months). Historically research into both viruses has primarily focused on their transcripts/proteins. However secondary somatic events are likely to be important as only a subset of infected individuals, following many decades of infection, develop a tumor. At the current time little is known about the landscape of somatic changes in BLV induced tumors and the timing of their occurrence. To address this we have carried out whole genome sequencing of BLV induced tumors from two cattle, and from five sheep with matched normal tissue. This revealed frequent aneuploidy, with orthologous regions of the genome involved in both species and elevated mitochondria DNA copy numbers in tumors. Recurrent structural variants (SVs) were seen affecting the tumor suppressors *CDKN2A* and *ARID1A*, both on OAR2. On average ~1400 somatic SNVs were observed in each ovine tumor, with high/moderate impact variants in known cancer drivers genes such as *KMT2A*, *ATRX* and *KRAS*. The five sheep were also sampled at regular time points, prior to leukemia onset, allowing us to examine tumor clone evolution. High throughput sequencing of proviral integration sites showed that the tumor clone represents only a small fraction of the infected cells for the majority of the disease, only expanding rapidly in the terminal stages. Low coverage sequencing of samples prior to tumor development indicates that aneuploidy of OAR9 is a feature of the majority of BLV infected clones. Preliminary nested PCR also showed that many SVs were present prior to the rapid expansion of the tumor clone. The results of ongoing work to track the emergence of both somatic SVs and SNV in the preleukemic stages of the disease will be presented.

**GENEPANEL.IOBIO -** AN INTERACTIVE WEB APPLICATION TO GENERATE LISTS OF PRIORITIZED GENES

Aditya Ekawade[1,2], Tonya Di Sera[1,2], Chase Miller[1,2,3], Alistair Ward[1,2,3], Matt Velinder[1,2], Gabor Marth[1,2,3]

[1]University Of Utah, Human Genetics, Salt Lake City, UT, [2]University Of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]Frameshift Genomics, Boston, MA

The field of medical genomics is rapidly expanding as a result of decreasing sequencing costs, and maturation of comprehensive genomic analysis pipelines. Growing interest in personalized medicine is helping to drive the adoption of clinical sequencing for patients with Mendelian disorders, cancer, prenatal screening, or pharmacogenomic purposes. Analysis of such data is a complex process demanding a wide range of technical expertise. The first step in this analysis process is to analyze those variants in genes that have been previously implicated in similar disorders, or associated with the phenotypes present in the patient who has undergone sequencing. A number of tools and databases exist, but compiling a comprehensive list of genes from these sources is far from straightforward, and no tools currently exist to offer this functionality for biomedical researchers and diagnostic clinicians.

Here, we present a new web-based application, ***genepanel.iobio*** **(http://genepanel.iobio.io)**, that focuses on the construction of a prioritized gene list to support bioinformaticians, clinical and diagnostic analysts, and physicians perform rapid analysis of patient data themselves. This app compiles comprehensive data from a number of different resources including commercially available panel tests and phenotype term-driven gene list generation tools. The NCBI's Genetic Testing Registry (GTR) compiles information on available gene panel tests provided by any commercial and academic vendors or diagnostic labs. Within genepanel.iobio, the user can search these gene panels based on a disorder of interest and the output will be a union of genes present in all available tests. The user can then filter these results to only include (or exclude) data from specific vendors, or identify genes that are shared across many such tests. The Phenolyzer tool from USC takes one or more phenotype terms and within seconds generates a prioritized list of genes based on association with the stated phenotype. These gene lists from the Phenolyzer can be filtered by the user and combined with genes from GTR to create comprehensive lists of genes for supporting genomic analysis. These gene lists can be easily integrated in the future with other IOBIO applications for a comprehensive analysis solution, or readily used by researchers and clinicians for diagnostic and gene discovery applications.

# SUPPORTING USERS OF THE 1000 GENOMES PROJECT DATA AND IMPROVING DATA RESOURCES IN THE INTERNATIONAL GENOME SAMPLE RESOURCE (IGSR)

Susan Fairley, Astrid Gall, Erin Haskell, Ernesto Lowy, Benjamin Moore, Emily Perry, Ian Streeter, Laura Clarke, Paul Flicek

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

The 1000 Genomes Project created the largest fully public catalogue of human genetic variation and a set of valuable reference resources that remain widely used.
The International Genomes Sample Resource (IGSR) exists to: 1) share new data generated on the 1000 Genomes samples, 2) make additional populations available alongside the existing 1000 Genomes populations and 3) ensure the continued usability of the original 1000 Genomes data, including updating it to GRCh38.
Updating 1000 Genomes data to GRCh38 has included aligning the sequence data to GRCh38 in an alt-aware manner and generating call sets directly on GRCh38 using BCFtools, FreeBayes and GATK's Unified Genotyper. Development of an integrated call set is in progress, with preliminary data from this process available on the project FTP site. An assembly mapping based 'liftover' of the phase three calls, generated by dbSNP and the European Variation Archive (EVA), is also available. This data can be browsed in Ensembl, which has extended its range of tools for working with 1000 Genomes data on GRCh38.
IGSR newly incorporates a wide range of data types generated by the Human Genomes Structural Variation Consortium (HGSVC) on 1000 Genomes samples. These include PacBio, 10X, Bionano and Oxford Nanopore data.
IGSR provides direct and individual support to users of the 1000 Genomes data and other newly added data sets via the helpdesk at info@1000genomes.org. In 2017 alone, hundreds of user enquiries covered a range of topics relating to both the data and options for accessing it. IGSR also continues to expand the data sets presented in the data portal on the IGSR website, facilitate searching for files among the >500,000 available on the project FTP site, and help users locate and use the available resources.
Finally, to facilitate the availability of newly created open genomic data, IGSR will provide data coordination work including assistance in ethical review and alignment of sequence data. IGSR is specifically interested in establishing new collaborations with projects based on openly consented genomic data with a strong desire to expand the range of populations represented.

# ANCIENT AFRICAN SUBSTRUCTURE INFERRED FROM WHOLE GENOME SEQUENCE DATA

Shaohua Fan[1], Derek E Kelly[1], Marcia H Beltrame[1], Matt Hansen[1], Swapan Mallick[3,4], Thomas Nyambo[5], Dawit Wolde Meskel[6], Gurja Belay[6], Nick Patterson[3], David Reich[2,3,4], Sarah A Tishkoff[7]

[1]University of Pennsylvania, Department of Genetics, Philadelphia, PA, [2]Harvard Medical School, Boston, Department of Genetics, Boston, MA, [3]Broad Institute of Harvard and MIT, Cambridge, MA, [4]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, [5]Muhimbili University of Health and Allied Sciences, Department of Biochemistry, Dares Salaam, Tanzania, [6]Addis Ababa University, Department of Biology, Addis Ababa, Ethiopia, [7]University of Pennsylvania, Department of Biology, Philadelphia, PA

Africa is the origin of modern humans within the past 300 thousand years ago (kya). To infer the complex demographic history of African populations and adaptation to diverse environments, we sequenced the genomes of 94 individuals from 44 indigenous African populations and compared to a set of 62 individuals from 32 west Eurasian populations from the Simons Genome Diversity Project. Phylogenetic analysis confirms that the San lineage is basal to all modern human population lineages. The phylogenetic locations of other African populations largely correlate with their current geographical locations, with the exception of the central African rainforest hunter-gatherer (CRHG) and some pastoralist populations. An early emergence of population structure was observed at ~200 thousand years ago (kya), corresponding to the time of origin of modern humans. The San and CRHG populations have maintained the largest effective population size compared to other populations prior to 60 kya. Using MSMC analysis, the San lineage and all non-Khoesan speaking populations diverged from their common ancestor at ~100-120 kya. In contrast, the divergence of African hunter-gatherer populations, including the San lineage, the CRHG and the Hadza and Sandawe, was within the past 66-82 kya, suggesting these populations could have originated from a historically more widespread population of hunter-gatherers. The divergence times of Niger-Kordofanian, Nilo-Saharan and Afroasiatic speaking populations were within the past ~22 to 41 kya. In the CRHG populations, the oldest divergence was observed between Eastern and Western RHG at ~36-51 kya; the time of divergence of the western CRHG was inferred to be ~12-18 kya. We observed signatures of positive selection at genes involved in muscle development, bone synthesis, reproduction, immune function, energy metabolism, and cell signaling, which may contribute to local adaptation of African populations.

# RUFUS: REFERENCE FREE VARIANT DETECTION IMPROVES ACCURACY AND SENSITIVITY

Andrew Farrell[1,2]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]USTAR Center for Genetic Discovery, Salt Lake City, UT

We developed a novel k-mer based variant detection tool, RUFUS, that vastly improves specificity and sensitivity for germline and somatic/tumor de novo mutations, and may reveal some of the missing heritability in many genetic diseases. RUFUS is based on direct k-mer comparison, removing the reference from variant detection, and any associated reference bias. This vastly improves the detection of medium sized (20-500bp) insertions/deletions (INDELs) that current methods are unable to reliably detect: small variant detectors (e.g. GATK, FreeBayes) are effective at finding 1-20bp events in read alignments; and structural variant callers (LUMPY, WHAM, etc.) are effective at >500bp events where insert size variations and read coverage anomalies can be confidently detected. As a result, medium length, INDELs have been missed by most sequencing studies. We are currently applying RUFUS to 519 family quartets (mother, father, autistic child, unaffected sibling), a total of 2,076 samples sequenced by the Simon's Foundation Autism Research Initiative to 30x whole genome coverage. This will be the largest de novo variation study to date, and combined with RUFUS's sensitivity for all variant types and sizes will provide the most complete picture of de novo variation ever constructed. Our preliminary data on a 40 family pilot study has shown that the rate of medium length de novo events is twice that of structural events (12 medium-length vs 6 SV events), suggesting that these events may be more common than previously thought.

In addition to increased sensitivity for variants of all sizes, RUFUS also shows far higher specificity over mapping based approaches. Previous research has suggested that the human per-nucleotide de novo SNV mutation rate is ~1.25 x 10-8, or roughly 75 mutations per generation. In our analysis of numerous disease family trio data sets at the University of Utah, RUFUS finds between 77 and 116 de novo mutations per child genome, including 74-101 SNV mutations; 98% also seen in mapping-based variant calls. Conversely, traditional mapping based methods call on average 150,000 de novo calls per child, dominated by mapping and reference errors, drowning out true variation, and post-processing and genome masking is necessary to improve these, still leaving thousands of de novo calls. RUFUS requires no filtering or masking of the genome, enabling true genome wide variant detection of all mutation types, at uniquely high specificity.

# RECONSTRUCTING THE SEQUENCE OF EVENTS: INTROGRESSION AND RAPID DIVERGENCE IN RECENTLY EMERGED TREE PATHOGENS

Anna Fijarczyk[1,5], Pauline Hessenauer[1], Hélène Martin[1,5], Louis Bernier[2], Philippe Tanguay[3], Richard Hamelin[4,1], Christian R Landry[1,5]

[1]Institut de Biologie Intégrative et des Systèmes, Université Laval, Département de Biologie, Québec, Canada, [2]Université Laval, Département des Sciences du Bois et de la Forêt, Québec, Canada, [3]Natural Resources Canada, Laurentian Forestry Centre, Québec, Canada, [4]The University of British Columbia, Department of Forest and Conservation Sciences, Vancouver, Canada, [5]PROTEO, The Quebec Network for Research on Protein Function, Engineering, and Applications, Québec, Canada

Pathogens constantly evolve in a co-evolutionary arms race with their hosts. Long-term selective pressures exerted by the host lead to evolution of increased virulence and/or transmission via molecular changes in protein-coding genes and gene duplications, accelerating the diversification and expansion of protein families. However, the pace at which these changes take place remains to be examined in detail. Mechanisms behind rapid adaptations include recombination between divergent strains, hybridization between closely related species, or horizontal gene transfer with unrelated species. Here we investigate the role of hybridization and genomic introgression in the emergence of three aggressive pathogenic fungi responsible for massive declines of elm trees around North America and Europe in the last century. We sequenced the genomes of 94 world-wide samples of three *Ophiostoma* subspecies causing two pandemics of Dutch elm disease, the latter being more dangerous. Our study revealed that the two most aggressive pathogen subspecies from the recent disease outbreak are in fact three widespread genetic lineages. Genome-wide analysis suggests that three *Ophiostoma* lineages diverged almost simultaneously, and quickly spread across continents to reach their current distributions. In spite of partial reproductive barriers, the species and subspecies hybridize with each other, leaving signals of recent introgression in their genome. Presence of introgression is likely a consequence of rapid spread, which brought distinct lineages together. On the other hand, patterns of divergence and differentiation along the genome also suggest that introgression might have happened before the divergence of lineages linked to second pandemic of Dutch elm disease. We put forward the hypothesis that ancient introgression triggered the emergence of aggressive *Ophiostoma* lineages and are currently testing this scenario. The role of hybridization in diversification of lineages and species is gaining attention, and can be potentially an important mechanism in the evolution of pathogenic fungi.

# GENETIC AND ENVIRONMENTAL EFFECTS ON GENE REGULATION IN THE VASCULAR ENDOTHELIUM

Anthony S Findley[1], Allison L Richards[1], Cristiano Petrini[1], Alexander S Shanku[1], Adnan Alazizi[1], Elizabeth Doman[1], Omar Davis[1], Yoram Sorokin[2], Nancy Hauff[2], Xiaoquan Wen[3], Roger Pique-Regi[1,2], Francesca Luca[1,2]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]Wayne State University, Department of Obstetrics and Gynecology, Detroit, MI, [3]University of Michigan, Department of Biostatistics, Ann Arbor, MI

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits. However, only a limited number of environmental factors are measured in GWAS. Controlling for, or accurately measuring, all possible environmental factors in a GWAS setting is a formidable challenge. Instead, molecular phenotypes (gene expression, chromatin accessibility) measured in tightly controlled cellular environments provide a more tractable setting in which to study gene-environment interactions (GxE) in the absence of other confounding variables.

Here we have exposed human umbilical vein endothelial cells (HUVECs) from 17 healthy donors to 3 treatments (dexamethasone, retinoic acid, and caffeine) and appropriate vehicle-controls for 6 hours. We genotyped and performed RNA-seq and ATAC-seq to model genetic and environmental effects on gene regulation and chromatin accessibility in the vascular endothelium, a common site of pathology in cardiovascular disease (e.g., atherosclerosis).

All three treatments induced significant alterations in the transcriptional and chromatin landscapes. Comparing each treatment to its appropriate control, we have identified 2879, 4874, and 5790 differentially expressed genes (FDR < 10%) and 419, 178, and 711 regions with differentially accessible chromatin (FDR < 10%) in response to dexamethasone, retinoic acid, and caffeine, respectively. We found that genes near regions of differentially accessible chromatin were more likely to be differentially expressed (OR = [3.41, 6.52], p < 10-16). Additionally, we have identified transcription factor footprints in each condition and observed an enrichment of the glucocorticoid receptor in response to dexamethasone and the retinoic acid receptor in response to retinoic acid. Using RASQUAL for joint ASE - QTL mapping, we found eQTLs for 574, 490, and 424 genes for dexamethasone, retinoic acid, and caffeine, respectively (FDR < 10%). eQTLs were enriched for SNPs predicted to alter transcription factor binding.

Our results indicate that changes in chromatin accessibility are a mechanism for gene expression response to treatment. We identified putative mechanisms for GxE that could be missed in studies of mixed environmental contexts. For example, we found that a SNP that was not identified as an eQTL in GTEx is associated with coronary artery disease risk in Cardiogram, and is also an eQTL for the LDL receptor in our caffeine-treated cells.

# THE INBRED MEDAKA KIYOSU PANEL

Tomas W Fitzgerald[1], Jakob Gierten[2], Felix Loosli[3], Jochen Wittbrodt[2], Ewan Birney[1]

[1]The European Bioinformatics Institute (EMBL-EBI), Birney research, Cambridge, United Kingdom, [2]Centre for Organismal Studies (COS), Wittbrodt, Heidelberg, Germany, [3]Karlsruhe Institute of Technology (KIT), Loosli, Karlsruhe, Germany

Over the last six years dedicated work has led to the establishment of the inbred medaka Kiyosu panel. Randomly selected mating pairs originating from an outbred wild population in Kiyosu Japan were used to start the inbreeding scheme and after 9 rounds of single brother-sister mating, 111 inbred lines remained from 83 different original wild founding breeding pairs. Thus 28 lines are “sib lines” to another line, a feature we aim to exploit in our statistical analysis. We have completed whole genome sequencing across the entire panel and called homozygous and heterozygous SNPs in each line against the improved PacBio derived medaka reference. Amazingly over 75% of the lines are greater than 80% homozygous providing a truly unique model organism resource. This level of homozygousity in over one hundred lines bred from a single wild population is unmatched in any other vertebrate model system and is similar to equivalent Drosophila resources. The isogenic nature of the Kiyosu panel promotes investigations into gene to environment interaction effects testing molecular measurements under tightly controlled environmental conditions.

For genetic association testing across the panel we have developed pilot phenotyping assays and analytical techniques for high throughput heart beat screening, behavioural assays, morphometric measurements and CT scanning of individual fish. Although whole organism phenotypes, such as heart rate or movement parameters are unlikely to provide specific loci from only panel phenotyping we aim to boost statistical power using carefully designed F2 cross strategies. We have developed an extensive simulation and statistical power assessment of the genetic, environmental and stochastic effects we might encounter, this simulation scheme handles polygenic phenotypes, multi-trait phenotypes and can simulate the results of phenotyping the Kiyosu panel and expected phenotyping of crosses. Multiple F2 crosses from the phenotypic extremes of the panel (for example 9 crosses between the 3 highest and 3 lowest individuals), focusing on lines with the strongest environmental or stochastic components is a particularly powerful approach, both for finding broad locus location and estimating effect sizes of loci discovered in the wild. Here we will present initial sequence analysis across the panel along with power analysis and pilot data for different F2 cross strategies for mapping discovered loci to the medaka genome.

# THE HUMAN CELL ATLAS DATA COORDINATION PLATFORM

Mallory A Freeberg[1], Human Cell Atlas Data Coordination Platform Team[1,2,3,4]

[1]EMBL-EBI, Human Cell Atlas, Hinxton, United Kingdom, [2]UCSC, Human Cell Atlas, Santa Cruz, CA, [3]Broad Institute, Human Cell Atlas, Cambridge, MA, [4]Chan Zuckerberg Initiative, Human Cell Atlas, Palo Alto, CA

The Human Cell Atlas (HCA) is taking a systematic, data-driven approach to create a reference map of all human cells. This Atlas will consider cell type alongside other facets of a cell's identity such as state, transitions between cell types, lineage, cell-cell interaction, and a cell's local neighbourhood. The HCA will be used as a basis for understanding human health and diagnosing, monitoring, and treating disease.

This massive undertaking requires an open, modular, and extensible approach to data coordination. The Broad Institute, Chan Zuckerberg Initiative, EMBL-EBI, and UCSC are building a Data Coordination Platform (DCP) to organise terabytes of data for billions of cells, across multiple modalities, generated by hundreds of labs around the world. The DCP will enable the community to innovate rapidly, without barriers to access, and facilitate computational researchers developing new analysis methods. The DCP has four components: the ingestion service, the data store service, the secondary analysis service, and the release service. All the software will be openly developed and licensed and released in forms which enable easy reuse in cloud infrastructures.

The *ingestion service* provides human support and software tools, as APIs and UIs, to enable data generators to
provide well-structured data and descriptions to the HCA.

The *data store service* provides multi-cloud-based storage for all raw data, metadata, and certain derived data. Users will be able to access the data directly via the consumer API or via visualization and analysis tools produced as part of the release service and third party tertiary portals.

The *secondary analysis service* provides a pipeline execution infrastructure for robust, community-vetted pipelines to be run in a standardized way. The results will be deposited back into the data store. The HCA analysis working group will identify which pipelines to run, ensuring there is at least one pipeline for each anticipated data type. All pipelines will be built using open source code and shared via containers to ensure the entire community can take advantage of this work.

The *release service* supports users discovering the data the Human Cell Atlas has collected and downloading results in matrix files suitable for downstream analysis. This service together with the consumer API will facilitate the scientific community using the HCA data to answer scientific questions.

Here we present an overview of the DCP and its components. For more information, please read our website: https://www.humancellatlas.org/

# IDENTIFICATION OF RARE-DISEASE GENES FROM RNA-SEQ OF UNDIAGNOSED CASES USING LARGE CONTROL COHORTS

Laure Fresard[1], Craig Smail[1,2], Kevin S Smith[1], Brunilda Balliu[1], Nicole M Ferraro[1,2], Nicole A Teran[1,3], Kristin Kernohan[4], Shruti Marwaha[5], Devon Bonner[5], Jean M Davidson[5], Jennefer Kohler[5], Dianna G Fisk[6], Megan Grove[6], Euan A Ashley[3,7], Kym Boycott[4], Jason D Merker[1,6], Matthew T Wheeler[5,6], Stephen B Montgomery[1,3]

[1]Stanford University, Pathology, Stanford, CA, [2]Stanford University, Biomedical Informatics Program, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA, [4]Eastern Ontario Research Institute, Children's Hospital, Ottawa, Canada, [5]Stanford University, Center for Undiagnosed Diseases, Stanford, CA, [6]Stanford University, Clinical Genomics Program, Stanford, CA, [7]Stanford University, Division of Cardiovascular Medicine, Stanford, CA

Large exome sequencing projects have uncovered plentiful rare protein-coding variants. Interpreting the consequences of these rare alleles has immediate bearing on understanding the role of genetics in individual health in the era of precision medicine. Outside of protein-coding genes, interpretation of rare alleles remains a considerable challenge. We have recently demonstrated that joint analysis of genomes and transcriptomes can identify genes containing rare regulatory and splicing variants in healthy individuals. Here, we extend this approach to individuals that are affected by rare genetic diseases. Specifically, we sequenced whole transcriptomes from peripheral blood of 61 cases from the UDN, Care4Rare and the Stanford Medicine Clinical Genomics Service. We developed a robust approach to compare rare disease cases to large existing sets of transcriptome-sequencing controls to identify disease genes as well as rare and causal non-coding variants. This effort has already culminated in the discovery and validation of rare disease variants in progressive myoclonic epilepsy and Carney Complex cases. Previous work has demonstrated the power of RNA-seq to help diagnose rare diseases in a tissue relevant to the disease category. Here we demonstrate that whole blood is a good surrogate tissue for rare disease diagnosis in diverse disease categories. We show an enrichment for expression outliers in loss of function intolerant genes in cases in comparison to healthy controls. We demonstrate that combining different levels of gene expression patterns (total expression, splicing, allele specific expression) together with genetic information can significantly narrow down candidate gene lists for follow-up studies. More generally, these approaches demonstrate how personal functional genomics when coupled with large-scale reference data can aid in the interpretation of impactful personal variants

# PROTEOGENOMIC CHARACTERIZATION OF HUMAN TISSUES REVEALS mRNA MOTIFS CONTROLLING PROTEIN ABUNDANCE

Basak Eraslan[1], Dongxue Wang[2], Hannes Hahne[2], Mirjana Gusic[3], Holger Prokisch[3], Bernhard Kuster[2], Julien Gagneur[1]

[1]Technical University of Munich, Faculty of Informatics, Garching, Germany, [2]Technical University of Munich, Chair of Proteomics and Bioanalytics, Freising, Germany, [3]Helmholtz Center Munich, Institute of Human Genetics, Neuherberg, Germany

Due to post-transcriptional regulation, variation in protein-to-mRNA (PTR) ratios across human genes span more than 2 orders of magnitude. Despite their importance in determining protein levels, a comprehensive catalogue of cis-regulatory elements controlling PTR and a quantification of their effects is still lacking. Here we analyzed the contribution of various mRNA and protein sequence elements to overall PTR ratio by the use of unpublished matching transcriptomics and proteomics data of 10,082 proteins across 29 human tissues. Our sequence-based model of PTR ratio includes well-known determinants, such as N-degrons, Kozak sequence, stop codon context, codon usage, micro-RNA motifs, and RNA-binding protein recognition sites. Additionally, we performed de-novo search of sequence motifs predictive of PTR ratios. This recovered well-known motifs including the Pumillo motif and the AU-rich element and revealed 5 novel motifs in 5'UTR and 5 novel motifs in 3'UTR. Moreover, our model leads to a definition of codon optimality based on a direct comparison of protein to RNA level that differs significantly from previous codon optimality measures such as genomic frequency and codon absolute adaptiveness. The vast majority of the sequence feature effects of our model reproduce in an independent dataset of matched transcriptomes and proteomes. Altogether, this study shows that a large fraction (~35% variance in PTR) can be predicted from mRNA and protein sequence and identify many new candidate human post-transcriptional motifs.

# SANDY: A STRAIGHTFORWARD AND STREAMLINED NEXT-GENERATION SEQUENCING READ SIMULATOR

Thiago A Miller, Fernanda Orpinelli, Pedro A Galante

Hospital Sirio Libanes, Bioinformatics, Sao Paulo, Brazil

Many next-generation sequencing (NGS) analyses rely on hypothetical models and principles that are not precisely satisfied in practice. Many times, a widely used pipeline in the literature (e.g., for SNVs calling or gene expression) produces unexpected results when applied in researcher's dataset. However, discovering if it is a consequence of technical mistakes during the analysis, a consequence of biological phenomenon, something upstream the bioinformatics analysis or even necessity of tuning parameter for the local data is extremely difficult. Simulated data, which provides positive controls would be a perfect way to overcome these difficulties. Nevertheless, most of NGS simulators are extremely complex to use, they do not cover all kinds of the desired features and/or are very slow to run. Here, we present "SANDY", a straightforward, easy to use, fast, and integrated set of tools to generate synthetic next-generation sequencing reads. SANDY mimic whole genome sequencing (including insertion of single, short and long genomic mutations), whole exome sequencing and RNA (short and long) sequencing. Sandy can be used therefore for benchmarking results of a variety of pipelines, including those for read alignment, calling of single/short/long genomic variation, gene expression, and de novo genomic or transcriptomic assembly, for example. Users can easily define information for their simulations, such as i) sequencing coverage or the number of reads; ii) read length; iii) sequencing type, single or paired-end; iv) and others, such as sequencing error rate, fragment length, stranded or not-stranded (for RNAseq) sequencing. Interestedly, users as well have the possibility to include their own sequencing quality profile (obtained from his/her own sequencer), which ensures greater reliability for simulated data. SANDY is optimized to run locally or on cloud computing, it is also publicly available and easy to install since it is available as a docker image.

# THE MOLECULAR REQUIREMENTS FOR EPIGENETIC ESTABLISHMENT OF CENTROMERES DEPEND ON THE UNDERLYING DNA

Glennis A Logsdon[1], Craig W Gambogi[1], Evelyne J Barrey[2], Patrick Heun[2], Ben E Black[1]

[1]Department of Biochemistry and Biophysics, Graduate Program in Biochemistry and Molecular Biophysics, and Epigenetics Institute, Perleman School of Medicine, University of Pennsylvania, Philadelphia, PA, [2]Wellcome Trust Center for Cell Biology, Institute of Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

Recent breakthroughs with synthetic budding yeast chromosomes expedite the creation of synthetic mammalian chromosomes and genomes. One substantial barrier to translating these breakthroughs to mammals is the centromere, the locus required for chromosome segregation at cell division. Unlike in budding yeast, mammalian centromeres require the epigenetic mark carried by nucleosomes containing the histone H3 variant, CENP-A. Prior human artificial chromosomes (HACs) required repetitive DNA encoding a binding site for the DNA sequence-specific binding protein, CENP-B. We developed two types of HACs that are completely independent of CENP-B: one with a repetitive centromere DNA template that is strongly stimulated by seeding CENP-A nucleosome assembly, and a non-repetitive template that does not require seeding and permits annotation of HAC copy number and organization. Together, the new HACs reveal an unexpected influence of DNA sequence in centromere formation.

# CONTRIBUTION OF RETROTRANSPOSITION TO DEVELOPMENTAL DISORDERS

Eugene J Gardner[1], Alejandro Sifrim[2], Giuseppe Gallone[1], Elena Prigmore[1], Helen V Firth[1,3], Matthew E Hurles[1], on Behalf of the Deciphering Developmental Disorders Study [1]

[1]Wellcome Sanger Institute, Human Genetics, Cambridge, United Kingdom, [2]KU Luven, Center of Human Genetics, Leuven, Belgium, [3]Cambridge University Hospitals NHS Foundation Trust, East Anglian Medical Genetics Service, Cambridge, United Kingdom

Mobile genetic Elements (MEs) are pieces of DNA which, through an RNA intermediate, can generate new copies of themselves within their host genome. Additionally, MEs can facilitate the duplication of non-ME transcripts, typically genes, through the mechanism of retroduplication. Combined, these two processes constitute what is known as retrotransposition (RT), and in humans several disorders can be attributed to such activity. However, the majority of these deleterious events have been discovered on a case-by-case basis and neither MEs nor gene duplications are routinely analysed as part of clinical sequencing. Likewise, large sequencing cohorts designed to elucidate the causes of congenital and developmental disorders (DDs) have neglected to identify pathogenic events attributable to RT-derived mutagenesis. As such, we have used computational approaches to identify RT events in 9,738 whole exome sequencing (WES) trios with DD-affected probands as part of the Deciphering Developmental Disorders (DDD) study. Through our analysis, we have discovered and genotyped 1,129 ME sites, of which ~20% directly impact coding sequence, and several hundred gene retroduplication events. We have also identified 9 de novo ME and 3 de novo gene retroduplications, 4 of which disrupt known DD genes and are likely causative of the patient's phenotype (0.04% of probands). We have also estimated the ME mutation rate in the human population to be ~1.5x10-11 mutations per bp per generation and have demonstrated our RT events have signatures of purifying selection equivalent to those of truncating mutations. Our study suggests that while the overall burden of RT-attributable disease is relatively low in the human population, it is nonetheless an important consideration when elucidating the genetic basis of DD in individual patients. Overall, our analysis represents the single largest interrogation of the impact of RT activity on the coding genome to date.

# IDENTIFICATION AND ANALYSIS OF SPLICING QUANTITATIVE TRAIT LOCI IN GTEx

Diego Garrido-Martín[1,2], Ferran Reverter[3], Miquel Calvo[3], Roderic Guigó[1,2]

[1]Centre for Genomic Regulation, Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat Pompeu Fabra, Experimental and Health Sciences, Barcelona, Spain, [3]Universitat de Barcelona, Statistics, Barcelona, Spain

The Genotype-Tissue Expression (GTEx) project has retrieved genetic and transcriptomic information across more than 50 tissues in hundreds of individuals, providing an unprecedented resource to address the identification of genetic variants associated with alternative splicing (splicing quantitative trait loci or sQTLs). To our understanding, splicing should be treated as a multivariate phenotype to be recapitulated completely. Hence, for sQTL mapping we developed sQTLseekeR, a method that addresses the identification of variants associated with changes in the relative abundances of a gene's transcript isoforms, relying on a nonparametric analogue to multivariate analysis of variance (MANOVA). Using GTEx v7 release (635 donors, 10,361 samples with both RNA-Seq and WGS data) we tested for association more than 17,000 protein coding genes and lincRNAs and almost 4 million *cis* variants, both SNPs and short *indels* (gene body plus 5Kb window upstream and downstream). On average per tissue we found 22,313 sQTLs (5% FDR) that affected 899 genes (sGenes). The ratio sGenes/expressed genes and the tissue sample size showed a substantial correlation. sQTLs were significantly more exonic and fell within splice sites, RNA-binding proteins' (RBP) binding sites and other functional elements more than non-sQTL SNPs. They were also enriched in GWAS hits, showing stronger associations than non-sQTLs with several traits. sQTLs displayed a substantial degree of sharing across tissues, and involved mostly complex changes in first/last exons and UTRs. Although most sQTLs have subtle effect sizes, around 14% present considerably large effects. sQTLs falling in splicing donor or acceptor sites modify splice site strength more than non-sQTLs, and incrementally with their effect size. GO enrichment analysis of sGenes revealed an enrichment in multiple metabolic and cellular processes, including RNA-processing, which might suggest some form of *trans* regulation. Furthermore, the proposed statistical framework is not restricted to splicing, and can be naturally generalized to study any other multivariate phenotype, being applicable to a wide variety of biological contexts.

# BIG BRAINS: WHAT HIGH-THROUGHPUT ENHANCER KNOCKOUTS REVEAL ABOUT HUMAN CORTICAL EVOLUTION

Evan Geller[1,2], James P Noonan[1,2,3]

[1]Yale University, Department of Genetics, New Haven, CT, [2]Yale University, Kavli Institute for Neuroscience, New Haven, CT, [3]Yale University, Ecology and Evolutionary Biology, New Haven, CT

The majority of genetic variation between humans and other primates resides within regions of gene regulation. It is hypothesized that genetic changes occurring within gene regulatory elements modified conserved developmental processes and contributed to human-specific biological phenotypes, such as the expansion and elaboration of the cerebral cortex. Multiple studies over the last decade have identified human-specific genetic changes, such as Human Accelerated Regions (HARs), that have been linked to changes in developmental gene regulation. However, the biological functions of these loci remain almost entirely unknown. To address this question, we used a high-throughput CRISPR-Cas9 knockout strategy in human neural stem cells to disrupt >50,000 potential transcription factor binding sites in >2,300 enhancers active in human cortical development. We targeted two classes of enhancers relevant to human evolution: human-gain enhancers showing increased epigenetic activity during corticogenesis and HARs that encoded enhancers active in the developing human cortex. Our assay quantified the effect of enhancer knockout on a critical phenotype during corticogenesis-- neural stem cell proliferation. A quantitative proliferation phenotype was obtained by measuring the abundance of each enhancer knockout at initial, intermediate, and final experimental time points by high-throughput sequencing. We found evidence of enhancer knockouts, including human-gain enhancers and HARs, with very strong effects on neural stem cell proliferation. Our genome-scale survey reveals the quantitative landscape of gene regulatory enhancer control of proliferation and utilizes advanced machine learning methods to interpret the biological impact of distinct enhancer classes on human cortical development and evolution.

# DIFFERENTIAL ALLELIC DROP-OUT WITH PARENT-OF-ORIGIN EFFECTS DUE TO G-QUADRUPLEXES AT IMPRINTED LOCI IN WHOLE GENOME SEQUENCE DATA FROM PCR LIBRARIES

Giulio Genovese[1,2], Chris Whelan[1,3], Robert E Handsaker[1,3], Seva Kashin[1,3], Steven A McCarroll[1,2]

[1]Broad Institute, Stanley Center for Psychiatric Research, Cambridge, MA, [2]Harvard Medical School, Department of Genetics, Boston, MA, [3]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA

Due to imprinting, the two copies of the human genome in each human cell are not interchangeable. Each copy, paternal and maternal, is required for proper development. However, most genotype experiments, from array genotyping to whole genome sequencing, are unable to retrieve these subtle systematic differences between paternal and maternal haplotypes, often referred to as differentially methylated regions (DMRs).

Using parental assignment of heterozygous alleles from whole genome sequence cohorts with trio design, we investigated the prevalence of differences directly or indirectly caused by DMRs that can be inferred from differences in coverage of the paternal and maternal alleles. Using whole genome sequence data of blood-derived DNA from libraries prepared using PCR at the Broad Institute, we identified several regions with excess paternal coverage. Most of these regions colocalize with DMRs known to be methylated in the maternal haplotype and they further colocalize with putative G-quadruplex structures, DNA secondary structures that are rich in guanines and believed to play a direct role in gene regulation. As CpG methylation is known to affect the thermal stability of G-quadruplexes, we hypothesize that the biases against the methylated allele in the PCR amplification are at the origin of the differential allelic drop-out (ADO).

We identify differential ADO at several known imprinted loci: GNAS, PLAGL1, NDN, RBMXL2, NPAS3, GRB10, ANO8, MEST, and others, including a few novel loci not previously known to be imprinted in humans. Most loci can be characterized as promoter regions of known imprinted genes. Consistent with our model implicating differential PCR amplification due to G-quadruplex formation, we identify that the effect is extremely punctuate and restricted to windows of a few hundred base pairs.

Our results provide an unbiased list of imprinted loci visible in blood-derived DNA and potentially shed new light on the effects of CpG methylation as a mechanism to control the stability of G-quadruplex structures in the human genome.

# A WEB TOOL FOR INTERPRETING GENOMIC PATIENT DATA IN THE CONTEXT OF LARGE DISEASE COHORT DATASETS

Stephanie Georges[1,2], Chase Miller[1,2,3], Alistair Ward[1,2,3], Tonya Di Sera[1,2], Gabor T Marth[1,2,3]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]Frameshift Genomics, Boston, MA

The medical genomics community is assembling vast, population-based and ailment-specific datasets in order to identify causative genetic variants and disease-associated genes. At present, scientists studying their own patient cohorts with such diseases must rely on indirect avenues to utilize the results from larger studies, e.g. use the ExAC/gnomAD database to look up variant population frequencies, and look up ClinVar for known pathogenic variants. Moreover, there are currently no tools that allow researchers and clinicians to answer subtle yet critical questions about patient variants in the context of the relevant, large research cohorts increasingly available to the community, e.g. whether a specific variant in the patient is enriched in a phenotypically similar subset of the large cohort.

To address this issue, we set out to create web-accessible, visually-driven software with a two-fold goal: to elucidate genetic variants enriched within a filtered dataset, and compare analyst-provided data to that of the enriched cohort. The cohort-gene.iobio application (http://cohort-gene.iobio.io) launches after the analyst selects a sample database, a gene or region of interest, and phenotypic filters for analysis. Within seconds of starting, cohort-gene.iobio compares sample variants to the selected dataset - user supplied or publicly available catalogs such as 1000G and ExAC - and generates annotations sourced from VEP, SIFT, PolyPhen, ClinVar, and others. Each group of samples is displayed within a single, cumulative variant track; variants are visualized to facilitate quantitative comparison between tracks, clearly and recognizably indicating enrichment in one cohort versus another. Additional annotation filters can be applied to variants while in the cohort-gene.iobio application, and multiple genes or regions can be further selected for simultaneous analysis. To give the analyst an indication of the quality of called variants, coverage tracks are displayed when possible. cohort-gene.iobio also provides convenient reporting functionality, saving visualization information in PDF format, and variant information in VCF format. In conjunction with our other iobio applications, cohort-gene.iobio offers powerfully complex variant processing in a streamlined, intuitive interface, and stands to help close the gap between intricate genomics data and directed patient care.

# INTEGRATIVE ANALYSIS OF ALLELE-SPECIFIC EXPRESSION, TRANSCRIPTION FACTOR BINDING, AND CHROMATIN STATE IN MULTIPLE HUMAN TISSUES.

Joel Rozowsky[1], Timur Galeev[1], Xiangmeng Kong[1], Min xu[1], Gamze Gursoy[1], Chengfei Yan[1], Alex Dobin[2], Anna Vlasova[3], Roderic Guigo[3], Michael Schatz[4], Thomas Gingeras[2], Mark Gerstein[1]

[1]Yale University, Program in Computational Biology & Bioinformatics, New Haven, CT, [2]Cold Spring Harbor Laboratory, ., Cold Spring Harbor, NY, [3]The Barcelona Institute of Science and Technology, DepartmeCentre for Genomic Regulation (CRGnts of Computer Science and Biology, Barcelona, Spain, [4]Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD

EN-TEx, a collaborative project between the ENCODE and the GTEx consortia, has employed an array of whole-genome sequencing technologies: Illumina short-read sequencing, PacBio long-read sequencing, 10X Genomics Chromium, and Hi-C, to identify and phase single nucleotide, short indel, and structural variants of four human donors. Samples from multiple tissues of these donors have been profiled with a wide range of assays, including RNA-seq, transcription factor and histone modification ChIP-seq, ATAC-seq, and DNase-seq.

Leveraging these high quality personal and functional genomics datasets, we investigate allele-specific expression, transcription factor binding, histone modification, and chromatin accessibility across a multitude of human tissues. Analyses based on allelic read counts are known to be very sensitive not only to the quality of variant calls but also to mapping biases. Thus, in order to eliminate the associated reference bias and to improve read alignment, especially, in the regions with high genomic variation, our pipeline utilizes diploid personal genomes. We constructed such personal genomes for each of the donors by incorporating their variants into the reference genome sequence. Overall, we identify and characterize thousands of heterozygous variants and hundreds of genes and peaks associated with allele-specific expression, binding, and chromatin state. Integrative analyses of personal variation, known quantitative trait loci, and the relationships of allelic imbalances in different assays reveal variants and mechanisms that may potentially be causing allelic skew of some of the detected allele-specific genes.

# CHIP-EAT: FROM RAW CHIP-SEQ READS TO HIGH QUALITY TFBS PREDICTION

Marius Gheorghe, Anthony Mathelier

Norwegian Centre for Molecular Medicine, University of Oslo, Oslo, Norway

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) represents the most popular experimental assay to identify the genomic regions, so called ChIP-seq peaks, where transcription factors (TFs) bind to the DNA in vivo. The ever increasing number of publicly available ChIP-seq data sets provides an unprecedented opportunity to develop computational tools to infer the precise locations of the TF binding sites (TFBSs) within the ChIP-seq peaks, by combining both computational and experimental evidence of direct TF-DNA interactions. While TFBSs are traditionally modelled through position weight matrices (PWMs), more advanced computational methods have been recently developed to incorporate nucleotide dependencies, variable spacing, and DNA conformation in their models. These methodologies highlight that a one-fits-all model for TFBS prediction is not applicable. In this scope, we have developed ChIP-eat, a uniform ChIP-seq data processing pipeline, from raw data to accurate, TF-specific TFBS prediction. With the latest release of the ReMap database (remap.cisreg.eu), we predicted ~80M binding regions in the human genome (hg38 version) for 485 transcriptional regulators, by uniformly processing 2,829 ChIP-seq data sets available from public sources. After the ChIP-seq peak calling, we used an entropy-based algorithm to predict direct TF-DNA interactions from four different TFBS modeling approaches, by computing the enrichment of the predicted TFBSs at the ChIP-seq peak-summits (where the highest number of reads mapped). Along with PWMs, we evaluated binding energy models, transcription factor flexible models, and DNA-shaped-based models for each ChIP-seq data set.

We applied ChIP-eat on 1,160 ENCODE and 1931 GEO and Array Express (AE) data sets covering a total of 496 distinct TFs. Out of those, 727 ENCODE and 1256 GEO and AE data sets mapped to a JASPAR profile, together covering 232 distinct TFs, and were subsequently processed for TFBS prediction. Our work culminates with the generation of a large, publicly available collection of uniformly processed ChIP-seq data sets from which we obtained ChIP-seq peaks and accurate TFBS predictions derived from multiple TFBS prediction models per data set. Altogether, we predict direct TF-DNA interactions covering about 4.3% of the human genome.

# FOLDING, UNFOLDING AND REFOLDING OF GENOMES

Job Dekker

Howard Hughes Medical Institute, University of Massachusetts Medical School, Program in Systems Biology, Worcester, MA

In order to understand how the genome operates, we need to understand not only the linear encoding of information along chromosomes, but also its 3-dimensional organization. The spatial organization of the genome is critical for gene regulation, genome stability and faithful transmission of chromosomes to daughter cells. Chromosome Conformation Capture-based technologies and live cell and high-resolution imaging approaches are now widely used to determine how cells fold their chromosomes, to discover the processes that drive the spatial organization of genomes and to identify the mechanisms by which this organization contributes to genome regulation and activity.

At the nuclear level, chromosomes are compartmentalized into large multi-Mb compartments that are either active and open or inactive and closed. At the scale of hundreds of Kb chromosomes form Topologically Associating domains (TADs). Gene regulation occurs mostly within TADs through long-range looping interactions between genes and regulatory elements. In mitotic cells chromosome conformation is completely different: inside compact metaphase chromosomes the genome folds as longitudinally compressed randomly positioned loop arrays, consistent with classical models proposed by the Laemmli lab.

I will present recent insights into the detailed spatial arrangement of the genome at different stages of the cell cycle, and the folding pathways by which these states interconvert. I will describe new folding intermediates and the roles of key protein complexes such as CTCF, cohesin and condensins. Finally I will discuss a common mechanism by which the genome folds, unfolds and refolds to facilitate gene regulation and chromosome transmission.

# INTEGRATIVE MULTI-OMICS ANALYSES OF iPSC-DERIVED BRAIN ORGANOIDS IDENTIFY EARLY DETERMINANTS OF HUMAN CORTICAL DEVELOPMENT

Anahita Amiri*[1], Gianfilippo Coppola*[1], Soraya Scuderi*[1], Feinan Wu*[1], Tanmoy Roychowdhury[3], Mark Gerstein[2], Nenad Sestan[4], Alexej Abyzov[3], Flora M Vaccarino[1,4]

[1]Yale University, Child Study Center, New Haven, CT, [2]Yale University, Molecular Biology and Biophysics, New Haven, CT, [3]Mayo Clinic, Department of Health Sciences Research, Rochester, MN, [4]Yale University, Neuroscience, New Haven, CT

Gene regulatory regions of the human genome active in the prenatal human cerebral cortex are thought to drive human brain evolution, and contain loci that confer risk for neuropsychiatric disorders. These stages are impossible to model in a longitudinal, dynamic fashion using postmortem brain tissue. Here, by comparing human forebrain organoids derived from induced pluripotent stem cells (iPSCs) and isogenic fetal cerebral cortex, we demonstrate that on the transcriptome and epigenome level organoids model embryonic and early fetal cerebral cortical development. By combined analyses of histone marks, transcriptome and chromatin conformation in organoids and fetal cortex, we reveal the longitudinal dynamics of transcripts and enhancer elements at stages that bridge neural stem cell proliferation with neurogenesis. We found that the transition from neural stem cells to cortical progenitors is characterized by the largest number of differentially expressed genes and differentially active enhancers, the majority of which were unique to this transition. We constructed networks of transcripts or enhancers modules, each exhibiting correlated patterns of expression/activity across samples. These modules could be grouped in "supermodules" with coordinated increases or decreases in activity across developmental transitions, suggesting co-regulation by common upstream mechanisms. Specific transcriptome and enhancer modules were strongly enriched with autism-associated and mental disability-related genes, suggesting the likely very early onset of these diseases and predicting genes and regulatory elements related to the disease onset.

*Authors contributed equally

# RADICL-SEQ: A NOVEL TECHNOLOGY FOR GENOME-WIDE MAPPING OF RNA-CHROMATIN INTERACTIONS

Alessandro Bonetti[1,2], Kosuke Hashimoto[1], Ana M Suzuki[1,3], Giovanni Pascarella[1], Erik Arner[1], Christopher Cameron[4], Shuhei Noguchi[1], Nicholas Luscombe[5], Mathieu Blanchette[4], Michiel De Hoon[1], Charles Plessy[1], Gonçalo Castelo-Branco[2], Valerio Orlando[6], Piero Carninci[1]

[1]RIKEN, Center for Life Science Technologies, Yokohama, Japan, [2]Karolinska Institutet, Department of Medical Biochemistry and Biophysics, Stockholm, Sweden, [3]Karolinska Institutet, Department of Medicine, Stockholm, Sweden, [4]McGill University, The School of Computer Science, Quebec, Canada, [5]The Francis Crick Institute, London, United Kingdom, [6]King Abdullah University of Science and Technology, Biological and Environmental Science and Engineering Division, Thuwal, Saudi Arabia

The potential regulatory functions of long non-coding RNAs (lncRNAs) are now broadly accepted. However, their functional characterization hasn't been advanced enough yet. Most lncRNAs are localized in the cell nuclei. In order to explore their regulatory functions, sophisticated technologies have been developed to analyze RNA-chromatin interactions in these years. However, genome-wide mapping of the interactions at high-throughput is still challenging.
To overcome this difficulty, we recently developed a novel technology, RNA and DNA Interacting Complexes Ligated and sequenced (RADICL-seq) that maps genome-wide RNA-chromatin interactions in intact nuclei. RADICL-seq utilizes proximity ligation methodology in the library preparation. It broadly identifies the target genomic regions for coding and non-coding RNAs and precisely map RNA and DNA with the directional information of the RNA tags.
We confirmed reproducible RNA-chromatin interactions detected by RADICL-seq identifying the genome occupancy of 14,000 transcripts including 1,000 lncRNAs. RADICL-seq also revealed specific patterns of RNA-chromatin interactions. Application of RADICL-seq in mouse embryonic stem cells (mESCs) revealed different chromatin interaction patterns for mRNAs and lncRNAs, uncovering the uncharted presence for intronic sequences mediating RNA-chromatin interactions. The annotation of RNA-DNA pairs mapped to the genome revealed that the RNAs primarily originated from genic regions with a dominant contribution from intronic reads, whereas the DNAs had an equal contribution from genic and intergenic regions. We also observed on average 25% of RNA-DNA pairs include RNA sequences intersecting with Repeat elements (REs). Remarkably, there was an intra-chromosomal pattern with different RE classes being enriched over specific distance ranges.

# CHARTING THE DIVERSIFICATION OF MAMMALIAN CELLS AT WHOLE ORGANISM SCALE

Jonathan A Griffiths[1], Blanca Pijuan-Sala[2,3], Fernando J Calero-Nieto[2,3], Carla Mulas[3], Wajid Jawaid[2,3,4], Carolina Guibentif[2,3], Ximena Ibarra-Soria[1], Hisham Mohammed[5], Jennifer Nichols[3], Wolf Reik[5,6,7], John C Marioni[1,6,8], Berthold Göttgens[2,3]

[1]CRUK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom, [2]Department of Haematology, University of Cambridge, Cambridge, United Kingdom, [3]Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom, [4]Department of Paediatric Surgery, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom, [5]Epigenetics Programme, Babraham Institute, Cambridge, United Kingdom, [6]Wellcome Trust Sanger Institute, Single-Cell Genomics Centre, Cambridge, United Kingdom, [7]Centre for Trophoblast Research, University of Cambridge, Cambridge, United Kingdom, [8]EMBL-European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom

Decision-making is a critical component of cellular behaviour, with errors giving rise to e.g., developmental failures or cancer. Embryonic development is an effective system for studying decision-making, as it provides a large set of robust, biologically relevant, relatively well-studied decision points. However, transcriptomic assays have historically been hamstrung by the extreme spatial complexity and molecular heterogeneity of the developing embryo.

We have captured 100,000 single cells for single-cell RNAseq from whole mouse embryos during gastrulation and organogenesis, spanning days 6.5 to 8.5 of development, including embryonic and extraembryonic tissues. Cells were sampled every six hours, providing a continuous molecular characterisation of these processes. We highlight the utility of this rich dataset in three ways. First, we reconstruct the development of the three germ layers and their lineages. Second, we identify non-negligible populations of rare cell types such as primordial germ cells. Third, we dissect one specific developmental process by considering the contribution of extraembryonic visceral endoderm to embryonic endoderm lineages.

# CLOCK-DEPENDENT CHROMATIN TOPOLOGY MODULATES CIRCADIAN TRANSCRIPTION AND BEHAVIOR.

Jerome Mermet*[1], Jake Yeung*[1], Clemence Hurni[1], Daniel Mauvoisin[1], Kyle Gustafson[1], Celine Jouffe[2], Damien Nicolas[1], Yann Emmenegger[3], Cedric Gobet[1,2], Paul Franken[3], Frederic Gachon[1], Felix Naef[1]

[1]Ecole Polytechnique Fédérale de Lausanne, School of Life Sciences, Lausanne, Switzerland, [2]Nestle Institute of Health Sciences, NIHS, Lausanne, Switzerland, [3]University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland

The circadian clock in animals orchestrates widespread oscillatory gene expression programs. Nearly half of all genes in the mouse genome oscillate with circadian rhythm somewhere in the body (Zhang et al 2014 *PNAS*). These regulatory programs underlie 24-hour rhythms in behavior and physiology, making it a powerful model to study the function of regulatory elements at the scale of single cells, tissues, and animal behavior. Studies have shown the possible role of transcription factors and chromatin marks in controlling cyclic gene expression (Fang et al 2014 *Cell*, Sobel et al 2017 *PLoS Biology*). However, how daily active enhancers modulate rhythmic gene transcription in mammalian tissues remains uncharted. Here, we discover oscillatory promoter-enhancer interactions along the 24-hour cycle in mouse liver and kidney, and these oscillations depended on the clock transcription factor BMAL1. Oscillations in promoter-enhancer interactions are accompanied by daily rhythms in H3K27ac and DNase-I hypersensitivity that are localized to the interacting regions, suggesting that coordinated epigenetic modifications underlie promoter-enhancer dynamics. Furthermore, deleting a contacted intronic enhancer element in the *Cry1* gene (*Cry1*Δe mutant) was sufficient to compromise the rhythmic chromatin contacts. The *Cry1*Δe mutant mice display reduced *Cry1* transcriptional burst frequency and, remarkably, a shortened circadian period of locomotor activity rhythms. Our results establish oscillating and clock-controlled promoter-enhancer looping as a new regulatory layer underlying circadian transcription and behavior.

# GENETIC AND EPIGENETIC FINE MAPPING OF COMPLEX TRAIT ASSOCIATED LOCI IN THE HUMAN LIVER

Minal Caliskan[1], Julian Segert[1], H. Shanker Rao[1], Andrea M Berrido[1], Marcia Holsbach Beltrame[1], Marco Trizzino[1], YoSon Park[1], Robert C Bauer[5], Nicholas J Hand[1], Kim M Olthoff[2], Abraham Shaked[2], Daniel J Rader[1,3], Barbara E Engelhardt[4], Christopher D Brown[1]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]University of Pennsylvania, Transplant Surgery, Philadelphia, PA, [3]University of Pennsylvania, Medicine, Philadelphia, PA, [4]Princeton University, Computer Science, Princeton, NJ, [5]Columbia University, Dept. of Medicine, New York, NY

The liver has a central role in detoxification of endogenous and exogenous toxins, synthesis of essential proteins, and regulation of carbohydrate, lipid, and drug metabolism. As such, the liver is associated with a diverse range of clinically important human traits and was recently reported as one of the most critical tissues for explaining the mechanisms at genome wide association study (GWAS) loci. In this study, we obtained genome-wide genotype, RNA-seq, and ChIP-seq data on H3K27ac and H3K4me3 histone modifications in up to 241 human liver tissues. We identified 131,293 and 68,600 genomic regions enriched for H3K27ac and H3K4me3 modifications, and identified cis-hQTLs for 910 and 96 of them, respectively. Despite significant overlap, we found that liver hQTLs are often not found in lymphoblastoid cell lines. In addition to performing the first hQTL study of the liver, we mapped cis-eQTLs for a total of 2,625 genes; liver cis-eQTLs were significantly more likely to overlap liver histone peaks than histone peaks identified in any other non-liver ENCODE cell type, reflecting the importance of annotating and mapping noncoding functional loci in the relevant tissue type. To fine-map causal regulatory variants with evidence of shared causality between expression and histone modification, we applied a Bayesian genetic colocalization approach and identified 68 eQTL-hQTL pairs with significant colocalization signal. Interestingly, histone peaks often do not colocalize with their nearest gene; we found a median number of two intervening genes between colocalized peak and gene pairs (range:0-16). Next, to help understand the mechanisms of GWAS loci we integrated our hQTL and eQTL results with GWAS summary statistics for nine liver related phenotypes including coronary artery disease, blood lipids, and blood pressure phenotypes. We identified 131 GWAS-eQTL and 15 GWAS-hQTL colocalizations. We prioritized GWAS-histone peak colocalizations for further functional characterization. With a combination of experimental approaches in multiple liver-related model systems, we are able to replicate the functionality of the fine-mapped GWAS-colocalizing histone peaks. These findings highlight the benefits of integrating multiple cellular traits for the identification and characterization of disease-causing variants and contribute to basic understanding of genetic and epigenetic regulation of gene expression in human liver tissue.

# DISSECTING TISSUE-SPECIFIC FUNCTIONAL NETWORKS ASSOCIATED WITH 16P11.2 RECIPROCAL GENOMIC DISORDER USING CRISPR ENGINEERED HUMAN iPS AND MOUSE MODELS

Parisa Razaz[1,2,3], Derek J Tai[1,2,3], Serkan Erdin[1,2,3], Tatsiana Aneichyk[1,2,3], Thomas Arbogast[4], Ashok Ragavendran[1], Alexei Stortchevoi[1,2], Benjamin B Currall[1,2,3], Celine E de Esch [1,2,3], Elisabetta Morini[1,2], Weiyuan Ma[1,2], Raymond J Kelleher[1,2], Christelle Golzio[4,5], Nicholas Katsanis[4], James F Gusella[1,2,3,6], Michael E Talkowski[1,2,3,6]

[1]Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital, Boston, MA, [2]Department of Neurology, Harvard Medical School, Boston, MA, [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Boston, MA, [4]Center for Human Disease Modeling, Duke University Medical Center, Durham, NC, [5]Institut de Génétique et de Biologie Moléculaire et Cellulaire, Médecine translationnelle et neurogénétique, Illkirch, France, [6]Department of Genetics, Harvard Medical School, Boston, MA

Reciprocal genomic disorders (RGDs) represent a recurrent class of copy number variants (CNVs) that collectively comprise a major contributor to neurodevelopmental disorders (NDD) and altered anthropometric traits. Here, we systematically dissected the functional networks associated with 16p11.2 RGD from transcriptome analyses of 70 mice with reciprocal CNV of the syntenic 7qF3 region across cortex, striatum, and cerebellum, as well as liver, white and brown adipose tissues in a subset of 16 mice (n=250 samples). We integrated these data with brain tissues from a *Kctd13* mouse model (a putative driver of 16p11.2 neuroanatomical phenotypes, n=50), and CRISPR-engineered, isogenic 16p11.2 iPSC-derived NSCs (n=25) and induced neurons (n=27). The strongest magnitude of effect sizes from 7qF3 were observed across brain regions by comparison to non-brain tissues (cortex 7qF3 region average p-value=8.80E-35; non-brain p=0.0013), reflecting the ~3x higher basal expression changes. Coexpression network analyses isolated a consistent module of 16p11.2 genes, as well as a module that was highly enriched for constrained genes (ExAC pLI≥0.9), autism-associated genes, early fetal development coexpression networks derived from BrainSpan, and neurological phenotypes and processes. Differentially expressed genes (DEGs) were enriched in this 'constrained' module network; moreover, DEGs from the *Kctd13* mouse coalesced into this same module (cortex enrichment p=7.82E-41), suggesting overlap in altered transcriptional networks between full length CNV and deletion of *KCTD13* alone. These analyses identify a tissue-specific impact of 16p11.2 RGD that converges on a module of co-expressed genes that are intolerant to genetic perturbation and associated with critical processes in human neurodevelopment.

# INFORMING HUMAN GENETIC VARIATION AND THERAPEUTIC ENTRY POINTS THROUGH MODIFIER SCREENS IN MICE.

Julie Ruston[1], Ashlee Dargie[1], Christine Taylor[1], Adebola Enikanolaiye[1], Monica J Justice[1,2]

[1]The Hospital for Sick Children, Genetics and Genome Biology, Toronto, Canada, [2]The University of Toronto, Molecular Genetics, Toronto, Canada

All genetic diseases are characterized by variability in the onset and expression of clinical features, most often due to second site gene modifiers. Identifying the modifiers that regulate this variation represents a transformative discovery for the disease they alter, shifting understanding of pathogenesis and providing avenues for diagnosis, prognosis, and therapy development. However, pinpointing the crucial genetic variants that modify inherited disease in human sequence is a daunting task. Mouse genetic modifier screens are an avenue to understand pathways that contribute to human disease presentation. I initiated a modifier screen in a mouse model for Rett syndrome, a neurological condition caused by mutations in methyl CpG binding protein 2 (MeCP2), which has no effective treatments. MeCP2 is a regulator of key activities in the brain and body, with mutations impacting both adult and childhood neuropsychiatric and immune disorders. Therefore, MeCP2 is an archetype for understanding key pathways that regulate brain and body function.

We used random unbiased mutagenesis with N-ethyl-N-nitrosourea (ENU) to isolate mutations that suppressed clinical signs and improved overall health in a MeCP2 mouse model. Whole exome sequencing of 91 lines carrying modifiers that improve health traits and prolong life has identified over 100 candidate genes. Many of the lines carry more than one modifier locus, while different mutations in some candidate genes are found in multiple independent lines, generating allelic series. One of our earliest modifiers pointed to lipid metabolism as being perturbed in Rett syndrome, suggesting possibilities for treating Rett through metabolic modulation. The remaining candidate genes fall into a limited number of pathways, which reveal new facets of MeCP2 biology. MeCP2 provides a bridge for the repressor complex NCoR1/SMRT/HDAC3 to regulate its targets on DNA, and the modifiers suggest that this complex is key to Rett syndrome pathology. Altogether, our data suggest that multiple factors will be required to reverse disease entirely, requiring combination therapies. A similar genetic approach could be exploited to identify unexplored genetic variation for other human diseases, informing disease presentation and opening a new field for translational discovery.

# THE GENOME10K VERTEBRATE GENOMES PROJECT: BUILDING *DE NOVO* REFERENCE GENOMES FOR ALL VERTEBRATE ORDERS

Arang Rhie[1], Shane A McCarthy[2], Olivier Fedrigo[3], William Chow[4], Zemin Ning[5], Joana Damas[6], Marcela Uliano-Silva[7], Martin Pippel[8], Sergey Koren[1], Kerstin Howe[4], Harris Lewin[6], Richard Durbin[2], Gene Myers[8], Adam M Phillippy[1], Erich D Jarvis[3]

[1]NIH, NHGRI, Bethesda, MD, [2]University of Cambridge, Department of Genetics, Cambridge, United Kingdom, [3]The Rockefeller University, The Vertebrate Genomes Laboratory, New York, NY, [4]Wellcometrust Sanger Institute, Genome Reference Informatics Team, Hinxton, United Kingdom, [5]Wellcome Trust Sanger Institute, High Performance Assembly Group, Hinxton, United Kingdom, [6]UC Davis, Evolution and Ecology, Davis, CA, [7]Berlin Center for Genomics, Biodiversity Research, Berlin, Germany, [8]Max Planck Institute, Molecular Cell Biology and Genetics, Dresden, Germany

The Genome10K Vertebrate Genomes Project (VGP) consortium is an international effort, spanning over 50 institutions on nearly all continents, that aims to create a digital open-access genome library of at least one high-quality, near-gapless, phased and annotated chromosomal-level assembly of all extant vertebrate species. The initial phase of this project is focused on finishing one species from each vertebrate order, totaling 260 individual species, to a quality standard of >1 Mb N50 contig size, >10 Mb N50 scaffold size, average bp quality >QV40, and 90% of the sequence assigned to chromosomes.
High-quality genome assemblies are needed to address fundamental questions in biology and disease, to identify species most genetically at risk for extinction, and to preserve genetic information for posterity. However, most vertebrate genomes are lacking such reference. Currently, only 9 vertebrate species have reference genomes meeting the VGP quality standard. This is reflecting the high cost previously required to build them. However, with the maturation of long-read sequencing and long-range scaffolding technologies, it is now possible to construct reference-grade assemblies, *de novo*, at reasonable cost. The VGP has begun collecting and sequencing ordinal samples using 4 such emerging technologies: PacBio long reads, 10X Genomics linked reads, Bionano optical maps, and Arima Genomics Hi-C libraries.
In parallel, the VGP assembly working group has been comparing and evaluating sequencing and assembly strategies using an initial set of species including mammals, birds, and fishes. The current assembly process involves contig generation using PacBio, followed by scaffolding using 10X Genomics, Bionano, and Hi-C. Draft assemblies are then evaluated for correctness using the gEVAL platform and genome-to-genome alignments. An improved, comprehensive assembly strategy is under continued development, and these new methods are aimed at better separation of haplotypes during assembly. In the interim, version 1 assemblies will be submitted to the public sequence repositories, and all raw data will be released as it is generated in coordination with DNAnexus and Amazon Web Services at the "genomeark" (http://genomeark.s3.amazonaws.com). This includes an initial batch of ~10 new ordinal genomes to be completed by the spring of 2018.

## *DE NOVO* ASSEMBLY OF MAMMALIAN GENOMES WITH CHROMOSOME-LENGTH SCAFFOLDS, FROM SHORT READS, FOR UNDER $1000

Olga Dudchenko[1,2,3,4], Muhammad S Shamim*[1,2,3,5], Sanjit S Batra*[1,6], Neva C Durand[1,2,3], Nathaniel T Musial[1,7], Ragib Mostofa[1,3], Melanie Pham[1,2,3], Brian Glenn St Hilaire[1,2,3], Weijie Yao[1,2,3], Elena Stamenova[1,8], Marie Hoeger[1], Sarah K Nyquist[1,9], Valeriya Korchina[1,10], Kelcie Pletch[11], Joseph P Flanagan[11], Arina D Omer[1,2,3], Erez Lieberman Aiden[1,2,3,4]

[1]Baylor College of Medicine, The Center for Genome Architecture, Houston, TX, [2]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, [3]Rice University, Department of Computational and Applied Mathematics, Houston, TX, [4]Rice University, Center for Theoretical and Biological Physics, Houston, TX, [5]Baylor College of Medicine, Medical Scientist Training Program, Houston, TX, [6]University of California, Department of Computer Science, Berkeley, CA, [7]University of Texas at Dallas, Erik Jonsson School of Engineering and Computer Science, Richardson, TX, [8]Broad Institute of MIT and Harvard, Epigenomics Program, Cambridge, MA, [9]Massachusetts Institute of Technology, Computational and Systems Biology, Cambridge, MA, [10]Des Moines University, DMU College of Osteopathic Medicine, Des Moines, IA, [11]Houston Zoo, Veterinary Clinic, Houston, TX

Hi-C contact maps are valuable for genome assembly (Lieberman-Aiden, van Berkum et al. 2009; Burton et al. 2013; Dudchenko et al. 2017). Recently, we developed 3D-DNA, an automated pipeline for using Hi-C data to assemble genomes (Dudchenko et al. Science 2017), Juicebox, a system for the visual exploration of Hi-C data (Durand, Robinson et al. Cell Systems 2016), and the Juicebox Assembly Tools, which provide a point-and-click interface for using Hi-C heatmaps to identify and correct errors in a genome assembly (Dudchenko et al., bioRxiv 2018). In this talk, we show that 3D-DNA and the Juicebox Assembly Tools greatly reduce the cost of accurately assembling complex eukaryotic genomes, enabling end-to-end genome assemblies from short reads alone. To illustrate, we generated *de novo* assemblies with chromosome-length scaffolds for three mammals: the wombat, *Vombatus ursinus* (3.3Gb), the Virginia opossum, *Didelphis virginiana* (3.3Gb), and the raccoon, *Procyon lotor* (2.5Gb). The only inputs for each assembly were Illumina reads from a short insert DNA-Seq library (300 million Illumina reads, maximum length 2x150 bases) and an *in situ* Hi-C library (100 million Illumina reads, maximum read length 2x150 bases), which cost <$1000. *These authors contributed equally to this work.

# MULTIPLE SELECTIVE SWEEPS REMOVED NEANDERTHAL ADMIXTURE ON THE X CHROMOSOME IN OUT-OF-AFRICA POPULATIONS

Laurits Skov, Moises C Marcia, Elise Lucotte, Mikkel H Schierup, Kasper Munch

Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

The dominating spread of modern humans out of Africa (60 ky) was associated with a number of dramatic events, including a severe population bottleneck and interbreeding with archaic species (Neanderthals and Denisovans). Several fossils in Israel dated to about 180 ky ago, suggest that this main exodus was preceded by earlier waves out of Africa. This is corroborated by genetic studies suggesting that an earlier out of Africa event (100-200 ky) may have been genetically replaced almost completely by the subsequent main exodus (60 ky). The X chromosome offers genetic information different from that of the autosomes in that it has a lower effective population size sensitive to sex-specific migration. More importantly, male hemizygosity exposes recessive variants to selection in males, and the interaction between the very different X and Y chromosomes may lead to unique opportunities for sexual antagonistic selection as well as meiotic drive in the form of direct competition between the X and Y for transmission in male meiosis. In an analysis of the Simons genome diversity data set, we discover megabase long segments of the X-chromosome that have lost most of the variation in out-of-Africa human populations. The loss of diversity is consistent with very strong selective sweeps occurring early in the out-of-Africa expansion. The genomic positions of these sweeps overlap hotspots of selection in the great apes, which may be subject to genomic conflict between the X and the Y chromosome (meiotic drive). Using a new approach to inferring individual Neanderthal-derived haplotypes, which do not rely on the Neanderthal reference genome, we discover that the swept regions are completely devoid of Neanderthal ancestry. The swept haplotypes show a closer relation to East African populations than the rest of the X chromosome. These findings prompt us to speculate that the swept regions are remnants of an initial out-of-Africa expansion population that was subsequently replaced by a later expansion. However, the initial expansion may have contained strong meiotic drivers that resisted this genetic replacement.

# EXPLORING THE JOINT DISTRIBUTION OF FITNESS EFFECTS FOR BENEFICIAL MUTATIONS IN YEAST

Lucas Herissant[1], Dave Yuan[2], Parris Humphrey[3], Milo Johnson[3], Atish Agarwala[4], Daniel Fisher[4], Michael Desai[3], Dmitri Petrov[2], Gavin Sherlock[1]

[1]Stanford University, Genetics, Stanford, CA, [2]Stanford University, Biology, Stanford, CA, [3]Harvard University, Organismic and Evolutionary Biology, Cambridge, MA, [4]Stanford University, Applied Physics, Stanford, CA

Pathogenicity, drug resistance and cancer progression are examples of mutation-driven processes, where increased selective advantage is conferred upon cells carrying beneficial mutations. However, while these mutations may be beneficial in one specific condition, they may be deleterious in other conditions, a phenomenon known as Antagonistic Pleiotropy (AP). AP is thought to lead to evolutionary trade-offs and the persistence of deleterious alleles. However, the prevalence of AP is unknown, nor is it known whether certain genes or pathways are more likely to be involved in AP, and under which conditions? We have set out to systematically determine how beneficial mutations selected in one environment fare in others.

We previously developed a lineage tracking system to follow the dynamics of adaptive evolution. By tracking the lineage tags over time, we showed that ~20,000 lineages gained a beneficial mutation during evolution over only 240 generations in limiting glucose. Furthermore, we were able to generate a distribution of fitness effects for these lineages (Levy, Blundell et al, Nature (2015)). By isolating and sequencing adaptive clones from independent lineages we found that the RAS/cAMP/PKA pathway and the Tor pathway are frequently targets for adaptation under the evolutionary conditions, uncovering almost 80 mutations in these pathways. In several cases where a gene has a paralog, beneficial mutations are recovered in one paralog significantly more frequently than that other. We have also found that even when mutations affect the same pathway, that the fitness conferred by mutations in a given gene tends to be specific for that gene, and distinct from the fitness effect of mutations in other genes in the pathway (Venkataram et al, Cell (2016)).

To generalize these findings, we have developed an augmented barcode system, whereby a second barcode can be used to encode the evolutionary condition. We have used this system to evolve both haploid and diploid yeast in several environments, then isolated adaptive clones from each of the conditions, pooled them, and remeasured their fitness across each of the conditions, to understand how the ways in which beneficial mutants may either be generalists or specialists, and the ways in which they might tradeoff.

# THE GENOME'S RESERVOIR OF BENEFICIAL PROTO-GENES.

Brian Hsu[1], Nelson Coelho-Castilho[2], Nikolaos Vakirlis[3], Trey Ideker[1], Anne-Ruxandra Carvunis[2]

[1]University of California, San Diego, Medicine, La Jolla, CA, [2]University of Pittsburgh School of Medicine, Computational and Systems Biology, Pittsburgh, PA, [3]Trinity College Dublin, Dublin, Ireland

One of the most astonishing phenomena in genome evolution is the *de novo* emergence of protein-coding genes in sequences that were previously non-genic. This radical transition requires non-genic sequences to become transcribed, to acquire open reading frames, and that the corresponding non-genic transcripts access the translation machinery. However, it is unlikely that the resulting polypeptides would spontaneously be fully integrated in cellular networks in an optimal fashion. Instead they are thought to represent intermediate "proto-gene" stages that expose the genome's reservoir of cryptic genetic variation to the action of natural selection. The majority will likely return to a non-genic state, but a subset may transition into *de novo* genes, for instance if their expression is beneficial to the organism. In support for this model, widespread translation of non-genic transcripts has recently been documented across numerous species. Whether these translation events really do carry adaptive potential has not yet been demonstrated.

Here, we experimentally assessed the fitness impact of proto-gene expression in the yeast *Saccharomyces cerevisiae*. We systematically overexpressed hundreds of naturally-occurring proto-genes and measured the fitness of each resulting strain relative to wild type in multiple environmental conditions. We found that overexpression of proto-genes provided the organism with a beneficial growth advantage significantly more often than the overexpression of conserved genes did across all conditions tested. The beneficial proto-genes we identified are short, uncharacterized coding sequences that are for the most part absent from closely related species. We propose that future regulatory mutations that naturally increase the expression level of proto-gene sequences may be adaptive, fix and participate in the birth of novel protein-coding genes.

CATALOG OF 91 MILLION VARIANTS EXTRACTED FROM WHOLE GENOME SEQUENCE OF 722 CANIDS REVEALS NEW VARIANTS ASSOCIATED WITH MORPHOLOGY, LIFE-SPAN AND BEHAVIOR.

Jocelyn Plassais, Brian W Davis, Danielle M Karyadi, Heidi G Parker, Alex Harris, Brennan Decker, Elaine A Ostrander

National Institutes of Health, National Human Genome Research Institute, Bethesda, MD

Human selection has divided domestic dogs into nearly 500 breeds, resulting in a host of population-enriched genomic changes. As a result, modern breeds are characterized by variation in behavior, temperament, disease susceptibility, and aesthetics. We have assembled a catalogue of 722 canine whole genome sequences (WGS), documenting over 91 million single nucleotide (SNV) and structural variants (SV), resulting in the largest catalog of genomic variation for any companion animal species. Using this resource, we undertook comprehensive genome-wide scans, inclusive of over 144 breeds, to identify loci associated with morphologic features, life-span and behavior. Building upon our previously published association study linking a small number of loci with reduced body size, analysis of our extensive catalog of genomic variation reveals a dozen newly associated genes (LCORL, ESR1, ZNF608, ADAMTSL3, ADAMTS9, HNF4G, R3HDM1) and associated mutations, explaining more than 85% of variation in breed height, weight, and muscle and fat distribution. In aggregate, these data reveal an increasingly complex pattern of genetic variation in specific transcripts, promoters and long non-coding RNA controlling both body size and shape. These data also reveal a strong negative correlation ($r = 0.76$) between body size and breed longevity for a subset of variants, with higher p-values observed when both weight and life-span data are combined. In addition to genes controlling appearance, our study revealed loci on canine chromosomes 10 and 24 associated with stereotypic dog behaviors such as boldness or aggressiveness, paralleling results from human psychiatric disorders. Finally, we demonstrate how this large catalogue is being used to both find and validate variants associated with common and rare disorders of interest to human health.

# 3D-MODELLING OF HI-C DATA TO INVESTIGATE THE SPATIAL ORGANISATION OF THE CANINE GENOME

Bobbie J Cansdale, Claire M Wade

University of Sydney, Faculty of Science, Sydney, Australia

The three-dimensional structure of the genome is non-random and important for several key biological processes including the regulation of gene expression. Determining this structure, as well as the sequence itself, is necessary to further genome biology research. Topologically associating domains (TADs) are a main feature of chromatin organisation. These are clusters of genes that are functionally co-regulated, with boundary regions enriched for features such as CTCF binding sites, transfer RNAs, and SINE retrotransposons. Chromosome conformation capture (3C) based approaches, including Hi-C, can provide valuable insight into the spatial organisation of chromatin fibre. Recently computational frameworks have become available to use this data to create 3D representations of the genome, providing novel insights compared to the standard interaction matrices alone.

Here we present the first investigation into the 3D structure of several phenotypic loci in the canine genome. The domestic dog is separated into pure breeding populations, with relatively reduced diversity within breeds, which provides clear population delimitation and a less variable genomic landscape in which to investigate a variety of genomic features. Canine phenotypic loci have been well described, making these useful candidate regions to study. We have analysed published canine Hi-C data with validated pipelines to identify TADs and produce 3D representations of several phenotypic loci to better understand how these prominent regions are organised. Knowledge of these structures will allow investigation as to how they relate to the nearby genes and other genomic features. Simultaneous investigation of the 3D model, interaction matrices, and linear tracks of features will enable us to gain a comprehensive view of these regions. This work provides unique insights into these loci not previously possible with standard sequence-based analysis.

# MEASUREMENT OF GENOME-WIDE SELECTIVE CONSTRAINT ON HUMAN GENE EXPRESSION

Emily C Glassberg*[1], Ziyue Gao*[2,3], Arbel Harpak[1], Xun Lan[2,4], 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 University, Stanford, CA, [4]Tsinghua University School of Medicine, Department of Basic Medical Sciences, Beijing, China

*authors contributed equally to the work.

Gene expression variation likely underlies phenotypic variation in human complex traits. Selection on fitness-relevant complex traits may therefore be reflected in constraint on gene expression levels. Here, we explore the effects of stabilizing selection on cis-regulatory genetic variation in humans. Analyzing patterns of expression variation at polymorphic gene duplicates, we find evidence for selection against large (1.5X) increases in gene expression. To further characterize the landscape of gene regulation, we detect 6,855 cis-eQTLs around 4,923 genes. Comparisons of eQTL allele frequencies and effect sizes across gene sets confirm the existence of tightly constrained genes; genes intolerant to loss-of-function variation are depleted for large-effect cis-eQTLs. However, the many common, modest-effect eQTLs observed in our study and others argues against strong, global constraint on expression levels. Additionally, we observe that genes with no significant eQTLs still display allele-specific expression. Within an individual, the degree of allele-specific expression is correlated with the number of cis-heterozygous sites, suggesting that rare and/or small-effect variants, undetectable as eQTLs, contribute to gene regulation. Finally, we combine genotype data with measurements of allele-specific expression to estimate that the effects of rare variants on gene expression are approximately 5X greater than those of common variants. We conclude that constraint on gene expression is present, but relatively weak at most genes, thus permitting high levels of gene-regulatory genetic variation.

# THE ROLE OF T CELL STIMULATION INTENSITY IN THE EXPRESSION OF IMMUNE DISEASE GENES

Dafni A Glinos[1], Blagoje Soskic[1], David M Sansom[2], Gosia Trynka[1]

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cellular Genetics, Cambridge, United Kingdom, [2]Institute of Immunity and Transplantation, University College London, Royal Free Campus, London, United Kingdom

Genome wide association studies for common immune diseases identified genetic variants in proximity to genes involved in T cell activation. T cells need two types of signals to undergo activation, from the T cell receptor (TCR) and the CD28 costimulatory molecule. We saugh to investigate the role of costimulation through CD28 in activation of CD4 naive and memory T cells. We stimulated CD4+ T cells from eight healthy donors with varying intensities of TCR and CD28 generating a total of seven conditions and performed RNA-seq, ATAC-seq and ChIPmentation assay for H3K27ac to profile gene expression regulation on sorted activated naïve and memory cells.

We build a linear model of gene expression along the increase of stimulus intensity and identified significantly different profiles of cell type specific response: 1,124 and 311 genes were TCR-sensitive in naive and in memory cells respectively; and 434 and 368 genes were CD28-sensitive in naive and in memory cells respectively. We observed overrepresentation of CD28-sensitive genes in the immune cell pathways, including the pro-inflammatory cytokines, such as IL6, IL13 and IL17F. The DNA replication pathway showed strong cell-type and stimulus-type specificity, CD28-sensitive in memory T cells while TCR-sensitive in naïve cells. These suggests that the main effector functions of memory cells, proliferation and cytokine production, are controlled by CD28 costimulation.

For 38% of the differentially expressed genes that changed specifically upon strong TCR we were able to nominate at least one candidate genomic region with ATAC or H3K27ac peak that could account for the gene regulation. These regions were enriched for Blimp-1 motifs in naive cells and IRF4-JUN-BATF motifs in memory cells.

Finally, we found that type-1 diabetes (T1D), rheumatoid arthritis (RA) and inflammatory bowel disease were all enriched for genes differentially expressed upon strong TCR by itself. We identified five genes that had a stimulus specific peak that overlapped with an immune SNP for the same disease, including IRF4 in RA and CTLA4 in T1D and RA.

This study represents the first effort to dissect the interplay between TCR and CD28 in T cell stimulation, and their role in the expression of genes that are affected by immune disease variants.

# EXPRESSION PATTERNS OF Y-CHROMOSOME GENES ACROSS HUMAN TISSUES AND INDIVIDUALS.

Alexander K Godfrey[1,2], David C Page[1,2,3]

[1]Whitehead Institute, Cambridge, MA, [2]Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, [3]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA

The human Y chromosome encodes 27 genes and multi-copy gene families in its male specific region (MSY). To gain deeper insight into their biological contributions throughout the body, we surveyed their quantitative expression profiles across 36 human tissues and 156 post-mortem donors using RNA-sequencing data from GTEx Consortium. Nearly all MSY genes show one of two characteristic expression profiles. Genes of one set are expressed predominantly in the spermatogenic cells of the testis. The 10 MSY genes of the other set are robustly expressed in all tissues examined. These more widely expressed genes are the differentiated, Y-specific homologs of regulatory genes on the X chromosome. Although these Y-linked homologs are, on average, expressed at 30% of levels of their X-linked counterparts, we identify genes and tissues where the Y homolog is more abundantly expressed, contributing up to 84% of the gene product in males. Finally, we show that the expression levels of MSY genes are largely conserved across human males whose Y chromosomes are typical of distinct populations. Our findings provide a foundation for exploring the broader contributions of Y-chromosome genes to human physiology and differences between the sexes.

# STUDY OF THE MITOTIC CHROMATIN SHOWS INVOLVEMENT OF HISTONE MODIFICATIONS IN BOOKMARKING AND REVEALS NUCLEOSOME DEPOSITION PATTERNS

Elisheva Javasky[1], Inbal Shamir[1], Shashi Gandhi[1], Shawn Egri[2], Oded Sandler[1], Noam Kaplan[3], Jacob D Jaffe[2], Alon Goren[4], Itamar Simon[1]

[1]The Hebrew University, Department of Microbiology and Molecular Genetics, Jerusalem, Israel, [2]The Broad Institute, Proteomics, Cambridge, MA, [3]Technion, Department of Physiology, Biophysics & Systems Biology, Haifa, Israel, [4]University of California San Diego, Department of Medicine, San Diego, CA

Mitosis encompasses key molecular changes including chromatin condensation, nuclear envelope breakdown and reduced transcription levels. Immediately after mitosis, the interphase chromatin structure is reestablished and transcription resumes. The reestablishment of the interphase chromatin requires 'bookmarking', i.e., the retention of at least partial information during mitosis. Yet, while recent studies demonstrate that chromatin accessibility is generally preserved during mitosis and is only locally modulated, the exact details of the bookmarking process and its components are still unclear. To gain a deeper understanding of the mitotic bookmarking process, we merged proteomics, immunofluorescence, and ChIP-seq approaches to study the mitotic and interphase organization in human cells. We focused on key histone modifications, and employed the HeLa-S3 cells as a model system. Generally, we observed a global concordance between the genomic organization of histone modifications in interphase and mitosis, yet the abundance of the two types of modifications we investigated was different. Whereas histone methylation patterns remain highly similar, histone acetylation patterns show a general reduction while maintaining their genomic organization. In line with a recent study demonstrating that minimal transcription is retained during mitosis, we show that RNA polymerase II does not fully disassociate from the genome, but rather maintains its genomic localization at reduced levels. Next, we followed up on previous studies demonstrating that nucleosome depleted regions (NDRs) become occupied by a nucleosome during mitosis. Surprisingly, we observed that the nucleosome introduced into the NDR during mitosis encompasses a distinctive set of histone modifications, differentiating it from the surrounding nucleosomes. We show that the nucleosomes in the vicinity of the NDR appear to both shift into the NDR during mitosis and undergo deacetylation. HDAC inhibition by the small molecule TSA reverts the deacetylation pattern of the shifted nucleosome. Taken together, our results demonstrate that the epigenomic landscape can serve as a major component of the mitotic bookmarking process, and provide evidence for a mitotic deposition and deacetylation of the nucleosomes surrounding the NDR.

# THE GENETIC BASIS OF MUTATION RATE VARIATION IN YEAST

Liangke Gou[1], Joshua S Bloom[1,3], Leonid Kruglyak[1,2,3]

[1]University of California, Los Angeles, Department of Human Genetics, Los Angeles, CA, [2]University of California, Los Angeles, Department of Biological Chemistry, Los Angeles, CA, [3]Howard Hughes Medical Institute, Los Angeles, CA

Mutations are the root source of genetic variation and underlie the process of evolution. Although the rate at which mutations occurs varies considerably between species, very little is know about the genetic and molecular basis of these differences, or differences within species. Here we leveraged the power of the yeast *Saccharomyces cerevisiae* as a model system to uncover natural genetic variants that underlie variation in mutation rate. We observed that mutation rate varies among natural yeast strains and is highly heritable ($H^2$=0.46). We developed and implemented a high-throughput fluctuation assay and used it to quantify mutation rates in natural yeast isolates and in 1008 segregant progeny from a cross between BY, a lab strain, and RM, a wine strain. We performed linkage mapping, and identified four quantitative trait loci (QTLs) underlying mutation rate variation between the two parental strains. We identified two causal genes, *RAD5* and *MKT1*, that contribute to mutation rate variation. These genes also underlie sensitivity to the mutagens 4NQO and MMS, suggesting a connection between spontaneous mutation rate and mutagen sensitivity.

# IMPLICATIONS OF POST-COLONIAL DEMOGRAPHIC STRUCTURE ON ASSOCIATION ANALYSES AND THEIR INTERPRETATION

Julie M Granka[1], Eurie L Hong[1], Kristin A Rand[1], Shiya Song[1], Daniel Garrigan[1], Jake K Byrnes[1], Catherine A Ball[1], Kenneth G Chahine[2]

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

Recent work has highlighted the need to evaluate the impact of fine-scale population structure on association analyses. Despite thousands of high-powered genome-wide association studies (GWAS), only about one-fifth of all studies include individuals of non-European descent. It is understood that GWAS conducted in European populations are not always generalizable to non-European populations, and that continental population structure is an important factor to consider. However, in the age of regionally-run genomic studies, it remains unclear whether fine- scale population structure, particularly more recent within-continental structure, contributes to bias in the interpretation of GWAS results.

We identify regional fine-scale structure, within continental ethnicity groups, in a subset of over one million AncestryDNA customers who have consented to research. In most cases, such regional United States population structure, identified to relate to post-colonial demography during the last several hundred years, is undetectable in a principal component analysis (PCA). However, we still observe regional allele frequency variation at the single-variant level.

To explore whether these regional allele frequency differences play a role in association analyses, we examine the frequency distribution and distribution of polygenic risk scores using previously-identified variants in the GWAS catalog. To account for the effects of continental population structure, we examine regional sub-population frequencies of variant subsets that have been identified in each specific continental population (i.e., SNPs identified in a GWAS of European ancestry for European-descent US subpopulations). While observed differences occur across only some regional groups, in some cases these differences could suggest residual confounding due to within-continental population structure.

It is unclear whether the observed differences can be attributed to true differences in disease risk, or whether they are solely an artifact of generalizing GWAS results to other, unstudied populations. Our results show that, even when considering very recent time scales, small frequency differences can impact GWAS conclusions and interpretations -- particularly in the transferability of results to similar, but unstudied, populations. This work highlights the fundamental need to understand the generalizability of GWAS results through the lens of both continental, and within-continental, population structure, particularly in large-scale analyses.

# HOMININ SELECTIVE PRESSURE DRIVES CELL TYPE SENSITIVE EPIGENOMIC SEGMENTATIONS AT BASE PAIR RESOLUTION.

Brad Gulko, Adam Siepel

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

Recent works like LINSIGHT and CADD leverage epigenomic properties to help quantify hominin selective constraint, identifying genomic loci important to human phenotypes. However, such methods are often linearly constrained and insensitive to cell type, obscuring epigenomic relationships important to tissue-specific biology. The relationship between procreative fitness, cellular epigenomes, and primary sequencer variation remains poorly explored.

Here we integrate selective pressure and diverse epigenomic properties to develop an interpretable, non-linear, generative model that infers 61 distinct patterns of epigenomic properties at base-pair resolution. Patterns are generated via a straightforward decision tree process maximizes information about weak, strong, and adaptive hominin selective pressures on the primary sequence positions exhibiting those patterns. Pattern driven genome segmentations associate each hg19 position in a human epigenome with one of 61 combinations of simple features such as: TFBS, RNA splicing, CDS, small RNA-seq, and chromatin state. Genomic positions evincing a pattern are collected across karyotype normal Roadmap cell types to calculate an impact score for that pattern, directly interpretable as probability of selective constraint under the INSIGHT model. The result is a scoring and segmentation of each Roadmap cell type describing both human functional genomics and hominin primary sequence constraint with base-pair resolution.

We leverage the generative properties of this model to separate primary sequence features like splice sites and TFB motifs, from tissue specific activity characterized by dynamic chromatin properties. This serves to identify shared regulatory structures, while simultaneously quantifying level of cell type specific activity. Using similar Roadmap cell types from different developmental stages, we analyze epigenomic trajectories characterizing important genomic properties of:
• Craniofacial enhancers,
• Autism associated genomic variation,
• Cancer-specific gene regulation.
In addition, we quantify the information theoretic contribution of each functional assay as a guide for biologists investigating epigenetic properties of novel cell types with limited experimental resources. Our scoring also identifies pathogenic regulatory loci with accuracy approaching state of the art classifiers. The epigenomic classes identified in this work provide a regulatory lexicon as a starting point for developmental regulatory network analysis and exploration of the regulatory basis for human disease.

Segmentations, covariates, and scoring are available via a UCSC genome browser track hub at http://compgen.cshl.edu/fitCons2 .

# ELUCIDATING THE GENETIC DEFECTS UNDERLYING RADIATION HYPERSENSITIVE PHENOTYPES THROUGH WHOLE GENOME AND EXOME SEQUENCING

Meenal Gupta[1], Xiangfei Liu[2], Sharon Teraoka[2], Patrick Concannon[2], Aaron Quinlan[1]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]University of Florida, Department of Pathology, Immunology and Laboratory Medicine, Gainesville, FL

Ionizing radiation is an effective therapeutic agent for cancer treatment as well as a potent carcinogen. Adverse reactions in patients owing to the underlying DNA repair defects are not uncommon. The present study aims to elucidate the genetic defects in such radiation sensitive individuals, which can shed light on novel proteins/pathways involved in DNA damage responses to radiation. Previous studies from our group have been successful in identifying and validating the role of new genes such as *MTPAP* and *ATIC* in radiation sensitivity (RS). Here, we performed whole genome and exome sequencing in a cohort of 80 radiation sensitive individuals to identify potential rare/novel loss of function (LoF) single nucleotide variants (SNVs) as well as copy number variants (CNVs). SNVs were annotated with functional consequence predictions using GEMINI and prioritized under the assumption of a recessive model. We did not detect any homozygous or compound heterozygous mutations in the known RS genes (*ATM* and *NBN*) in this cohort. Rather, we were able to identify three novel frameshift, potential LoF variants in three genes (*LIMCH1*, *CTRB1* and *SLC2A5*) not known to be involved in RS or DNA damage pathways. We also identified two novel, possibly deleterious missense mutations in DNA damage repair pathway genes (*KLF4*, *MCPH1*). In addition, we have also adopted an epistatic model to identify rare/novel heterozygous mutations in gene-pairs which might disrupt essential protein-protein interactions (PPIs) among RS genes. This approach led to identification of potentially deleterious novel and rare mutations in *ATM-ATR*, *TP53-TP53BP1* and *BRCA2-BARD1* gene pairs. Apart from SNVs, we have also detected large homozygous and heterozygous CNVs on chromosomes 2,14 and 16 in four individuals. These include deletions and duplications affecting atleast 136 genes. The putative novel radiation sensitivity genes identified are being functionally validated by a two step process. Initially the genes are knocked down using shRNA in a cell line of normal radiosensitivity followed by testing for increased susceptibility to killing by ionizing radiation. Genes passing this screen are then further tested by knockdown followed by colony survival assays. In conclusion, our robust methodology to detect and validate RS gene candidates can provide new insights into the mechanisms underlying radiation hypersensitivity.

# A REFERENCE HAPLOTYPE PANEL FOR GENOME-WIDE IMPUTATION OF SHORT TANDEM REPEATS

Shubham Saini[1,2], Ileena Mitra[2,3], Melissa Gymrek[1,2]

[1]University of California San Diego, Department of Computer Science and Engineering, La Jolla, CA, [2]University of California San Diego, Department of Medicine, La Jolla, CA, [3]University of California San Diego, Bioinformatics and Systems Biology Program, La Jolla, CA

Short tandem repeats (STRs) are involved in dozens of Mendelian psychiatric and neurological disorders and have been implicated in a variety of complex traits in humans. However, existing technologies have not allowed for systematic STR association studies. Genotype array data is available for hundreds of thousands of samples, but is limited to variation in common single nucleotide polymorphisms (SNPs) and does not adequately capture more complex variants like STRs. Here, we leverage next-generation sequencing from 479 families along with existing bioinformatics tools to phase STRs onto SNP haplotypes and create a genome-wide reference haplotype panel. Imputation using our panel achieved an average of 98% concordance between true and imputed STR genotypes in an external dataset. We tested our panel on STRs with known associations with complex traits or Mendelian disorders, and found that our panel accurately recovered known association signals better than the best tag SNPs. Imputed STRs capture on average 20% more variation in STR allele length compared to individual common SNPs, highlighting a limitation of standard genome-wide association studies. Our framework will enable testing for STR associations with hundreds of traits across massive sample sizes without the need to generate additional data.

# EXPLORING THE FUNCTION OF NON-CODING VIRAL TRANSCRIPTS IN BOVINE LEUKEMIA VIRUS INDUCED LEUKEMIA

Vincent Hahaut[1], Maria Artesi*[1], Keith Durkin[1], Natasa Arsic[2], Philip Griebel[2], Michel Georges[1], Anne Van den Broeke[1,3]

[1]University of Liège, Unit of Animal Genomics, Liège, Belgium, [2]University of Saskatchewan, VIDO-InterVac, Saskatoon, Canada, [3]Institut Jules Bordet, Laboratory of Experimental Hematology, Brussels, Belgium

The deltatretroviruses Human T-cell leukemia virus and its close relative Bovine Leukemia Virus (BLV) produce a chronic infection in their respective hosts that evolves into full-blown leukemia/lymphoma in ~5% of individuals following several years of latency. The early and chronic stages of infection are characterized by the presence of multiple clones of varying abundance, each uniquely identified by their proviral integration site in the genome, making it possible to track them over time. After a protracted latency period, for unknown reasons one of these clones rapidly expands, leading to the abrupt accumulation of malignant cells. While not a natural host, it is possible to experimentally infect sheep with BLV and in contrast to the situation in cattle and human, all infected sheep generally develop B-cell leukemia/lymphoma after a much shorter latency.

Although BLV encodes proteins that participate to the leukemogenic process, BLV induced tumors typically lack detectable viral mRNA produced from the viral positive strand. Using high-throughput sequencing based methods, we recently identified two classes of viral non-coding transcripts that are expressed at all stages of the disease: (i) a cluster of ten abundant RNA POL III dependent microRNAs transcribed from the positive strand, and (ii) viral antisense transcripts (AS-1 and -2) originating in the 3' Long Terminal Repeat of the provirus. We also demonstrated that BLV produces antisense RNA dependent chimeric transcripts that affect cancer driver genes located in the vicinity of the provirus.

To explore the function of BLV non-coding transcripts we have infected sheep with wild-type (WT) and mutant BLV proviruses. We have been following these animals for > 4 years with regular samplings. This resulted in a large collection of samples from before infection to the aggressive leukemia. We examined the genome-wide distribution of proviral integration sites and measured the abundance of the corresponding clones in WT and mutant BLV infected animals over time. This was achieved by applying an improved HTS-based clonality method to longitudinal samples covering each stage of the disease. Doing so, we were able to explore the presence of hotspots of proviral integration and track the expansion and persistence of infected clones over time. Altogether, this work revealed novel insights into the clonal architecture and its evolution during progression of deltaretrovirus induced leukemia.

# AN ALIGNMENT AND REFERENCE FREE APPROACH TO DECONVOLVE LINKED-READS FOR METAGENOMICS

David C Danko, Dmitrrii Meleshko, Daniela Bezdan, Chris Mason, Iman Hajirasouliha

Weill Cornell Medicine of Cornell University, Institute for Computational Biomedicine, New York, NY

Emerging linked-read technologies (aka read-cloud or barcoded short-reads) have revived interest in standard short-read technology as a viable way to understand large-scale structure in genomes and metagenomes. Linked-read technologies, such as the 10X Chromium system, use a microfluidic system and a set of specially designed 3' barcodes (aka UIDs) to tag short DNA reads which were originally sourced from the same long fragment of DNA; subsequently, these specially barcoded reads are sequenced on standard short read platforms. This approach results in interesting compromises. Each long fragment of DNA is covered only sparsely by short reads, no information about the relative ordering of reads from the same fragment is preserved, and typically each 3' barcode matches reads from 2-20 long fragments of DNA. However, the cost per base to sequence is far lower than single-molecule long read sequencing systems, far less input DNA is required, and the error rate is that of standard short-reads.

Linked-reads represent a new set of algorithmic challenges. In this talk, we formally describe one particular issue common to all applications of linked-read technology: the deconvolution of reads with a single barcode into clusters that correspond to a single long fragment of DNA when a reference genome is not available. We introduce Minerva, A graph-based algorithm which approximately solves the barcode deconvolution problem for metagenomic data (where reference genomes may be incomplete or unavailable). Additionally, based on evaluations on two primary real data sets from mock communities, we demonstrate that deconvolved barcoded reads significantly improve downstream results by improving the specificity of taxonomic assignments, and by improving the ability of topic models to identify clusters of related sequences. In particular, using our technique, we were able to rescue a large number of reads from unspecific taxonomic assignments. To the best of our knowledge, we are the first to describe the problem of barcode deconvolution in metagenomics. Our tool is open source and freely available at:
https://github.com/dcdanko/minerva_barcode_deconvolution

# MECHANISTIC INSIGHT INTO THE BIOACTIVITY OF A NOVEL BACTERIAL PROTEIN FOR THE TREATMENT OF INFLAMMATORY BOWEL DISEASE

Roberta Hannibal[1], Cheryl-Emiliane Chow[2], Andrew Goodyear[3], Yingwu Li[3], Tarunmeet Gujral[3], Shoko Iwai[2], Andrew Han[1], Laurens Kruidenier[3], Todd DeSantis[2], Karim Dabbagh[1,2,3]

[1]Second Genome, Inc, Microbiology, South San Francisco, CA, [2]Second Genome, Inc, Informatics, South San Francisco, CA, [3]Second Genome, Inc, Discovery, South San Francisco, CA

Changes in gut microbiota composition and activity are associated with a wide variety of disorders, including inflammatory bowel disease (IBD). However, there is limited understanding of how bacteria-derived molecules regulate host physiology. To develop novel therapies, Second Genome has implemented a unique microbiome discovery platform that associates human clinical phenotypes with key bacterial molecules. By comparing the microbiome of colonic mucosal biopsies from healthy subjects and individuals with ulcerative colitis, we identified bacterial secreted proteins associated with health and disease. Proteins were heterologously expressed, purified, and screened in a panel of *in vitro* assays. We then tested the top *in vitro* hits in a mouse model of IBD. One protein, SG-P413, improved gut barrier function and clinical endpoints *in vivo*. To connect this activity to a molecular pathway, we performed RNA sequencing. We identified a number of affected host pathways and are currently determining whether the effect of SG-P413 is direct or indirect. Understanding how microbial proteins impact specific host cell types and their underlying molecular pathways is the most direct route towards developing a novel class of therapeutics.

# SYSTEMATIC INTEGRATION OF EPIGENOMES VIA IDEAS PAINTS THE REGULATORY LANDSCAPE OF HEMATOPOIESIS

Ross C Hardison[1], Cheryl A Keller[1], Guanjue Xiang[1], Lin An[1], Elisabeth Heuston[2], Jens Lichtenberg[2], Belinda M Giardine[1], David Bodine[2], Yu Zhang[1]

[1]Penn State University, BMB and Statistics, University Park, PA, [2]National Institues of Health, NHGRI, Bethesda, MD

We integrated genome-wide epigenetic and transcript data for twenty cell types covering hematopoietic stem cells, multilineage progenitors, and mature cells across the blood cell lineages of mouse. The epigenomic data included chromatin accessibility (ATAC-seq and DNase-seq), CTCF occupancy, and histone H3 modifications associated with transcriptional activity (K4me1, K4me3, K27ac, K36me3) or repression (K27me3, K9me3). The epigenomic data were integrated using the Integrative and Discriminative Epigenome Annotation System (IDEAS, PMIDs: 27095202 and 28973456), which learns all common combinations of features (chromatin states) simultaneously in two dimensions - along chromosomes and across cell types. This system preserves position-specific information to provide more consistent views across cell types and better discrimination of differences in the regulatory landscape. The result is a segmentation that effectively paints the regulatory landscape in readily interpretable views, revealing constitutively active or silent loci as well as the loci specifically induced or repressed in each stage and lineage. Nuclease accessible DNA segments in active chromatin states were designated candidate *cis*-regulatory elements (ccREs) in each cell type, providing one of the most comprehensive catalogs of hematopoietic ccREs to date. Studying ccREs within this context of systematic integration across a well-known differentiation process allows us to follow their history across cell types (birth, persistence, loss), their shifts between chromatin states during differentiation, and their occupancy by different transcription factors. We leveraged both the associated gene expression data and whole-genome chromatin interaction frequency data for selected cell types to assign the ccREs to candidate target genes. These maps of ccREs within the cell-specific regulatory landscape are being used to build models that should reveal consistent rules for hematopoietic gene regulation. The data have a variety of sources, generated within our own labs as well as acquired from public resources, and we developed novel methods for more effective normalization across heterogeneous datasets. This work is part of a project called VISION, an international, multi-lab project that aims to provide a ValIdated Systematic IntegratiON of epigenomic data in mouse and human hematopoiesis. The data, integrated segmentation results, ccRE catalog, chromatin interaction maps, and other resources are available from our website **http://usevision.org**.

# TRANSCRIPTIONAL COMPLEXITY OF NON-CODING GENOMIC REGIONS ASSOCIATED WITH BRAIN PHENOTYPES

Simon A Hardwick[1,2], Sam Bassett[2], Martin A Smith[1,2], Nenad Bartonicek[1,2], Dominik Kaczorowski[1], Tim R Mercer[1,2,3], John S Mattick[1,2]

[1]Garvan Institute of Medical Research, Genomics & Epigenetics, Sydney, Australia, [2]University of NSW, Faculty of Medicine, Sydney, Australia, [3]Altius Institute for Biomedical Sciences, Genome Science, Seattle, WA

Genome-wide association studies (GWAS) have uncovered many genetic variants associated with brain phenotypes, including neurological and behavioural traits. However, unexpectedly, the majority of GWAS single nucleotide polymorphisms (SNPs) identified to date are located in intronic and intergenic regions of the genome, confounding their functional evaluation. Long non-coding RNAs (lncRNAs) exhibit a staggering diversity of transcript isoforms in human brain, but often at low expression levels that are below the limit of detection of traditional RNA sequencing (RNA-seq). Accordingly, we designed a set of probes to capture transcripts from intronic and intergenic brain phenotype-associated haplotype blocks identified by GWAS, and performed targeted RNA-seq on samples taken from eight different brain regions from three neurologically healthy donors. We developed a robust analytical pipeline for transcript assembly, annotation and quantification, finding multi-exonic transcripts for 824/1,023 (~81%) haplotype blocks targeted. The vast majority (~74%) of transcripts assembled are novel, suggesting that existing databases are incomplete. Many of these novel transcripts are brain region-specific, evolutionarily conserved, and enriched for epigenetic signatures of active transcription and enhancer activity. We present these transcriptomes as an atlas that can be used to connect gene expression with neurological and behavioural traits. Additionally, a selection of differentially expressed transcripts are examined in detail as examples of mining the atlas and potential candidates for further study.

# PRIMATE INTRA- AND INTER-SPECIES CHROMOSOME 19 VARIATION IN THE CONTEXT OF THE REGULATORY METHYLOME

R. Alan Harris[1,2], Muthuswamy Raveendran[1,2], Kim C Worley[1,2], Jeffrey Rogers[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Baylor College of Medicine, Molecular and Human Genetics, Houstonj, TX

Human chromosome 19 has many unique characteristics including gene density more than two fold higher than the genome-wide average. It contains 20 large tandemly clustered gene families and has the highest GC content of any chromosome, especially in regions outside of those gene clusters. The high GC content and concomitant high number of CpG dinucleotides raises the possibility chromosome 19 exhibits higher levels of nucleotide diversity both within and between species, and may possess greater variation in DNA methylation that regulates gene expression.

We examined GC and CpG content of chromosome 19 orthologs across representatives of the primate order. In all 12 primate species with suitable genome assemblies, chromosome 19 orthologs have the highest GC content of any chromosome. CpGs and CpG islands are also more prevalent in chromosome 19 orthologs than other chromosomes. GC and CpG content are generally higher outside of the gene clusters. Intra-species variation measured in the number of SNPs is most prevalent on chromosome 19 and its orthologs in human common dbSNP, rhesus, crab eating macaque, baboon and marmoset datasets. Inter-species comparisons based on phyloP conservation show accelerated nucleotide evolution for chromosome 19 promoter flanking and enhancer regions. These same regulatory regions show the highest CpG density of any chromosome suggesting they possess considerable methylome regulatory potential. Previously published primate sperm methylation data show chromosome 19 promoter flanking regions to have the lowest methylation levels of any autosome, but chromosome 19 promoter and enhancer regions show average methylation.

The pattern of high GC and CpG content of chromosome 19 orthologs, particularly outside gene clusters, is present from human to mouse lemur representing 73 million years of primate evolution. Much CpG variation exists both within and between primate species with a portion of this variation occurring in regulatory regions. What maintains this unusually high CpG content across primates despite the tendency of methylated cytosines to mutate to thymines remains an open question.

# A MAP OF CONSTRAINED CODING REGIONS IN THE HUMAN GENOME.

James M Havrilla[1,2], Brent S Pedersen[1,2], Ryan M Layer[1,2], Aaron R Quinlan[1,2,3]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]USTAR Center for Genetic Discovery, Human Genetics, Salt Lake City, UT, [3]University of Utah, Biomedical Informatics, Salt Lake City, UT

Interspecies sequence conservation summarizes the degree of genetic constraint over vast evolutionary periods. In contrast, catalogs of genetic variation from thousands of exomes and genomes enable the inference of more recent constraint from extreme paucities of genetic variation. While existing techniques such as pLI and RVIS summarize constraint for entire genes, it is clear that singular gene-wide metrics do not capture the variability in constraint that exists within each protein coding gene. To address this limitation, we have charted a detailed map of constrained coding regions (CCRs) within the human genome by leveraging coding variation observed among 123,136 humans from the Genome Aggregation Database (gnomAD). Constrained coding regions arise when the observed distance between missense variants in gnomAD — a proxy for constraint — is much greater than expected.

We demonstrate that, as expected, our most constrained coding regions are significantly enriched (154-fold) for ClinVar pathogenic variants over benigns in CCRs greater than the 95th percentile. Moreover, pathogenic de novo variants, identified in individuals with developmental delay, severe intellectual disability, and epileptic encephalopathy, are substantially (6.35-fold) enriched in CCRs in the 95th percentile when compared to benign de novo variants from unaffected siblings of autism patients. The most constrained CCRs are found in many genes known to be associated with severe autosomal dominant phenotypes and important cellular function. CCRs also reveal protein domain families under extreme constraint, suggest unannotated functional domains from the degree of constraint, and perform as well as or better than existing tools for prediction of ClinVar pathogenicity, and better than existing tools in the prioritization of de novo variation in studies of disease. We also explore whether highly constrained regions overlap residues and domains demonstrated to affect interaction with other proteins.

Finally, we hypothesize that CCRs under the highest constraint hold the promise of revealing coding regions under extreme purifying selection and genes with yet unobserved human phenotypes owing to embryonic lethality. Therefore, our map of constrained coding regions provides a course from which to identify new genes underlying human disease phenotypes and regions crucial for protein interaction.

# dbVar: TOWARD A HUMAN STRUCTURAL VARIATION REFERENCE SET (ALPHA RELEASE)

Tim Hefferon, John Lopez, John Garner, Lon Phan

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

Advances in genomic technologies have revealed structural variations (SV) to be prevalent in all human DNA, and research increasingly implicates SV in phenotypic diversity and disease. Within the next few years millions of genomes will be sequenced, leading to the discovery of millions of SVs that will need to be analyzed to understand their functional impacts. A critical step in variant analysis and interpretation is comparing SV identified in an individual or population with sets of known variants in public databases such as dbVar, to gain biological insights from existing variant annotation and to identify novel variants. This process can be especially difficult without a reference set of variants, representing the sum of current knowledge, to provide a context in which novel variation can be understood. To address this gap, dbVar proposed the creation of a structural variation reference catalog with rich genomic and biological annotations for use in SV analysis, annotation, and related workflows.

dbVar is an NCBI database of human genomic structural variation that contains more than 5 million submitted SVs from 157 human studies, including data from large diversity projects such as the 1000 Genomes Project and the global population CNV survey (Sudmant et al., 2015), and from clinical sources such as ClinVar and ClinGen. Using this large collection of variants, we created two SV datasets. One is an aggregated list of variants in dbVar (i.e. CNV, indel, etc.) derived from large-scale published studies that were conducted by consortia or large collaborations, included high numbers of tested samples, were genome-wide in scope, and applied multiple rigorous methods and analyses.

The second reference dataset is a list of genomic intervals named Structural Variant Clusters (SVC) containing high concentrations of putative ‘common’ structural variation across the genome. SVC are generated using open source software (https://github.com/NCBI-Hackathons/Structural_Variant_Comparison, Phan et al., 2017). SVs from both reference datasets were annotated using available data on genes, molecular consequence, clinical significance, dosage sensitivity, and regulatory regions, as well as relevant genomic structural features such as repetitive regions, segmental duplications, and assembly-assembly alignment anomalies. The reference data and annotations are intended to facilitate the integration and comparison of dbVar SV data with other genome annotations (such as disease phenotype and population frequencies) and to provide insights into the impact of the SV on biological functions.

The resulting alpha-release datasets are available in VCF and tab-delimited formats. Our goal is to update these files on a regular basis, incorporating new variant submissions, genomic features, Gene, RefSeq, ClinVar and other annotation information as is becomes available. We encourage users to test these files and to provide feedback to dbVar at dbvar@ncbi.nlm.nih.gov.

**Acknowledgments**
Work at NCBI is supported by the NIH Intramural Research Program and the National Library of Medicine.

# INFERENCE OF SELECTIVE SWEEPS BASED ON THE ANCESTRAL RECOMBINATION GRAPH IN RAPID AVIAN RADIATION

Hussein A Hejase [1], Leonardo Campagna [2,3], Ilan Gronau [4], Melissa Hubisz[5], Irby J Lovette [2,3], Adam Siepel [1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cornell University, Department of Ecology and Evolutionary Biology, Ithaca, NY, [3]Cornell Laboratory of Ornithology, Fuller Evolutionary Biology Program, Ithaca, NY, [4]Herzliya Interdisciplinary Center, Efi Arazi School of Computer Science, Herzliya, Israel, [5]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY

A central problem in evolutionary biology is to understand the genetic basis behind divergent traits. Genomic scale data is key to identifying differentiated loci among recently diverged taxa, where regions in the genome that show high differentiation among species could contain key loci that shape an adaptive radiation. In this work, we analyze a collection of rapid radiation events in eight sympatric bird species known as southern capuchino seedeaters in the genus Sporophila. These southern capuchinos have high levels of phenotypic diversity in terms of male plumage and song, yet show extremely low levels of genetic differentiation, due to both genetic admixture and incomplete lineage sorting. This genetic homogeneity led us to hypothesize that the observed phenotypic diversity may be the result of strong selection acting at a few key loci. Previous studies have attempted to analyze these genomes by using simple summary statistics (e.g. $F_{ST}$) to identify highly differentiated loci that are under selection. Here, we perform direct evaluation of hard and soft selective sweeps in southern capuchinos using a machine learning classifier that utilizes an ensemble of decision trees (i.e. random forests). Our method, ARG-Sweep, is similar to methods such as S/HIC, SFselect, and evolBoosting, in that it aggregates summary statistics as informative sequence features for prediction (e.g. number of segregating sites, Tajima's D statistic, site frequency spectrum, length distribution between segregating sites, identity-by-state tract length distribution, and linkage disequilibrium distribution). However, ARG-Sweep differs from these recent methods by utilizing distinguishing local topological features of the ancestral recombination graph (ARG) (e.g. outlier clusters of coalescent events and coalescent time to most recent common ancestry) as inferred by ARGweaver. ARG-Sweep then learns patterns of variation in the feature set to discriminate between hard selective sweeps, soft selective sweeps, and neutral regions. Using simulations, we demonstrate that ARG-Sweep can accurately predict the location of hard and soft selective sweeps with an accuracy as high as 85% based on demographic scenarios of the southern capuchino clade. Our analysis provides new insights into how selection shaped regions of the southern capuchino seedeaters genomes, and provides better understanding on the driving force behind these high levels of phenotypic diversity observed among these different bird species.

# ENRICHED LOSS-OF-FUNCTION VARIANTS IN *ANGPTL4* AND *ANGPTL8* ASSOCIATE WITH LOWER RISKS OF CARDIOMETABOLIC DISEASES

Pyry Helkkula[1], Ida Surakka[2,1], Mervi Alanne-Kinnunen[1], Aki Havulinna[1,3], Tuomo Kiiskinen[1], Olli Raitakari[4], Terho Lehtimäki[5], Johan Eriksson[6], Minna Männikkö[7], Seppo Koskinen[3], Veikko Salomaa[3], Hannele Laivuori[1], Elisabeth Widen[1], Mark J Daly[8], Aarno Palotie[1,8], Samuli Ripatti[1]

[1]Institute for Molecular Medicine Finland, Helsinki, Finland, [2]University of Michigan, Ann Arbor, MI, [3]National Institute of Health and Welfare, Helsinki, Finland, [4]University of Turku, Turku, Finland, [5]University of Tampere, Tampere, Finland, [6]University of Helsinki, Helsinki, Finland, [7]University of Oulu, Oulu, Finland, [8]Massachusetts General Hospital, Boston, MA

Population isolates, such as Finland, show an enrichment of loss-of-function (LoF) variants with potential impact on phenotypes and disease risk. We tested if some of the LoF variants were associated with blood lipids in Finns. Our study consisted of 3,412 autosomal LoF variants with MAF>1.0e-03 in 2,860 genes. We tested their association with LDL-C, HDL-C and triglycerides in the population-based FINRISK cohort (N=23,960) with genome-wide genotyping and imputation using a Finnish imputation reference panel.

We found high-effect lipid-modulating LoF variants in four genes without previously reported studies of human knockouts. Protein-disrupting variants in *ANGPTL4* (p.Cys80ValfsTer12, MAF=4.8e-03, effect=-0.29 SD, *P*=2.7e-06) and *ANGPTL8* (p.Gln131Ter, MAF=1.9e-03, effect=-0.61 SD, *P*=2.9e-08) were associated with lower triglyceride levels. Angptl4 and Angptl8 are *LPL*-inhibiting proteins, that play an essential role in triglyceride trafficking during fasting and feeding respectively. The association for *ANGPTL4* and *ANGPTL8* was stronger in individuals with longer and shorter fasting times respectively, which is in line with murine models. We also observed LoF variants with substantial HDL-increasing effects in *LIPG* (c.98-1G>A, MAF=1.7e-03, effect=0.63 SD, *P*=5.6e-11) and *CETP* (c.118+4_118+7delAGTA, MAF=1.5e-03, effect=1.1 SD, *P*=1.6e-20).

Pooling individuals from five Finnish population-based cohorts (N=38,618) we observed a protective effect against coronary artery disease of the *ANGPTL4* (OR=0.57, *P*=8.2e-03) and *ANGPTL8* (OR=0.24, *P*=6.1e-03) LoF alleles. The *ANGPTL8* truncating variant was also associated with lower risk of diabetes (OR=0.47, *P*=0.042). In a phenome-wide scan across EHR data, the *ANGPTL4* LoF allele associated with an increased risk of lymphatic disorders: Non-Hodgkin lymphoma (OR=3.04, *P*=0.018) and lymphoma (OR=2.77, *P*=0.028). The therapeutic safety of *ANGPTL4* inhibition has been questioned due to lymphatic lipid accumulation in animals.

A phenome-wide screen of the *ANGPTL8* LoF mutation showed increased odds of multiple metabolic and psychiatric endpoints, including thyroid gland disorders (OR=3.60, *P*=1.9e-04), inflammatory liver disease (OR=8.40, *P*=0.012) and major depression (OR=2.20, *P*=0.018). Our study shows the potential in combining genome data with rich nationwide health registry data in an isolated population such as Finland to learn about potential drug targets and their possible safety.

# A FIRST GENERATION ATLAS OF IN VIVO MAMMALIAN CHROMATIN ACCESSIBILITY AT SINGLE CELL RESOLUTION

Andrew J Hill[1], Darren A Cusanovich[1], Delasa Aghamirzaie[1], Riza M Daza[1], Hannah A Pliner[1], Joel B Berletch[2], Galina N Filippova[2], Lena Christiansen[3], William S DeWitt[1], Choli Lee[1], Samuel G Regalado[1], David F Read[1], Frank J Steemers[3], Christine M Disteche[2], Cole Trapnell[1], Jay Shendure[1,4]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]University of Washington, Department of Pathology, Seattle, WA, [3]Illumina, Advanced Research Group, San Diego, CA, [4]Howard Hughes Medical Institute, Seattle, WA

A common feature of regulatory DNA in eukaryotic genomes is increased chromatin accessibility. Although chromatin accessibility has been extensively surveyed in mammalian tissues and cell lines, these data are confounded by cell type heterogeneity and in vitro culturing, respectively. Here we applied a single cell combinatorial indexing assay for transposase-accessible chromatin with high throughput sequencing (sci-ATAC-seq) to diverse tissues of the house mouse, Mus musculus. We profiled genome-wide chromatin accessibility in ~100,000 single cells derived from 17 samples representing 13 mouse tissues (whole brain, forebrain, cerebellum, heart, thymus, large intestine, small intestine, testes, spleen, bone marrow, lung, kidney, liver) at 8 weeks of age. Across cells from all tissues, we identified 85 distinct patterns of chromatin accessibility. These data define, for the first time, the in vivo landscape of the regulatory genome for common mammalian cell types at single cell resolution. We show how these data are useful for cataloging ~400,000 differentially accessible candidate regulatory elements; for linking distal regulatory elements to their target genes; for defining the transcription factor grammar that specifies each cell type; and for identifying in vivo correlates of heterogeneity in chromatin accessibility within a cell type (e.g., tissue-of-origin, differentiation dynamics, spatial coordinates). Finally, by intersecting mouse chromatin accessibility with the results of human genome-wide association studies, we identify cell-type-specific enrichments of the heritability signal for hundreds of complex traits. In the coming years, single cell profiles of the in vivo chromatin landscape will be a key component of organism-scale cell atlases. As with the human genome, we anticipate that comparison to model organisms such as the mouse will be an essential lens for interpreting the evolution and function of human cell types.

# FUNCTIONAL INTERPRETATION OF GENETIC VARIANTS USING DEEP LEARNING

Gabriel E Hoffman

Department of Genetics and Genomics, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY

Identifying causal variants underling disease risk and the adoption of personalized medicine are currently limited by the challenge of interpreting the functional consequences of genetic variants. Predicting the functional effects of protein-coding variants is now routine in research and increasingly in some clinical applications. Yet the vast majority of variants in the human genome are non-coding, and predicting the functional consequence of these non-coding variants and prioritizing these variants for functional validation remains a major challenge.

Non-coding variants contribute to disease risk by regulating gene expression, and this effect is driven in large part by the genetic regulation of transcription factor binding and histone modification. Predicting the functional effect of non-coding variants on the epigenome will distinguish benign variants from variants with the potential to confer disease risk. Deep convolutional neural networks have recently been used to develop predictive models linking the genome sequence to splicing (1), protein binding (2), and the discrete presence or absence of a signal from epigenomics assays (3, 4).

Here we introduce a deep learning framework for functional interpretation of genetic variants (DeepFIGV). We develop predictive models of quantitative epigenetic variation in DNA-seq and H3K27Ac, H3K4Me3, and H3K4Me1 histone modifications from 75 lymphoblastoid cell lines (LCL). Modeling quantitative variation, integrating whole genome sequencing, and training the models on many experiments from the same cell type and assay improves the power to identify variants with functional effects on the epigenome.

The vast majority of variants have no predicted effect, yet variants with large predicted effects are enriched in transcription factor binding sites, CpG islands, promoters and 5' UTR. On average, rare variants have larger effect sizes than common variants, consistent with negative selection against variants that effect the epigenome. DeepFIGV variant scores are highly concordant with expression QTLs, and the scores accurately predict allele specific binding of transcription factors in independent experiments. Moreover, variants with large predicted effects are enriched for variants associated with autoimmune disease.

This application of the DeepFIGV framework illustrates the value of interpreting the functional consequences of non-coding variants, and sets the stage for applying DeepFIGV to epigenomic data from cell types and tissues relevant to additional diseases.

Finally, we have developed a public resource of the DeepFIGV predicted functional scores for 500 million variants.

1. H. Y. Xiong, et al. Science. 347, 1254806–1254806 (2015).
2. B. Alipanahi, et al. Nature Biotechnology. 33, 831–838 (2015).
3. J. Zhou, O. G. Troyanskaya. Nature Methods. 12, 931–934 (2015).
4. D. R. Kelley, et al. Genome Research. 26, 990–9 (2016).

# RECONSTRUCTION OF SUBCLONAL TUMOR EVOLUTION FROM RAPID AUTOPSY DATA REVEALS NOVEL PATTERNS OF AGGRESSIVE METASTATIC COLONIZATION

Xiaomeng Huang[1,2], Yi Qiao[1,2], Samuel W Brady[1], Adam Cohen[1], Andrea Bild[3], Gabor T Marth[1,2]

[1]University of Utah, Salt Lake City, UT, [2]USTAR Center for Genetics Discovery, Salt Lake City, UT, [3]City of Hope, Duarte, CA

Metastatic breast cancer is an advanced-stage disease in which the cancer cells spread to distant organs. To understand the patterns of metastatic colonization in a patient who presented with aggressive disease, we collected tumor biopsies at initial diagnosis and at mastectomy necessitated by the patient's relapse; as well as twenty-six metastatic tumors across seven organs and two normal tissue control skin biopsies via a rapid autopsy procedure within hours after the patient's death.

All biopsy samples were subjected to 60X Illumina whole genome sequencing. Our analysis revealed extensive chromosomal changes including amplifications, deletions, LOH and translocations, as well as known driver mutations in RB1, TP53, and PTEN in all biopsy samples. We used the CNV and LOH data to reconstruct the phylogenetic relationships among the tumor samples. Using an extension of our published SubcloneSeeker algorithm, we utilized the somatic SNV allele frequencies in copy number-normal regions of the tumor genomes to refine these phylogenetic relationships, and to construct a detailed map of subclonal expansion that led to metastatic colonization of distal organs.

Our analysis indicates that the primary breast tumor collected at diagnosis already had metastatic potential, and contained 85% of all the mutations that were present in the later metastases. Subclonal analysis indicates early metastatic escape into the lung, well before the mastectomy procedure could have saved the patient. We identified four distinct waves of metastatic colonization, first into the abdominal organs, then two separate waves into the lymph nodes, and then the brain and bones. Detailed subclonal analysis in this aggressive tumor reveals both monoclonal and polyclonal seeding of specific metastatic sites. We also observed, for the first time in clinical setting, a novel seeding pattern: metastatic "recolonization" of an already established metastatic site in the lung. The lung plays a central role in the metastatic spread of the tumor, serving as an "incubator" and a "jumping board" in the colonization of new organs and new sites. Independent transcriptomic analysis conducted on the basis of bulk RNA-Seq data at many of the biopsied metastatic sites confirmed the phylogenetic relationship derived from the bulk whole-genome DNA sequencing data. This analysis also revealed that genes promoting brain and bone metastasis were over-expressed in brain and bone metastatic samples.

The high number of biopsied sites in this study, 30 in all, allowed us to reconstruct the evolution of the aggressive disease in our patient with unprecedented resolution, and to identify characteristic patterns of metastatic colonization. If confirmed in additional, similarly high- resolution datasets, these patterns will lead to better understanding of the metastatic process, and guide effective clinical intervention.

# A SURVEY OF 220,000 PEOPLE FROM ALL OF CHINA DEMONSTRATES THAT COMPLEX POPULATION GENETIC PROCESSES HAVE SHAPED CHINESE GENOMES

Zhuoyi Huang[1,2], Navin Rustagi[2], Desheng Liang[3,4], Feng Tian[1], Jiani Li[2], Xiaoyan Ge[2], Fan Xia[2], Xiaoming Liu[5], Yu Zhang[1], Kun Wang[1], Jinchuan Xing[6], Heshan Lin[7], Li Jin[8], Yiping Shen[9], Lynn B Jorde[10], Lingqian Wu[3,4], Fuli Yu[1,2]

[1]Berry Genomics, Beijing, China, [2]Baylor College of Medicine, Houston, TX, [3]Central South University, Center for Medical Genetics, Changsha, China, [4]Hunan Jiahui Genetics Hospital, Changsha, China, [5]University of Texas Health Science Center at Houston, Houston, TX, [6]The State University of New Jersey Rutgers, Piscataway, NJ, [7]Alibaba Cloud Computing, Hangzhou, China, [8]Fudan University, Shanghai, China, [9]Guangxi Maternal and Child Health Hospital, Nanning, China, [10]University of Utah, Salt Lake City, UT

The population structure of Chinese has only been coarsely studied by HGDP, HapMap, and 1000 Genomes due to limitations in sampling bias of ethnicities and geography, sample size, genomic coverage, or allele frequency. Therefore, the complexity of demographic process in shaping Chinese genomes is poorly understood. In our study, we interrogated 220,000 Chinese individuals from 31 provinces in China through the Non-Invasive Prenatal Test, performed whole genome sequencing and surveyed the allele frequency at provincial resolution for 25.8 million SNVs. We developed a novel machine learning based informatics method and a Bayesian framework for variant calling and allele frequency estimation, and achieved high accuracy with false discovery rate <3%, and accurate allele frequency estimation when compared to ExAC and gnomAD. We applied Euclidean distance and leveraged clustering algorithm to systematically quantify genetic distance in a locus-by-locus manner among populations from each province in China. From the distribution from the whole genome, we identified significant outliers of genes or variants. These loci are correlated with geographical distributions, suggesting demographics (peopling, recent admixture) and adaptations were in the play in shaping Chinese genomes. EDAR, SLC24A5 and OCA2 showed that the observed allele frequency distribution reflected large-scale demographic migration and recent local ethnicity admixture within China. The patterns of allele frequencies in HFE, ADD1 and ABCC11 demonstrate pronounced clustering phenomena, due to adaptations such as lifestyles and food resources. We also discovered significantly elevated frequencies of pathogenic variants for recessive disorders in specific provinces due to founder effects - beta thalassemia (65x higher in Guangxi) and glucose-6-phosphate dehydrogenase deficiency (16x in Guangxi and Guangdong). Based on these novel datasets and new knowledge, we proposed a unified evolutionary model to account for the genetic diversity in modern Chinese genomes.

# MULTI-POLYGENIC RISK SCORING TO DEFINE ANOREXIA NERVOSA SUBTYPES

Christopher Hübel[1,2], Héléna A Gaspar[1], Jonathan R Coleman[1], Shing Wan Choi[1], Saskia Hagenaars[1], Kirstin Purves[1], Ken B Hanscombe[1,3], Paul O'Reilly[1], Cynthia M Bulik[2,4], Gerome Breen[1]

[1]King's College London, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [2]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Stockholm, Sweden, [3]King's College London, Department of Medical and Molecular Genetics, London, United Kingdom, [4]University of North Carolina at Chapel Hill, Department of Psychiatry, Chapel Hill, NC

Anorexia nervosa (AN) is a heritable psychiatric and metabolic disorder with high medical morbidity and mortality. Diagnostically, AN is currently divided into restrictive (AN-R) (i.e., weight loss via caloric restriction and/or increased exercise) and binge/purging subtypes (AN-BP)(i.e., restriction coupled with binge eating and/or purging behaviors). The subtypes differ in regard to comorbid psychopathology and somatic comorbidities, including bone marrow dysfunction and hypoglycemia. These subtypes are based on clinical observation, rather than underlying biology.

We are calculating ~100 polygenic risk scores (PRS) derived from genome-wide association studies of biomarkers, personality traits, and somatic and psychiatric illnesses in 500 AN cases selected from the UK Biobank (www.ukbiobank.co.uk). We then will cluster individuals based on PRSs using unsupervised learning techniques including machine learning and clustering to generate biologically informed subtypes of AN. This approach will require replication in independent datasets. Next steps will include linking phenotypic characteristics related to course of eating disorder, medical and psychiatric comorbidity, and outcome, to the polygenic profiles to further characterize genetically derived subtypes. Genetically-informed subtypes of AN have the potential to be more nuanced and biologically informed subcategorizations that can more reliably predict course of illness and prognosis and be used to tailor interventions to avoid adverse outcomes.

# GENOTYPE-FITNESS MAPPING IN CANCER CELL LINES USING CRISPR-CAS9

Elizabeth Hutton[1], Xiaoli Wu[3,2], Timothy Somerville[2], Bin Lu[2], Sofya Polyanskaya[2], Yusuke Tarumoto[2], Yali Xu[2,3], Yuhan Huang[2], Christopher Vakoc[2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cold Spring Harbor Laboratory, Cancer Center, Cold Spring Harbor, NY, [3]Stony Brook University, Department of Molecular Biology, Stony Brook, NY

Popular high-throughput genetic screens using CRISPR-Cas9 allow direct testing of the activity of genetic elements in their endogenous chromatin context. Reliable measurements of variant effect from these screens can quantify the impact of extremely deleterious variants which are not present in viable organisms, but these assays present unique challenges in signal extraction. Our work uses a probabilistic graphical model to analyze cell growth and proliferation in modified cell lines. This allows us to quantitatively evaluate the effects of variants within their genomic context, identifying essential gene domains and epistatic interactions. Using our hierarchical framework to capture multiple levels of CRISPR-specific technical noise, we examined 6 independently designed CRISPR libraries resulting in over 100 separate high-throughput assays. Our improved estimation of tissue-specific fitness effects provides information on the extreme end of the fitness distribution, and can be complemented with classical signatures of natural selection for essential pathway analysis, tissue-specific function prediction, and epistatic interactions specific to different cancer types.

# PHENOMEWIDE CONSEQUENCES OF GENE EXPRESSION VARIATION IN HUMANS

Alvaro N Barbeira, Rodrigo Bonazzola, Jiamao Zheng, Milton Pividori, Hae Kyung Im

University of Chicago, Medicine, Chicago, IL

Unprecedented advances in genome technology and the power of genome wide association studies have enabled the discovery of thousands of loci robustly associated with complex diseases and traits. Transcriptome imputation approaches such as PrediXcan prioritizes genes that are likely to mediate these associations seeking to advance our understanding of the mechanism. The human knockdown project promises to characterize the function of every human gene. We reasoned that small variation in gene expression levels due to genetic variation coupled with phenotype data can provide complementary and orthogonal evidence for the function of genes. With that goal in mind, we have performed a phenomewide (~3000 phenotypes) survey of the associations between imputed expression levels and phenotypes and made the full set of results publicly available, accessible interactively and programmatically. High level of replication between independent studies demonstrate the robustness of our results. Interestingly, we found that monogenic disease genes are enriched among significant associations for related traits, suggesting that smaller alterations of these genes may cause a spectrum of milder phenotypes. Consistent with the sharing across tissues of eQTL, we found that combining evidence across tissues increases our ability to detect associated genes. In summary, our catalog contributes to the characterization of the function of genes in humans by integrating genotype and phenotype data from millions of individuals and reference transcriptome datasets.

# NEW METHODS FOR DETECTING MOUSE T CELL RECEPTOR REPERTOIRE WITH SINGLE CELL GENE EXPRESSION

Sadahiro Iwabuchi[1], Hitomi Okada[1], Shugo Deshimaru[2], Shinichi Hashimoto[1]

[1]Kanazawa University, Department of Integrative Medicine for Longevity, Graduate School of Medical Sciences, Kanazawa, Japan, [2]University of Tokyo, Department of Molecular Preventive Medicine, Faculty of Medicine, Tokyo, Japan

The huge diversity of the T cell receptor (TCR) recognizes many types of "self" or "non-self" immune responses, and it is important task to understand antigen specific TCRα/TCRβ chains in the pathological conditions. A variable (V), joining (J) and constant (C) regions constitute TCRα chain and TCRβ chain is made up of V, diversity (D), J and C region. The most part of V regions is belonged to complementary determinant regions 1 (CDR1) and CDR2, and CDR3 has variable V(D)J regions. Recently, some methods for identifying CDR3 have been developed, but it is still challenging to know the pair of CDR3 in TCRα (VJ) and TCRβ (VDJ) from the same T cell. Here, we show our new barcode beads for detecting mouse CDR3 subtypes, especially, the pair of TCRα/TCRβ chains in a single T cell. The series of process in the methods are based on our previous strategy of single-cell transcriptome analysis by using barcoding micro-beads which can characterize complex heterogeneous patterns in cancer tissues by examining thousands of cells per experiment. One representative result indicated that new barcode beads could identify 16% of TCR repertoire from the cells such as the regulatory T cell or cytotoxic T lymphocyte. Among them, about 1.5% of TCRα/TCRβ pair was also detected; for example, TCRα (V2, J39) and TCRβ(V6-4, D1, J2-1) chains were belonged to a single T cell. Further improvement for barcode beads itself and the methods of sequencing are needed, but this single-cell transcriptome analysis with TCR repertoire could contribute to understand what the property of T cell in the pathological condition has what the repertoire of TCRα/TCRβ chains, resulting in the development of new biomarkers or antigen specific drugs for the immunological disorders.

# DYNAMIC HUMAN ENVIRONMENTAL EXPOSOME REVEALED BY LONGITUDINAL PERSONAL MONITORING

Chao Jiang, Xin Wang, Xiyan Li, Jingga Inlora, Ting Wang, Qing Liu, Michael Snyder

Stanford University, Genetics, Stanford, CA

Human health is heavily dependent upon environmental exposures, yet the diversity and variation in human environmental exposures is poorly understood. To this end, we have developed a novel highly sensitive method to monitor personal airborne biological and chemical exposures (collectively referred to as the environmental exposome) longitudinally by integrating a wearable device and multiple-omics measurements. By following personal exposomes of 15 individuals for up to 890 days and over 66 distinct geographical locations, we demonstrated that individuals are potentially exposed to thousands of pan-domain species and thousands of chemical compounds, including insecticides and carcinogens. In aggregate, over 2500 species were identified with great intraspecies diversity. We found that personal biological and chemical exposomes are highly dynamic and vary spatial-temporally, even for individuals located in the same general geographical region. We were able to construct a season-predictive model based on the pan-domain genera profile. Integrated analysis of biological and chemical exposomes revealed strong location-dependent relationships. Finally, we built an exposome interaction network and demonstrated the presence of distinct yet interconnected human- and environment-centric clouds, depicting extensive inter-species relationships derived from various interacting ecosystems such as human, flora, pets and arthropods. Overall, we describe a method to capture and analyze personal environmental exposures, and demonstrate that human exposomes are diverse, dynamic, spatiotemporally-driven networks that have the potential to impact human health.

# THE EMBL-EBI GENOME EDITING CATALOGUE

Thomas Juettemann, Sybilla Corbett, Myrto Kostadima, Fiona Cunningham, Daniel R Zerbino, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Recent discoveries in the area of genome engineering have revolutionized biomedical research. The available technologies have significantly changed our approach to studying cell function, with an impact on biomedical sciences similar to that of PCR and high-throughput sequencing. Specifically, the techniques derived from the discovery of the CRISPR/Cas9 system are ushering in a new wave of laboratory protocols that are undergoing rapid experimentation and modification[1]. These include a broad range of techniques from gene editing[2] to transcriptional activation and repression[3], as well as imaging experiments[4]. Protein activity can be controlled in both time[5] and space[6], further expanding the potential uses of this approach. A rapid increase in published experiments using CRISPR-related proteins has been driven by past successes and the increasing availability of commercial reagents.

Scientific databases provide an invaluable service, collecting high-quality, coherent data sets and streamlining information retrieval in a stable format. EMBL-EBI provides a range of databases for information sharing. Curation is required to capture the wealth of information generated by CRISPR experiments in a systematic fashion, enabling its visibility, discoverability and re-usability. To this end we are developing the EMBL-EBI Genome Editing catalogue to collect, curate and convey the large amounts of data arising from these techniques.

Having created a prototype website, we are continuing the ongoing process of assessing the depth of experimental data to capture, the set of suitable visualisations for experimental output and the systems that will facilitate submissions to the archive from the authors themselves. The standard operating procedures and best practices have been developed in line with other databases at EMBL-EBI, such as the GWAS Catalog[7], and through discussions with different groups of CRISPR users. All submitted datasets will be required to provide a common set of parameters, allowing for cross-comparison irrespective of their origin. By including data from genome-wide pooled screens and single gene experiments we hope to provide an integrated resource that will aid research and facilitate new discoveries.

1. Jinek, M. et al. Elife 2013, (2013)
2. Tzelepis, K. et al. Cell Rep. 17, (2016)
3. Gilbert, L. A. et al. Cell 159, (2014)
4. Qin, P. et al. Nat. Commun. 8, (2017)
5. Liu, K. I. et al. Nat. Chem. Biol. 12, (2016)
6. Nihongaki, Y., Kawano, F., Nakajima, T. & Sato, M. Nat. Biotechnol. 33, (2015)
7. MacArthur, J. et al. Nucleic Acids Res. 45, (2016)

# A NEW WORKFLOW BUILT ON WHOLE-GENOME PHYLOCSF DISCOVERS HUNDREDS OF HUMAN PROTEIN-CODING GENES, EXONS, AND PSEUDOGENES, SHEDDING LIGHT ON MANY DISEASE ASSOCIATIONS

Jonathan M Mudge*[1], Irwin Jungreis*[2,3], Toby Hunt[1], Jose M Gonzalez[1], James Wright[6], Mike Kay[1], Claire Davidson[1], Stephen Fitzgerald[5], Ruth Seal[1], Susan Tweedie[1], Liang He[2,3], Robert M Waterhouse[4], Yue Li[2,3], Elspeth Bruford[1], Jyoti Choudhary[6], Adam Frankish[1], Manolis Kellis[2,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom, [2]MIT, Computer Science and Artificial Intelligence Lab, Cambridge, MA, [3]Broad Institute, Cambridge, MA, [4]University of Lausanne, Department of Ecology and Evolution, Lausanne, Switzerland, [5]Wellcome Trust, Sanger Institute, Hinxton, United Kingdom, [6]The Institute of Cancer Research, Royal Cancer Hospital, London, United Kingdom

*Co-first author

The most widely appreciated role of DNA is to encode for protein, yet cataloging all human protein-coding genes and the variety of translated transcripts remains challenging. Previously, we developed PhyloCSF to identify the evolutionary signature of coding regions using multi-species genome alignments. Here, we report whole-genome PhyloCSF datasets for human, mouse, fly, worm, and mosquito, and use machine learning to find and prioritize candidate novel coding regions. As part of the GENCODE gene annotation project, we analyse over 1000 high-scoring human PhyloCSF regions, and confidently describe 144 protein-coding genes and 171 pseudogenes previously missing from the GENCODE geneset, as well as additional coding sequences within 301 existing coding genes or pseudogenes. The majority have not been described previously. Most were intergenic and required simultaneously extending the GENCODE transcript catalog using modern transcriptomics data. The translational potential of all of our novel protein-coding genes has been independently assessed by comparative annotation and analysis of other vertebrate genomes, which is essential to remove spurious ORFs and to correctly distinguish coding from pseudogenic sequence. We have confirmed the translation of several of the novel coding genes using mass spectrometry. We use SNVs to find evidence that purifying selection in our novel coding regions has continued in the human lineage. We estimate there are unlikely to be more than a few dozen human protein-coding genes yet to be discovered that bear the characteristic evolutionary signature of coding sequence. Finally, we use our improved annotations to find that 120 trait-associated variants previously annotated as noncoding are in fact missense coding variants, offering clues to their mechanisms. Altogether, our PhyloCSF resources will be an important tool for researchers seeking to interpret these genomes, while our novel annotations present exciting loci for further experimental characterisation.

# GENOME-WIDE META-ANALYSIS OF POLYCYSTIC OVARY SYNDROME IN WOMEN OF EUROPEAN ANCESTRY IDENTIFIES NOVEL LOCI

Tugce Karaderi*[1,2], Felix R Day*[3], Michelle R Jones*[4,5], Cindy Meun*[6]

[1]The Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [2]Eastern Mediterranean University, Department of Biological Sciences, Famagusta, Cyprus, [3]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, [4]Cedars-Sinai Medical Center, Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, Los Angeles, CA, [5]Cedars-Sinai Medical Center, Bioinformatics and Computational Biology Research Center, Los Angeles, CA, [6]Division of Reproductive Medicine, Department of OBGYN, Erasmus MC - University Medical Center, Rotterdam, Netherlands

On behalf of the PCOS Consortium.
*Authors contributed equally.

Polycystic ovary syndrome (PCOS) is a common complex disorder causing reduced fertility affecting 5-15% of reproductive-aged women worldwide. Characterized by metabolic disturbances, hyperandrogenism and chronic anovulation, its etiology is largely unknown but with a clear genetic component. To date, genome-wide association studies (GWAS) have delivered 16 PCOS loci. Here, we perform a large GWAS meta-analysis of PCOS in up to 10,074 cases and 103,164 controls of European ancestry (NIH criteria, 2,540 cases/15,020 controls; Rotterdam criteria, 2,669 cases/17,035 controls; self-reported, 5,184 cases/82,759 controls from 23andMe). This genomic control corrected fixed-effects meta-analysis included 10,637,747 imputed variants excluding markers with minor allele frequency <1% and imputation quality <0.3. We identified 15 independent loci ($P<5x10^{-8}$), four of which were novel. All but one association are consistent across the case definitions [near *GATA4/NEIL2*, $OR_{self-reported}$=1.08 (1.03-1.13); $OR_{Rotterdam}$=1.21 (1.14-1.28); $OR_{NIH}$=1.33 (1.26-1.41)]. We find significant genetic correlations ($P<8.9x10^{-4}$) with obesity, fasting insulin, type 2 diabetes, high-density lipoprotein cholesterol, menarche timing, triglycerides and cardiovascular risk factors. In Mendelian randomization analyses, both obesity ($P=1.6\times10-^{23}$) and fasting insulin ($P=1.7\times10^{-5}$) seem to play a causal role in PCOS, independent of each other. We see potential neuroendocrine (*FSHB, ZBTB16, GATA4/NEIL2, DENND1A, RAB5B, TOX3*), metabolic (*THADA*) and developmental (*YAP1, ERBB4, ERBB3, MAPRE1*) components to PCOS. Further characterization of the observed PCOS associations in relevant subphenotypes (sex hormone levels, hyperandrogenism, amenorrhea, polycystic ovarian morphology and ovarian volume) will provide more details about the potential functions of these variants. This large-scale study implicates a role of neuroendocrine and metabolic mechanisms in PCOS and is advancing our knowledge about its genetic architecture and etiology.

# THE SPECTRUM OF LOSS OF FUNCTION TOLERANCE IN THE HUMAN GENOME

Konrad J Karczewski[1,2], Laurent Francioli[1,2], Kaitlin E Samocha[1,2], Beryl Cummings[1,2], Daniel Birnbuam[1,2], Mark J Daly[1,2], Daniel G MacArthur[1,2]

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

Deciphering the function and essentiality of genes in the genome is a central problem in human genetics. Large-scale exome and genome sequencing panels, such as the Exome Aggregation Consortium (ExAC) and the Genome Aggregation Database (gnomAD), have provided a glimpse into standing genetic variation, including loss-of-function (LoF) variation. The presence of LoF variants at high rates suggests a gene's redundancy, while a significant depletion suggests strong selective pressures against these variants, and thus, the gene's essentiality. Understanding where each human gene lies along the spectrum between these extremes is important for prioritizing candidate disease genes and the development of inhibitory therapeutics.

Using a mutational model, which accurately fits observed neutral synonymous variation, to generate expected numbers of variants for each gene in the genome, we previously found that in a dataset of 60706 exomes from ExAC, 3230 genes were found to be significantly depleted (constrained) for LoF variation, even in the heterozygous state. Here, we apply an expanded constraint model to high-confidence LoF variants in 123136 exomes from gnomAD. With this increased power, we identify approximately 15% more constrained genes, capturing more known disease genes and putative essential genes. Furthermore, we incorporate allele frequency information to distinguish weak selection against heterozygous variation from strong selection against homozygous variation. Finally, we investigate the extent of population-specific constraint across major continental groups. These results show how large-scale datasets can reveal natural selection on each gene, even before the responsible phenotypic consequences are discovered, improving interpretation of disease variants and therapeutic targets.

# INSECTOR : A WEB-SERVER TO IDENTIFY OLFACTORY RECEPTOR GENES FROM INSECT GENOMES

Snehal D Karpe, Murugavel Pavalam, R Sowdhamini

National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, India

Few protein families are extremely diverse and huge leading to complications in gene annotations of the newly sequenced genomes. Classical examples for this are the protein families of olfactory receptors (ORs). ORs are membrane protein receptors that determine the diversity of the smells perceived by an organism and hence can range from 10s to 1000s per genome as per the requirements of each species. Comprehensive identification of these genes from sequenced genomes of insects will help in the development of future repellents and pesticides which are important for human health worldwide. From our previous analysis (1) it was discovered that the common genome annotation pipelines tend to provide partial coverage and erroneous gene boundaries for huge protein families like insect ORs. Such diverse families are currently corrected using time-consuming and error-prone manual curation. Here we present a web-server, 'insectOR' (http://caps.ncbs.res.in/insectOR/), which make this process simpler and semi-automated. The pipeline can be readily used to identify OR gene loci from genomes of economically important insects like bees or pests like mosquitoes, bed-bugs, cockroaches, etc. In short, when provided with thousands of alignments of well-annotated insect OR proteins to that of the genome of interest, the insectOR pipeline filters the best alignments, stitches the consecutive fragments and provides better gene models for OR coding genes. These can be further validated within the web-server with the optional analysis for presence of consensus transmembrane helices, motifs and 7tm_6 Pfam domain characteristic of insect ORs. InsectOR was tested on ORs from a solitary bee genome previously discovered from our lab (2) and on ORs from *Drosophila melanogaster*. It outperformed two general annotation pipelines (MAKER and NCBI eukaryotic genome annotation) at nucleotide level precision and sensitivity to detect ORs. Though insectOR is not tested outside insect OR protein family, the core pipeline should be able to work on any medium sized DNA/genome and for any other diverse protein family. Hence its use can also be extended to the identification of genes from other important protein families (e.g. GPCRs) from individual human chromosomes as well.

1. Karpe,S.D., Jain,R., Brockmann,A. and Sowdhamini,R. (2016) Identification of complete repertoire of *Apis florea* odorant receptors reveals complex orthologous relationships with *Apis mellifera*. Genome Biology and Evolution, 8, 2879–2895.
2. Karpe,S.D., Dhingra,S., Brockmann,A. and Sowdhamini,R. (2017) Computational genome-wide survey of odorant receptors from two solitary bees *Dufourea novaeangliae* (Hymenoptera: Halictidae) and *Habropoda laboriosa* (Hymenoptera: Apidae). Scientific Reports, 7, 10823.

# COMBINATORY USE OF TWO scRNA-SEQ ANALYTICAL PLATFORMS REVEALS THE HETEROGENOUS TRANSCRIPTOME RESPONSE IN LUNG ADENOCARCINOMA CELL LINES

Yukie Kashima, Yutaka Suzuki

University of Tokyo, Graduate School of Frontier Sciences, Kashiwa, Japan

Single-cell RNA-seq (scRNA-seq) is a powerful tool for revealing heterogeneity in cancer cells. However, each of the current single-cell RNA-seq platforms appears to have inherent advantages and disadvantages. Namely, the micro-chamber-based platform provides sufficient sequence coverage per cell but sparse coverage of the cellular population. The opposite is observed for the micro-droplet-based platform.

We combined the data from these two platforms to reveal variable transcriptome responses to an anti-cancer drug using five lung adenocarcinoma cell lines. Using the micro-chamber-based and micro-droplet-based platform, we generated scRNA-seq datasets. Based on the obtained datasets, we demonstrate that it is possible to estimate missing values for expression information in the micro-droplet dataset by statistical inference of the micro-chamber dataset. We further demonstrate that even in the cases where precise information for an individual gene cannot be inferred, it is possible to analyze the activity of given transcriptional modules. Interestingly, we found that two distinct transcriptional modules associated with the Aurora kinase and DUSP genes are aberrantly regulated in a minor population of cells, and thus, may contribute to the possible emergence of dormancy or eventual drug resistance within the population.

Combining the different single-cell RNA-seq platforms can be one effective approach to obtaining complete information for both continuous expression varieties and a sufficient size of the cellular population to understand transcriptional heterogeneity in cancers.

# DTH - DIRECTORY TO TRACK HUB

Hideya Kawaji[1,2]

[1]RIKEN, Advanced Center for Computing and Communication, Yokohama, Japan, [2]RIKEN, Preventive Medicine and Diagnosis Innovation Program, Wako, Japan

Track Data Hub (Bioinformatics. 30:1003-5, 2014) is a mechanism to share and visualize genomic profiles and annotations over the internet remotely, for example via the UCSC Genome Browser Database and the ENSEMBL genome browser. Besides preparation of data files, it requires generation of configuration files that organize data by grouping a set of profiles into a meaningful chunk with customized graphic parameters. While data file preparation can be automated as a part of data processing, configuration requires manual curation to reflect focus of the data in the context of study. Such curation becomes time consuming in particular cases where the number of genomic profiles increased. Here, I introduce a framework to generate configuration files for track data hub, based on a directory structure containing data files, DTH (Directory to Track Hub). Since data files are organized in a meaningful chunk under a directory in general, the tool enables us to generate reasonable configurations effectively.

# DIFFERENTIAL MUTATION ANALYSIS ACROSS GENE SETS IN CANCERS

Katarzyna Z Kedzierska[1,2], Nathan Sheffield[1], Aakrosh Ratan[1]

[1]University of Virginia, Center for Public Health Genomics, Charlottesville, VA, [2]Warsaw University of Technology, Department of Drug Technology and Biotechnology, Warsaw, Poland

Cancer is a heterogeneous disease influenced by complex interactions between the inherited genotype, the acquired mutations, and the cell microenvironment. Studies such as The Cancer Genome Atlas have focused on identifying genes that could be driving the tumoral phenotype by investigating somatic mutations - those present in tumor cells and absent in normal, adjacent tissue. The most common methods to determine such cancer drivers focus on finding genes that exhibit a higher than expected mutation rate. MutSigCV, for example, generates a background model for mutations that includes patient and genomic position based factors to estimate the probability that a base is mutated by chance. Based on those probabilities, each gene is assigned a score which is later compared against a threshold to select driver genes. Another approach is adopted by TUSON, which uses silent and benign mutations to model the background frequency. However, those methods are limited by the background mutation frequency which tends to be overestimated, whereas the functional mutation rate - underestimated.

Instead, we apply differential mutation analysis - a more direct method to account for the background rates. The approach has been introduced in Przytycki and Singh, 2017 and leverages publicly available variant calls from the 1000 Genomes Project. The authors rank-normalize genes based on the number of non-silent mutations in both cancer patients and healthy individuals and then identify genes with higher mutational burden in tumors. We extend the approach to gene sets to identify pathways with a significant mutational burden. Additionally, we incorporate informative priors such as scores of deleteriousness (i.e., CADD scores) to allow for better prioritization of driver gene candidates. This is especially beneficial in highly mutated cancers where methods identify many driver genes. We also perform the analysis with regard to ancestry and find instances where certain driver genes appear to play a role in specific subpopulations.

As an example, we identify gene sets related to hypoxia and RET pathway as significantly altered in PCPG samples. Among others, in colon adenocarcinoma after incorporating CADD scores, CDH10 is assigned a higher rank than by merely aggregating non-silent mutations. Almost all the mutations in this gene are localized in cadherin domains which have been shown to be of relevance to epithelial-mesenchymal transition (EMT). Furthermore, we show genes that are only significant in some subpopulations. For instance, alterations in RUNX1T1 and NFE2L2 in ESCA are more relevant in the east and south Asian population.

In summary, we present an approach to aggregate mutation burden across not only individual genes but also gene sets. Additionally, our approach leverages biological and populational information allowing for a more detailed understanding of cancer biology.

# HIGH-RESOLUTION GENOME-WIDE FUNCTIONAL DISSECTION OF TRANSCRIPTIONAL REGULATORY REGIONS IN HUMAN

Xinchen Wang[1,2], Liang He[2], Alham Saadat[2], Li Wang[1], Melina Claussnitzer[2], Manolis Kellis[1,2]

[1]Broad Institute, Broad, Cambridge, MA, [2]Massachusetts Institute of Technology, Computer Science, Cambridge, MA, [3]Beth Israel Deaconess Medical Center, Medicine, Boston, MA

Genome-wide epigenomic maps revealed millions of regions showing signatures of enhancers, promoters, and other gene-regulatory elements. However, high-throughput experimental validation of their function and high-resolution dissection of their driver nucleotides remain limited in their scale and length of regions tested.

Here, we present a new method, HiDRA (High-Definition Reporter Assay), that overcomes these limitations by genome-wide selection of accessible regions, cloning into self-transcribing episomal reporter constructs, massively-parallel measurements of reporter gene expression, and computational deconvolution of overlapping segments into high-resolution inferences.

We used HiDRA to test ~7 million DNA fragments preferentially selected from accessible chromatin in the GM12878 lymphoblastoid cell line. By design, accessibility-selected fragments were highly overlapping (up to 370 per region), enabling us to pinpoint driver regulatory nucleotides by exploiting subtle differences in reporter activity between partially-overlapping fragments, using a new machine learning model.

Our resulting maps include ~65,000 regions showing significant enhancer function and enriched for endogenous active histone marks (including H3K9ac, H3K27ac), regulatory sequence motifs, and regions bound by immune regulators. Within them, we discover ~13,000 high-resolution driver elements enriched for regulatory motifs and evolutionarily-conserved nucleotides, and help predict causal genetic variants underlying disease from genome-wide association studies.

Overall, HiDRA provides a general, scalable, high-throughput, and high-resolution approach for experimental dissection of regulatory regions and driver nucleotides in the context of human biology and disease.

# IDENTIFYING THE GENETIC AND ENVIRONMENTAL DETERMINANTS OF GENE EXPRESSION VARIATION IN AFRICANS

Derek E Kelly[1,2], Nicholas G Crawford[1], Yue Ren[3], Renata A Rawlings-Goss[1], Gregory R Grant[1], Meredith Yeager[5], Stephen Chanock[4], Alessia Ranciaro[1], Simon Thompson[1], Jibril B Hirbo[6], William Beggs[1], Thomas B Nyambo[7], Sabah A Omar[8], Dawit O Meskel[9], Gurja Belay[9], Christopher D Brown[1], Hongzhe Li[3], Sarah A Tishkoff[1]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]University of Pennsylvania, Genomics and Computational Biology, Philadelphia, PA, [3]University of Pennsylvania, Biostatistics, Philadelphia, PA, [4]National Institutes of Health, Division of Cancer Epidemiology and Genetics, Rockville, MD, [5]Frederick National Laboratory for Cancer Research, Frederick, MD, [6]Vanderbilt University, Medical Genetics, Nashville, TN, [7]Muhimbili University, Biochemistry, Dar es Salaam, Tanzania, [8]Kenya Medical Research Institute, Center for Biotechnology Research and Development, Nairobi, Kenya, [9]Addis Abada University, BiologyAddis Abada, Ethiopia

Gene regulation plays a predominant role in human evolution and complex traits, and high throughput methods are making the measurement of expression variation routine. As with many fields of genomics, however, studies have failed to capture the breadth of global genetic and environmental diversity by focusing primarily on Western individuals of European descent. Studies of Africans, who harbor the most genetic variation in the world and are exposed to a diversity of environmental variables and diets, are necessary to complete our understanding of how evolution has shaped human genetic and phenotypic diversity. To identify genetic and environmental contributors to gene expression variation in whole blood we have collected RNA sequencing data from 171 individuals representing 9 diverse African populations. Differential expression analyses uncover genes correlated with ancestry and environmental variables, clustering individuals by ancestry and diet. Combining expression data with SNP genotypes, we also map cis-eQTLs in our samples. The majority of identified eQTLs replicate in Europeans, though they can often be mapped to a more narrow credible set owing to the shorter tracks of linkage-disequilibrium in Africa. Conversely, those that fail to replicate are enriched for variants that are at moderate frequency in Africa and are low frequency or monomorphic in Europe. Using these eQTLs, allele frequency differentiation and scans of natural selection identify candidate genes and pathways that may have undergone positive selection in these populations. This study represents the most diverse study of gene expression in Africans to date and highlights the need to extend genomics studies to non-European populations.

# CTCF BINDING EVOLUTION HELPS MAINTAIN THE INTEGRITY OF TOPOLOGICALLY ASSOCIATING DOMAINS

Elsa Kentepozidou[1], Maša Roller[1], Christine Feig[2], Sarah Aitken[2], Ximena Ibarra[2], Duncan Odom[2,3], Paul Flicek[1,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom, [2]University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom, [3]Wellcome Trust Sanger Institute, Wellcome Trust Sanger Institute, Cambridge, United Kingdom

Topologically associating domains (TADs) are a fundamental component of the 3D chromatin structure in eukaryotes and play an essential role in transcriptional regulation. Despite their importance and high conservation across species, the mechanisms underlying their formation and evolution remain elusive. It is clear that the CCCTC-binding factor (CTCF) is an important player in TAD formation, although not sufficient to demarcate TAD boundaries alone. Additionally, studies investigating the impact of CTCF depletion on TAD formation have been inconclusive thus far.

We have investigated the role of CTCF binding in the formation and evolution of TADs by mapping CTCF binding genome wide using ChIP-seq in livers of five mouse species. Our results indicate that TAD boundaries commonly include both highly conserved and evolutionary dynamic CTCF bindings and that these TAD boundary-associated CTCF sites are often located more closely to each other than are CTCF binding sites in other regions of the genome. We hypothesize that these clusters of conserved and dynamic CTCF binding sites are critical to the evolutionary and functional stability of TAD borders and that the apparently “spare” CTCF sites are able to preserve CTCF binding if other sites are disrupted. We conclude that CTCF binding evolution plays an important role in maintaining TAD stability.

# FAST VISUAL EXPLORATION OF HUNDREDS OF GENOME-WIDE DATASETS

Peter Kerpedjiev, Danielle Nguyen, Nils Gehlenborg

Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Since the completion of the Human Genome Project and the rapid adoption of next generation sequencing technologies the quantity of genome-wide assays has increased rapidly. Genome browsers are used to visualize this data along the length of the genome where zooming and panning operate along the genomic axis. Multiple datasets are typically treated as individual entities and pulled from separate files. This can create a bottleneck for data retrieval, requiring separate access to each file. By combining separate sources of data, storing them in a single file and aggregating across multiple levels of resolution, we can rapidly retrieve hundreds to thousands of individual tracks, render them as textures and support fast, seamless zooming and panning across both the genomic and track axes.

We demonstrate an implementation of this type of data in HiGlass (http://higlass.io) where we show the contents of 256 epigenomic tracks zoomable from gene to genome resolution. Clearly visible are peaks common to all datasets as well as peaks particular to clusters of data or even individual samples. The ability to continuously zoom out to genome scale without losing context makes it easy to see larger scale features such as gene deserts, telomeres, centromeres and unresolved regions of the genome. The ability to zoom in along the track axis makes it possible to home in to a handful of tracks and examine their signal at a given genomic location in more detail.

Unlike contact matrices which are scaled uniformly across both x and y axes and regular genomic tracks which are scaled along the x axis, multi-sample data can be scaled independently across both axes. To facilitate efficient data retrieval and smooth zooming across large genomic distances, we aggregate along the genomic axis and store data at multiple resolutions. The second axis, the samples, remains unaggregated so that operations such as reordering and clustering can be performed without significant computational overhead.

By adding a comprehensive set of visual encodings, we can use the same data format and infrastructure to display a wide range of of genomic and epigenomic data. A stacked bar chart scaled according to the total column-wise values can be used to display chromatin state model data as epilogos (https://epilogos.altiusinstitute.org/). Standard heatmaps and row-wise line and bar charts can be used to show coverage tracks. More creative renderings can be used to develop sequence-based tracks such as motif logos.

Enabling all of these applications is a data format consisting of multi-resolution 2D data as well as the ability to dynamically render it using a variety of different visual encodings. This makes it possible to rapidly retrieve, display and rearrange hundreds to thousands of genome-wide samples as well as to explore them from the scale of genes to the scale of the genome.

# ESTIMATING CELL TYPE ABUNDANCE IN GTEX ENABLES INSIGHTS INTO CELLULAR MECHANISMS AND ORIGINS OF eQTLs

Sarah Kim-Hellmuth[1], François Aguet[2], Meritxell Oliva[3], Manuel Muñoz-Aguirre[4], Jie Quan[5], Valentin Wucher[4], GTEx Consortium[2], Hualin S Xi[5], Barbara E Stranger[3], Tuuli Lappalainen[1], Roderic Guigó[4], Kristin G Ardlie[2]

[1]New York Genome Center, -, New York, NY, [2]Broad Institute, -, Cambridge, MA, [3]University of Chicago, -, Chicago, IL, [4]Centre for Genomic Regulation, -, Barcelona, Spain, [5]Pfizer Inc, -, Cambridge, MA

The Genotype-Tissue Expression (GTEx) project has identified expression quantitative trait loci (*cis*-eQTLs) for the majority of genes across a range of human tissues. However, the interpretation of these eQTLs has been limited by the complex cell type composition of GTEx tissue samples. Here, we applied gene expression-based deconvolution approaches to estimate the cellular composition of GTEx tissues, based on gene expression profiles of 17,382 RNA-seq samples from 838 individuals across 49 tissues (v8 data release). We found that highly abundant cell types were robustly estimated across multiple methods and were consistent with cell types expected to be present in different tissues. Cell type composition explained a large part of inter-individual variation in gene expression, and provided insights into cellular events associated to pathological changes in GTEx individuals. To demonstrate that computationally inferred cell type abundances can be used as proxies to identify cell-specific eQTLs, we used monocyte estimates from GTEx whole blood RNA- seq data (670 individuals) to identify cell type-interacting eQTLs (i-eQTLs). We found 835 genome-wide significant (FDR 5%) i-eQTLs interacting with monocyte abundance; these showed high replication of eQTLs identified in purified monocytes. Notably, 25% of monocyte i-eQTLs had a nominal whole blood *cis*-eQTL p-value > 0.05, implying that a significant fraction of cell type-specific eQTLs is undetectable when studying heterogeneous bulk tissue expression. Finally, colocalization analysis of monocyte i-eQTLs with multiple immune-related diseases not only pinpointed the cellular origin of known whole blood eQTLs, but identified novel colocalized loci that were masked in bulk tissue.
Taken together, our results emphasize the importance of identifying genetic effects on gene expression at the cellular level, and demonstrate that cell type-specific effects can be uncovered in bulk tissues by integrating increasingly available single-cell and cell type-specific transcriptomic data. The GTEx tissues provide a unique opportunity to expand this approach to immune and stromal cell types in tissues where cell-specific eQTLs have not yet been characterized.

# IDENTIFYING SIGNATURES OF DIVERGENCE IN REGULATORY DNA ACROSS MAMMALS

James King[1,2], Vahan B Indjeian[3], Boris Lenhard[1,2]

[1]MRC London Institute of Medical Sciences, Computational Regulatory Genomics Group, London, United Kingdom, [2]Imperial College, Faculty of Medicine, Institute of Clinical Sciences, London, United Kingdom, [3]Genentech Inc., San Francisco, CA

A subset of conserved non-coding elements, which otherwise have high sequence similarity across mammals, have rapidly accumulated substitutions in humans. These Human Accelerated Regions, which tend to be enriched near developmental genes, have been proposed to have played roles in the evolution of human-specific phenotypes, including increased brain size and bipedalism. To explore how similar processes are shaping mammalian evolution, we extend the study of Accelerated Regions (ARs) to other mammals, and demonstrate that a significant proportion of conserved non-coding elements are accelerated in at least one of the lineages considered. We find an enrichment for elements that are independently accelerated in multiple lineages, uncovering conserved elements that are more prone to variations in evolutionary rate throughout mammalian evolution, and hypothesise that these regions may in some cases underlie convergent evolution. We show that ARs in different lineages are distributed similarly around developmental loci, often clustering around skeletal patterning genes, indicating a role for ARs in innovation of morphological novelty. In contrast to many other mammals, we find that human and other primate ARs are often more clustered around genes heavily expressed in the developing brain, including *TENM3* and *NPAS3*.

# SINGLE CELL RNASEQ AND IMMUNOFLUORESCENCE IMAGING OF JOINT STROMA IN RHEUMATOID ARTHRITIS REVEALS SPATIAL ORGANIZATION OF FIBROBLASTS REMODELED UNDER INFAMMATION.

Ilya Korsunsky[1,2,3], Kevin Wei[4], Michael Brenner[4], Soumya Raychaudhuri[1,2,3,4]

[1]Brigham and Women's Hospital, Division of Genetics, Boston, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Harvard Medical School, Department of Biomedical Informatics, Boston, MA, [4]Brigham and Women's Hospital, Division of Rheumatology, Immunology, and Allergy, Boston, MA

The stromal cell compartment consists of heterogeneous populations of spatially organized fibroblasts and mesenchymal cells that mediate normal and pathological processes. Rheumatoid arthritis (RA) is an autoimmune disease characterized by neovascularization, infiltration of immune cells, and expansion of stromal cell populations in the joint tissue. In a recent study, we identified the expansion of a transcriptionally defined population of fibroblasts, marked by CD90 expression. To further investigate the transcriptional and spatial heterogeneity of these fibroblasts in the context of the whole stromal compartment, we performed single cell RNAseq on over 50,000 stromal cells as well as immunohistochemistry (IHC) staining on patient derived tissue sections. Fibroblasts are spatially organized in the joint, from the tissue surface to the interstitium, with some populations surrounding deep interstitial blood vessels. Under the hypothesis that this localization is encoded in transcriptional programs, we performed an unbiased ordering of the cells along a smooth trajectory. Genes most correlated with position on this gradient, including CD90, are known markers for distinguishing surface from deep fibroblasts. IHC staining confirmed that trajectory associated genes formed a smooth gradient along an anatomical axis, from the surface to perivascular fibroblasts. Moreover, this spatial signature is lost after a single passage in culture, highlighting the role of the stromal microenvironment in maintaining fibroblast identity.
Further trajectory analysis revealed that fibroblasts were better organized into a branching structure than a single gradient. Deep interstitial fibroblasts bifurcated into two distinct branches. We found one branch to be enriched for genes involved in antigen presentation. The inflammed RA joint is characterized by infiltrating immune cells, absent in non-inflammed tissue. Fibroblasts in this immunogenic branch may be responsible for stimulating T cells to promote inflammatory activity. Trajectory analysis from non-inflammed joint fibroblasts found an absence of branching, as well as of expression of immunogenic genes. IHC showed that the expression of these immunogenic genes was localized to perivascular fibroblasts, associated with a distinct subset of blood vessels. Expression of these genes was missing in tissue slices from non-inflammed tissue. Combined, the results suggest that a distinct population of deep interstitial fibroblasts are associated with inflammatory remodeling in the RA joint.

# EVOLUTION OF NLR IMMUNE RECEPTORS IN FLOWERING PLANTS

Paul C Bailey[1], Erin Baggs[1], Christian Schudoma[1], Elisha Thynne[2], William Jackson[1], Gulay Dagdas[2], Matthew Moscou[2], Wilfried Haerty[1], Ksenia V Krasileva[1,2]

[1]Earlham Institute, Organisms and Ecosystems, Norwich, United Kingdom, [2]The Sainsbury Laboratory, TSL, Norwich, United Kingdom

The immune system of plants is innate yet capable of recognizing a diverse range of pathogens. A major class of intracellular plant immune receptors is characterized by common nucleotide binding and leucine rich repeat domains (NLRs). NLRs deploy a variety of mechanisms to rapidly diversify and evolve new recognition specificities. A recent discovery is the evolution of NLRs through fusions with other plant genes that resemble host targets and serve as 'baits' for pathogen effectors. Such NLRs with integrated domains (NLR-IDs) often require a helper NLR receptor to initiate signalling.

We have examined the evolution of NLRs across more than 50 genomes of flowering plants and showed variability in the rate of evolution between NLR clades. A few NLRs are extremely conserved from basal plants to monocots and dicots. However, the majority of NLRs show lineage-specific expansion. While NLR-IDs are present in all flowering plants and fusions occur at low rate across most lineages, we have shown the presence of a major clade in grasses that is a hotspot for new fusions. The mechanism of NLR-ID formation involves inter-chromosomal DNA exchange despite little evidence for the involvement of transposable elements. The genomic regions surrounding NLR-IDs are evolutionary dynamic undergoing frequent gain/loss of exogenous genes. We continue to examine the effect of genomic architecture on NLR evolution and are beginning to unravel how domestication of crops, such as maize, contributed to the evolution of the plant immune system.

# ADVANCING COLLABORATIVE RESEARCH IN SYSTEMS BIOLOGY USING OPEN-SCIENCE, CYBERINFRASTRUCTURE OF KBASE

Vivek Kumar[1], Sunita Kumari[1], Doreen Ware[1,2], Priya Ranjan[3], Nomi Harris[4], Bob Cottingham[3], Christopher Henry[5], Adam Arkin[4]

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

The U.S. Department of Energy (DOE) has invested substantially in environmental and biological system science research to investigate the complex interplay between biological and abiotic processes that influence soil, water, and environmental dynamics of our biosphere. The community that has grown around these efforts has recognized the need to lower the barrier to accessing computational tools, data, and results, and to work collaboratively to accelerate the pace of their research. The DOE Systems Biology Knowledgebase (KBase, kbase.us) is a free, open-source software and data platform designed to provide these needed capabilities.

KBase currently has over 160 scientific analysis tools in areas such as (meta)genome assembly, contig binning, genome annotation, sequence homology analysis, tree building, comparative genomics, metabolic modeling, community modeling, gap-filling, RNA-seq processing, and expression analysis. Users can build and share sophisticated workflows by chaining together multiple apps–for example, one could predict species interactions from metagenomic data by assembling raw reads, binning assembled contigs by species, annotating genomes, aligning RNA-seq reads, and reconstructing and analyzing individual and community metabolic models.

Its web-based user interface supports building, sharing, and publishing reproducible, annotated analysis workflows with integrated data. Additionally, KBase has a software development kit that enables the community to add functionality to the system. Visit http://kbase.us/ to learn how KBase might be useful in your research.

# SINGLE-BASE RESOLUTION OF AUTOIMMUNE DISEASE ASSOCIATIONS USING MOLECULAR PHENOTYPES

Kousik Kundu[1,2], Stephen Watt[1], Alice Mann[1], Katrina M De Lange[1], Louella Vasquez[1], BLUEPRINT Consortium[1], Lu Chen[1], Jeffrey C Barrett[1], Carl Anderson[1], Nicole Soranzo[1,2]

[1]Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]University of Cambridge, Haematopieisis, Cambridge, United Kingdom

In-depth understanding of molecular mechanisms of disease informs the development of new therapeutic approaches. Autoimmune diseases collectively affect almost 10% of the world's population. To date, close to a thousand genetic loci have been associated with the risk of autoimmune diseases through genome-wide association studies. Characterising the causal genetic variants, putative effector genes and molecular mechanisms underpinning these associations is the necessary next step to harness the power of these genetic discoveries. Here we extend the evaluation of molecular QTLs generated as part of the BLUEPRINT project (www.blueprint-epigenome.eu) to systematically map molecular mechanisms and causal genetic variants at 14 different autoimmune diseases with publicly-available summary statistics. We first recomputed molecular QTLs for high-resolution genetic, epigenetic, and transcriptomic profiling in three primary human immune cell types (i.e., monocyte, neutrophil, and T-cell) using a denser genotype map. We then used colocalisation analysis to identify shared genetic effects between each molecular and disease trait, testing 13,295 overlapping ($r2 \geq 0.8$) disease-molecular trait pairs. Among them, 10,758 (~81%) showed high posterior probability of colocalisation ($PP \geq 0.99$), corresponding to 170 unique, non-HLA loci across all diseases, of which approximately half were shared between two or more diseases. We next sought to test the relative resolution of disease and molecular QTLs for defining credible sets of causal variants at colocalised loci. We optimized a Bayesian fine-mapping framework for analysis of disease and molecular QTL summary statistics, and applied it to fine-map associations at 332 independent disease-QTL colocalised loci. We show that fine-mapping analysis applied to multiple layers of molecular traits systematically improves resolution of causal variants compared to disease summary statistics alone. On average, molecular QTLs yielded smaller 95% credible sets and a more resolved distribution of posterior probabilities for causal variants within each set. For example, we were able to resolve causal credible sets (>95% posterior probability) to less than 20 variants for approximately 65% of loci using molecular QTLs, compared to 45% based on disease summary statistics alone. More importantly, molecular QTLs provide interpretable molecular mechanisms at a large set of putative causal variants. For example, fine mapping of the *ITGA4* locus associated with inflammatory bowel disease yields smaller credible sets for expression (n=2 variants), H3K4me1 (n=3) and H3K27ac (n=5) QTLs compared to use of disease summary statistics alone (n=11), and highlights a putative role for one promoter variant affecting CEBPB binding. Overall, our analysis clearly demonstrates how the use of molecular data empowers the interpretation of disease associations.

# A CRISPR/CAS9-MEDIATED SCREEN OF CANDIDATE GENES IN THE REGULATION OF OPTIC FISSURE CLOSURE

Shyam Lakshmanan, Sunit Dutta, Tiziana Cogliati, Brian P Brooks

National Institutes of Health, National Eye Institute, Bethesda, MD

Purpose: Closure of the optic fissure, an opening at the ventral side of the developing vertebrate eye, is required for normal eye development and function. Failure of the two edges of the fissure to fuse together results in a potentially blinding congenital ocular condition known as coloboma, which accounts for ~10% of childhood blindness. Although the embryology of optic fissure closure is well understood, the gene networks underlying this process are largely unknown. By using laser capture microdissection (LCM) of tissues from optic fissure margins in mice embryos (E10.5-E12.5) and microarray, we identified 164 annotated candidate genes that are dynamically expressed during optic fissure closure. In order to investigate the specific roles of the candidate genes with no previous association with coloboma, we are generating knockout (KO) zebrafish lines using CRISPR/Cas9-mediated genome editing technology and screening for ocular coloboma and other eye abnormalities.

Methods: We used two different approaches to generate CRISPR-KO lines. 1) Target-specific single sgRNAs were designed and co-injected with Cas9 protein to introduce indels. 2) Pairs of sgRNAs were designed to create large exonic deletions (30-1100 bp) in the coding sequence. Efficiency of the designed sgRNAs were assessed and verified using CRISPR Somatic Tissue Activity Test (CRISPR-STAT), a fluorescent PCR-based method. Sanger Sequencing was used to evaluate the nature of the CRISPR-mediated large deletions.

Results: Genomic DNA isolated from F0 embryos injected with two sgRNAs and Cas9 were found to contain large deletions in the exonic regions of the targeted genes, suggesting that this deletion strategy works in the somatic tissue.

Conclusions: Injected F0 embryos will be raised to adulthood and backcrossed to identify founders exhibiting germline transmission. F1 embryos will be assessed for coloboma and other eye abnormalities. This screen will expand our understanding of the molecular mechanisms involved in optic fissure closure.

# IN VIVO DEPLOYMENT OF A MASSIVELY PARALLEL REPORTER ASSAY FOR THE VALIDATION OF ENHANCERS ACTIVE IN POSTNATAL BRAIN DEVELOPMENT

Jason T Lambert, Jessica L Haigh, Iva Zdilar, Tyler W Stradleigh, Alex S Nord

University of California Davis, Department of Psychiatry and Behavioral Sciences, Department of Neurobiology, Physiology and Behavior, Davis, CA

Enhancers recruit tissue- and cell-type-specific transcription factors to drive specific expression patterns. The regulatory activity of enhancers is thought to produce and organize gene expression patterns producing the vast diversity of cell types that make up tissues and organs in animal development. Various means of assessing chromatin state have proven useful in predicting the presence of enhancers, but it remains expensive and labor intensive to functionally validate predicted enhancers *in vivo*. However, recent advances in massively parallel reporter assays (MPRAs) make possible the large-scale screening of enhancers. We present progress toward functionally validating a set of predicted disease-relevant human enhancer sequences in the neonatal mouse brain via the MPRA approach using recombinant adeno-associated virus as an expression vector. Moving forward, this approach will allow us to validate and dissect the regulatory function of enhancers across multiple time-points and cell types in postnatal brain development.

# SINGLE-CELL MULTI-OMIC ANALYSIS OF INTRATUMORAL HETEROGENEITY AND IMMUNE LINEAGES IN PATIENT TUMORS FOR PRECISION ONCOLOGY

Billy T Lau[1], Anuja Sathe[2], Noemi Andor[2], Hanlee P Ji[1,2]

[1]Stanford University, Stanford Genome Technology Center, Department of Biochemistry, Palo Alto, CA, [2]Stanford University, Division of Oncology, Stanford, CA

Intratumoral heterogeneity and immune infiltrates are the predominant indicators predictive of cancer patient outcomes and thus motivates their comprehensive characterization. Intratumoral heterogeneity drives Darwinian evolutionary responses against therapy. Infiltrating immune cell lineages, distinguished by distinct transcriptional phenotypes, affects cancer progression and response to therapy. Both intratumoral heterogeneity and the associated immunologic status are driven at the genomic and transcriptional level. However, conventional bulk next-generation DNA and RNA sequencing is unable to resolve the complex genomic and transcriptomic contributions from individual cells.

In this study, we use massively parallel single-cell DNA and RNA sequencing with the goal of a complete understanding of a tumor. Freshly excised patient gastric and colorectal tumors were dissociated into suspensions for parallel single-cell DNA-Seq and RNA-Seq. Our study enabled (1) statistically robust assessment of intratumoral heterogeneity; (2) determination of differentially actionable events in the context of precision oncology; and (3) an integrative analysis of both genomic and transcriptional components of a tumor.

Single-cell DNA-Seq on dissociated patient-derived tumor cells yielded the separation of malignant, genomically-unstable cells from euploid cells. Genome-wide copy number calls and ploidy were determined for each cell. Importantly, the increased cellular throughput led to statistically robust inferences of phylogenetic relationships between subclones. The identification of differential CNVs across subpopulations enabled the potential assessment of whether particular targeted therapies would be successful; in one patient, we observed that differential amplification of regions in chromosome 7 and 8 across subpopulations resulted in significant implications for the selection of therapeutic targets.

Single-cell RNA-Seq yielded a medley of subpopulations with distinct transcriptional phenotypes. We clearly resolved multiple subsets of tumor epithelial cells reflecting transcriptional heterogeneity as well as stromal components including endothelial, muscle, fibroblast and immune cells. We also assessed each tumor's potential for immune checkpoint blockade therapy. Matching the gene expression of different clusters to chromosomal regions also yielded the potential correspondence of transcriptional subpopulations to those yielded by single-cell DNA-Seq, thus yielding a more complete and integrative landscape of patient tumors.

# MINING THOUSANDS OF GENOMES FOR RELIABLE STRUCTURAL VARIANT POPULATION ALLELE FREQUENCY ESTIMATES

Ryan M Layer[1,2], Brent S Pedersen[1,2], Aaron R Quinlan[1,2,3]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]University of Utah, Biomedical Informatics, Salt Lake City, UT

SNV population frequency is vital to our assessment of variant pathogenicity. Unfortunately, there is no adequate frequency resource for structural variants (SVs). While SNV detection considers every base in every sample, the number of possible SV configurations makes it intractable to interrogate all possible SVs. Instead, SV detection methods cluster evidence into sets of likely variants, then filter these sets to maximize sensitivity and specificity. We show that this filtering makes it challenging to use SV callsets to determine the prevalence of variants observed in new patients since it is impossible to tell if the variant is absent from the population or if it was filtered. We need a method to provide a full accounting of SVs among the data generated from projects such as TOPMed and the Centers for Common Disease Genetics.

To fix this we built STIX, an index that enables searching for SVs in thousands of samples across multiple cohorts. STIX reports the per-sample count of all concurring evidence, from which we can distinguish between common and rare SVs (deep evidence in many samples vs. no evidence). To test STIX, we indexed 2,504 genomes from the 1,000 Genomes Project (1KG) and 253 genomes from the Simons Foundation Genome Diversity Project (SFGD), then quantified the frequency of 40,746 cancer-related deletions from the COSMIC SV database. While only 0.9% of the COSMIC SVs appeared in the 1KG SV call set, STIX was able to identify evidence for 16.2% of the COSMIC SVs across the two cohorts (7.3% in 1KG and 8.8% in SFGD). 5.1% of these SVs appeared to be common and occurred in at least 10% of the samples, making it unlikely that theses SVs play a major role in cancer. Searching a STIX index of healthy genomes for SVs seen in patients has greatly reduced the number of candidate SVs that we considered in our disease studies (e.g., Treacher Collins syndrome, ALS). We are also exploring creating disease-specific STIX indices (e.g., breast cancer tumors) to help us identify recurring mutations.

STIX will also be useful for large-scale SV genotyping and can empower true population scale SV detection by jointly considering all samples at once, both of which are important to large sequencing projects.

# GENOTYPE AND DISEASE SHAPE MOLECULAR CO-REGULATION

Amanda J Lea[1], Meena Subramaniam[2], Arthur Ko[3], Päivi Pajukanta[3], Noah Zaitlen[2], Julien F Ayroles[1]

[1]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [2]University of California: San Francisco, Institute for Human Genetics, San Francisco, CA, [3]University of California: Los Angeles, Human Genetics, Los Angeles, CA

Correlation between gene transcripts or protein products is a fundamental feature of co-regulated networks. However, we do not understand how individual-specific factors (e.g., genotype or disease) may affect this correlation structure. Here, we present a flexible approach for asking whether the degree of correlation between two measures is predicted by a variable of interest.

First, we used extensive simulations to validate our approach. Second, we tested the hypothesis that disease perturbs co-regulation, by asking whether pairwise metabolite correlations systematically varied between healthy individuals and those with metabolic syndrome (using blood-derived NMR data from The Cardiovascular Risk in Young Finns Study, YFS; n=1564). We found that, overall, metabolite pairs tended to be similarly correlated in healthy people and in those with metabolic syndrome ($R^2=0.62$, $p<10^{-16}$). However, for a subset of metabolite pairs, disease status had strong effects: 1528 pairs were more correlated in healthy individuals relative to those with metabolic syndrome, and 619 metabolite pairs showed the opposite pattern (FDR<1%). This represents a 2.20x enrichment ($p<10^{-16}$) of metabolite pairs that become 'dysregulated' (i.e., that lose correlation) following the onset of disease. Further, metabolites that exhibited this pattern were strongly enriched for key functional classes, namely apolipoproteins (odds ratio=1.94, $p=2.34\times10^{-13}$), cholesterols (odds ratio=1.39, $p=1.58\times10^{-15}$), and small molecules involved in energy metabolism (odds ratio=1.43, $p=2.75\times10^{-13}$).

Finally, we asked whether genetic variation can affect the magnitude or direction of pairwise gene expression correlations, using whole blood-derived expression data from the Netherlands Study of Depression and Anxiety (NESDA; n=2477). Here, we found 484 unique 'correlation QTL', in which a genetic polymorphisms affects the degree of correlation between two gene pairs (FDR<10%). We replicated our correlation QTL in a separate set of NESDA participants (63% and 88% replicated at Bonferroni and FDR thresholds, respectively) and in the YFS cohort (0% and 64% replicated at Bonferroni and FDR thresholds, respectively). Further, we show that genes involved in correlation QTL are enriched for known transcription factors ($p=1.7\times10^{-3}$) and their targets ($p=6.7\times10^{-4}$), providing insight into their mechanistic underpinning. Together, our results reveal a new layer of biological regulation, by demonstrating that both genetic variation and metabolic syndrome contribute to inter-individual differences in molecular co-regulation.

# A GRAPH-BASED FRAMEWORK FOR UNIFIED IDENTIFICATION OF SHORT AND STRUCTURAL GENETIC VARIANTS IN WHOLE-GENOME SEQUENCING DATA

Dillon H Lee, Yi Qiao, Andrew Miller, Alistair Ward, Gabor Marth

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

Current variant calling practices are largely restricted to detection of short sequence variants such as SNPs and INDELs. Although several state-of-the-art tools exist for short-variant detection (e.g. GATK, FREEBAYES, SNPTOOLS), these tools often produce divergent variant callsets, especially INDELs, and it is often very difficult to validate and reconcile these differences. Furthermore, while it would be highly desirable to detect larger structural variants (SV), existing SV detector packages are typically difficult to integrate and result in callsets with a high false positive rate that must be validated by expert reviewers.

Here we present GRAPHITE (https://github.com/dillonl/graphite) a highly innovative approach for addressing these current limitations. In our approach, we start with a collection of sequence variants e.g. candidate variant calls generated by multiple alternative methods, resulting in a variant callset that is typically of high sensitivity, but low specificity. From this set of variants we then apply a novel variant adjudication procedure which allows us to identify both true positive and false positive calls. This is accomplished by constructing a graph using the linear reference genome as a backbone and allelic variants as branching nodes. Using our previously developed graph mapping algorithm, we re-map all reads from each of the samples contributing to the candidate calls. Since the graph contains paths through all alleles, reads can map through all putative variants, providing an effective platform to confirm or reject candidate alleles. This procedure results in a substantial improvement in specificity and allows for a more disciplined integration of callsets generated by high-performance calling software.

Visualizing variant callsets is critical for understanding complex INDELs and SVs. Current practices for visualizing complex INDELs and SVs involve viewing read alignment (BAM) files in the Integrative Genomics Viewer (IGV) and using careful sequence and breakpoint bookkeeping. These techniques are time consuming and can be difficult to ascertain true genetic structure. GRAPHITE addresses these issues by providing output which takes advantage of graph-based realignments. This is accomplished by producing the paths through the graph as FASTA output as well as the realigned reads in BAM format. This output can be visualized in the well-established and commonly used IGV viewer program. This provides the user with a clear view of how each read aligns to the proposed variants, and is particularly powerful when visualizing highly complex INDELs, duplications and large structural variants.

# A NOVEL APPROACH FOR DISCOVERING ONCOVIRUSES IN HUMAN CANCERS USING WHOLE-GENOME SEQUENCING

Xun Chen[1], Jian Cao[2], Dawei Li[1,3,4]

[1]University of Vermont, Department of Microbiology and Molecular Genetics, Burlington, VT, [2]Yale University, Department of Pathology, New Haven, CT, [3]University of Vermont, Department of Computer Science, Burlington, VT, [4]University of Vermont, University of Vermont Cancer Center, Burlington, VT

Oncoviral infection is responsible for 12-15% of human cancer cases. Convergent evidence from epidemiology, pathology, and oncology suggest that additional cancer viral etiologies remain to be undiscovered. Oncoviral profiles can be obtained from cancer genome sequencing data; however, wide-spread virus contamination and non-causal viruses complicate the process of identifying genuine oncoviruses. We propose a novel strategy to address these challenges by performing a virome-wide screening of "early stage" clonal viral integrations. To implement this strategy, we developed VIcaller, a novel platform to identify viral integration events that are derived from any known viruses and shared by a large proportion of tumor cells using whole-genome sequencing (WGS) data. The sensitivity and precision of our platform were validated with both simulated and benchmark cancer datasets. By applying this platform to The Cancer Genome Atlas bladder cancer WGS datasets, we have demonstrated that the strategy and platform are capable of identifying new cancer viral etiologies. This is the first study to systematically investigate the strategy of virome-wide screen of clonal integrations for identifying oncoviruses. We have demonstrated that clonal viral integrations provide the strongest evidence for discovering cancer viral etiologies.

# PARASITE INDUCED CHANGES IN THE GUT MICROBIOME AND THEIR PATHOPHYSIOLOGICAL IMPLICATIONS

Robert W Li[1], Yueying Wang[2], Joe F Urban, Jr.[3], Peter Geldholf[4]

[1]USDA-ARS, Animal Genomics and Improvement Laboratory, Beltsville, MD, [2]Henan Agricultural University, Zhengzhou, China, [3]USDA-ARS, Diet, Genomics and Immunology Laboratory, Beltsville, MD, [4]Ghent University, Faculty of Veterinary Medicine, Ghent, Belgium

Gastrointestinal parasitic infection has a profound impact on the composition and structure of the gut microbiome in many host-parasite systems. In this study, alterations in the porcine gut microbiome induced by infection with the parasitic nematode Ascaris suum were characterized to understand how these changes may affect intestinal inflammation and host immunity. Compared to uninfected controls, A. suum infection significantly increased the level of fecal short-chain fatty acids (SCFA), and significantly decreased microbial diversity indices, including Chao1, Shannon, and Simpson, especially in colon contents. Among the 16 phyla identified in colon contents, the relative abundance of Spirochaetes and Planctomycetes was significantly decreased by infection. The abundance of approximately 29% of genera detected was significantly altered by infection regardless of worm burden, including some of the most predominant genera, such as Prevotella and Faecalibacterium (absolute Linear Discriminant Analysis LDA score > 2.0). At the species level, infection significantly changed the relative abundance of 179 OTUs. Moreover, the infection markedly impacted the metabolic potential of the proximal colon microbiome, where the relative abundance of 58 metabolic pathways, including carbohydrate metabolism and amino acid metabolism, was affected. A disruption in microbial co-occurrence networks was detected during the infection with A. suum. Furthermore, a Mantel test suggested that node connectivity of the OTUs assigned to the class Mollicutes had a significant correlation with fecal parasite egg counts ($P < 0.05$) and the OTUs belonging to Betaproteobacteria had a strong and significant correlation with total SCFA, especially levels of acetate and propionate ($r = 0.87$; $P < 0.05$). Together, novel insights into the metabolic potential of the porcine gut microbiome in response to infection with A. suum could facilitate the development of alternative strategies to control parasitic helminth infections in farm animals and humans.

# SEROTYPE-SPECIFIC EVOLUTIONARY PATTERN OF ANTIMICROBIAL-RESISTANT *SALMONELLA ENTERICA*

Jingqiu Liao[1,2], Renato Oris[1], Laura Carroll[1], Jasna Kovac[3], Hongyu Ou[4], Martin Wiedmann[1]

[1]Cornell University, Department of Food Science, Ithaca, NY, [2]Cornell University, Field of Microbiology, Ithaca, NY, [3]Pennsylvania State University, Department of Food Science, State College, PA, [4]Shanghai Jiao Tong University, Department of Life Sciences and Biotechnology, Shanghai, China

The emergence of antimicrobial-resistant (AMR) strains in the important human and animal pathogen *Salmonella enterica* poses a growing threat to public health. Here, we study the genome-wide evolution of 90 *S. enterica* AMR isolates, representing one host adapted serotype (*S.* Dublin) and two broad host range serotypes (*S.* Newport and *S.* Typhimurium). Results show that AMR *S.* Typhimurium had a large effective population size, a large and diverse genome, AMR profiles with high diversity, and frequent positive selection and homologous recombination. Evolutionary pattern observed in *S.* Typhimurium are consistent with multiple emergence events of AMR strains and/or ecological success of this serotype in different hosts or habitats. AMR *S.* Dublin population showed evidence for a recent population bottleneck, and their genomes were characterized by a larger number of genes and gene ontology terms specifically absent from *S.* Dublin and a significantly higher number of pseudogenes as compared to other two serotypes. The recent population bottleneck and genome decay in AMR *S.* Dublin are congruent with its narrow host range. In addition, our data indicates a strong association between genome content of *S. enterica* and serotype. Approximately 50% of accessory genes, including specific AMR and putative prophage genes, were significantly over- or underrepresented in a given serotype. About 65% of the core genes showed phylogenetic clustering by serotype, including the AMR gene *aac(6')-Iaa*. Finally, our results suggest the important role of positive selection and phage-mediated homologous recombination in the evolution of AMR *S. enterica*. While cell surface proteins were shown to be the main target of positive selection, some proteins with possible functions in AMR and virulence also showed evidence for positive selection. Recombination mainly acted on prophage-associated proteins.

# REGULATION OF CELL TYPE-SPECIFIC ENHANCER COMMISSIONING BY GROWTH FACTOR SIGNALING

Emi Ling*, Thomas Vierbuchen*, Marty G Yang, Christopher J Cowley, Cameron H Couch, David A Harmin, Michael E Greenberg

Harvard Medical School, Department of Neurobiology, Boston, MA

* authors contributed equally

During development, a limited set of signaling pathways specify hundreds of distinct cell types. Translation of these generic signals into cell type-specific transcriptional responses underlies this striking diversification of cell fates. A conserved feature of multiple transcriptional programs that signal through one such pathway (growth factor signaling via Ras/MAPK) is the rapid transcription of a universal group of early-response gene transcription factors (ERG TFs) that control activation of late-response genes (LRGs), which are cell type-specific and mediate appropriate changes in cellular behavior. This general signaling mechanism achieves specificity through selective binding of ERG TFs to cell type-specific LRG enhancers that are thought to be commissioned by cell type-specific TFs during differentiation, but this model has not been rigorously tested and mechanisms of enhancer commissioning remain poorly understood.

To investigate these issues, we previously characterized the *cis*-regulatory elements that control growth factor-dependent transcriptional responses in fibroblasts from two genetically divergent mouse strains as a "mutagenesis screen" to identify TFs necessary for enhancer commissioning and Ras/MAPK signaling-dependent activation (Vierbuchen*, Ling* *et al.* 2017). These experiments led to the surprising finding that the ERG TFs Fos/Jun (AP-1) are in many cases required for enhancer commissioning by cell type-specific TFs. We showed that AP-1 TFs bind to nucleosome-occupied enhancers and directly recruit the SWI/SNF (BAF) chromatin remodeling complex to remodel nucleosomes and facilitate and/or stabilize binding of cell type-specific TFs. To comprehensively interrogate the *cis*-regulatory logic of stimulus-responsive enhancers *in situ*, particularly how AP-1 is targeted to specific enhancers, we are currently mapping functional stimulus-responsive enhancers in fibroblasts derived from crosses between C57BL/6J and four distinct wild-derived inbred mouse strains. Taken together, these findings suggest a new model for how growth factor signaling can drive cellular differentiation by facilitating enhancer commissioning via AP-1 TFs, and should provide further insight into the *cis*-regulatory logic of stimulus-responsive gene expression programs.

# GLOBAL lncRNA PROTEOGENOMICS WITH RIBOSEQ AND MASS SPECTROMETRY PINPOINTS PERSISTENT RIBOSOMAL IN-FRAME MIS-TRANSLATION OF STOP CODONS AS AMINO ACIDS IN MULTIPLE OPEN READING FRAMES OF A HUMAN LONG NON-CODING RNA.

Leonard Lipovich[1], Pattaraporn Thepsuwan[1], Anton S Goustin[1], Jason Herschkowitz[2], Juan Cai[1], Donghong Ju[1], Noah Alexander[3], Matthew McKay[3], Anne H Prather[4], Christopher E Mason[3], James B Brown[5]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University at Albany - SUNY, Cancer Research Center, Albany, NY, [3]Weil Cornell Medicine, Department of Physiology and Biophysics, New York, NY, [4]Star Rose Arts, Seattle, WA, [5]Lawrence Berkeley National Laboratory, Molecular Ecosystems Biology, Berkeley, CA

Two-thirds of human genes do not encode known proteins. Aside from relatively few non-coding RNA (ncRNA) genes with well-characterized functions, the ~40,000 non-protein-coding genes remain poorly understood, and a role for their transcripts as de-facto unconventional messenger RNAs has not been formally excluded. Ribosome profiling (Riboseq) predicts translational potential, but without independent evidence of proteins matching long ncRNA (lncRNA) open reading frames (ORFs), ribosome binding of lncRNAs does not prove translation. We previously mass-spectrometrically documented translation of specific human lncRNAs (Bánfai et al 2012 Genome Research 22:1646). We now examined lncRNA translation in human MCF7 cells, integrating strand-specific Illumina RNAseq, Riboseq, and mass spectrometry in biological quadruplicates performed at two core facilities (BGI, China; City of Hope, USA). UCSC Genome Browser-assisted manual annotation of imperfect (tryptic-digest-peptides)-to-(lncRNA-three-frame-translations) alignments revealed three peptides hypothetically explicable by “stop-to-nonstop” in-frame replacement of stop codons by amino acids in two ORFs of the lncRNA MMP24-AS1. To search for this phenomenon genomewide, we designed and implemented a pipeline matching tryptic-digest spectra to wildcard-instead-of-stop versions of repeat-masked, six-frame, whole-genome translations. Along with singleton stop-to-nonstop events at four other lncRNAs, we identified 24 additional peptides with stop-to-nonstop in-frame substitutions from multiple plus-strand MMP24-AS1 ORFs. Only UAG and UGA, never UAA, stop codons were impacted. All MMP24-AS1-matching spectra met the same significance thresholds as known-protein signatures. Targeted resequencing of MMP24-AS1 genomic DNA and cDNA from the same samples did not reveal any mutations, polymorphisms, or sequencing-detectable RNA editing. This unprecedented apparent gene-specific violation of the genetic code highlights the importance of matching peptides to whole-genome, not known-genes-only, ORFs in mass-spectrometry workflows, and suggests a new mechanism enhancing the combinatorial complexity of the proteome.
[Funding: NIH Director’s New Innovator Award 1DP2-CA196375 to LL.]

# FAST, MEMORY-EFFICIENT DECOMPOSITION OF PROHIBITIVELY LARGE GENETIC RELATEDNESS MATRICES

Zhi Xiong[1], Qingrun Zhang[2], Alexander Platt[3], Gustavo de los Campos[4], Quan Long[2]

[1]Shantou University, Computer Science, Shantou, China, [2]University of Calgary, Biochemistry and Molecular Biology, Calgary, Canada, [3]Temple University, 4. Center for Computational Genetics and Genomics, Philadelphia, PA, [4]Michigan State University, Epidemiology and Biostatistics, East Lansing, MI

A computational tool, OCMA (Out-of-Core Matrices Analyzer) is developed to relieve the current bottleneck of analyzing too large a genetic matrix, paving the path to the future precision medicine using lightweight hardware, e.g., personal computers.

The current and emerging biobanks enable researchers to leverage the genomic information of up to tens or even hundreds of thousands subjects to carry out various analyses. For many popular population-based analyses, e.g., principle component analysis (PCA), heritability estimation, and phenotype predictions using mixed models, the first-line characterization is the eigen-decomposition of the Genetic Relationship Matrix (GRM) that records pair-wise genetic similarities between all participating subjects. Thanks to the ever-growing size of the current biobanks, the normal size of GRM is reaching the genomic data itself! However, current popular tools need to load a GRM into the main memory. As a result, they are not scalable to large or even moderately sized GRMs without a high-performance computing cluster.

Using state-of-the-art computer technologies, we developed OCMA that can, using a personal computer, solve a moderate GRM (N=10,000) in 84 seconds and a huge GRM (N=100,000) in 5 days. OCMA is supported by two fundamental innovations. First, it utilizes memory mapping that moves the computation out of memory to the disk (a term called Out-of-Core in computer science). Second, it adopts a specific function in the Intel Math Kernel Library that offers tremendously fast calculation.

By envisioning the future use of genomic information to predict patients' traits such as disease risk and drug response on a daily basis, we believe OCMA has moved the first step towards the future personalized medicine based on lightweight infrastructure.

# USING DEEP LEARNING TO MODEL THE HIERARCHICAL STRUCTURE AND FUNCTION OF A CELL

Jianzhu Ma[1], Michael K Yu[1], Samson Fong[1], Keiichiro Ono[1], Eric Sage[1], Barry Demchak[1], Roded Sharan[2], Trey Ideker[1]

[1]University of California, San Diego, School of Medicine, San Diego, CA, [2]Tel Aviv University, Blavatnik School of Computer Science, Tel Aviv, Israel

Deep learning has revolutionized the field of artificial intelligence by enabling machines to perform human activities like seeing, listening and speaking. In modern ANN architectures, the connections between neurons as well as their strengths are subject to extensive mathematical optimization, leading to densely entangled network structures that are neither tied to an actual physical system nor based on human reasoning. Consequently, it is typically difficult to grasp how any particular set of neurons relates to system function. These are so-called 'black boxes', in which the input/output function accurately models an actual system but the internal structure does not. Such models, while undoubtedly useful, are insufficient in cases where simulation is needed not only of system function but also of system structure.

Here I will present a new computational framework, called DCell, an interpretable or 'visible' neural network (VNN) simulating a basic eukaryotic cell. The structure of this model is formulated from extensive prior knowledge of the cell's hierarchy of subsystems documented for the budding yeast Saccharomyces cerevisiae, drawn from either of two sources: the Gene Ontology (GO), a literature-curated reference database from which we extracted 2526 intracellular components, processes, and functions; or CliXO, an alternative ontology of similar size inferred from large-scale molecular datasets rather than literature curation. Subsystems in these ontologies are interrelated through hierarchical parent-child relationships of membership or containment. Such hierarchies form a natural bridge from variations in genotype, at the scale of nucleotides and genes, to variations in phenotype, at the scale of cells and organisms.

The function of DCell is learned during a training phase, in which perturbations to genes propagate through the hierarchy to impact parent subsystems that contain them, giving rise to functional changes in protein complexes, biological processes, organelles and, ultimately, a predicted response at the level of cell growth phenotype. We use the biological hierarchy to directly embed the structure of a deep neural network, enabling transparent biological interpretation. Unlike standard ANNs, DCell's simulations were tied to an extensive hierarchy of internal biological subsystems with states that could be queried. This 'visible' aspect raised the possibility that DCell could be used for in-silico studies of biological mechanism, of which we focused on four major types: 1) Explaining a genotype-phenotype association. 2) Prioritizing all important mechanisms in determination of phenotype overall. 3) Characterization of the genetic logic implemented by a process. 4) Discovery of new biological processes and states.

# INTEGRATING GENETIC VARIATION WITH METABOLITE ABUNDANCE AND GENE EXPRESSION IN *POPULUS TREMULA*

Niklas Mähler[1], Kathryn M Robinson[1], Torgeir R Hvidsten[2,1], Nathaniel R Street[1]

[1]Umeå University, Umeå Plant Science Centre, Umeå, Sweden, [2]Norwegian University of Life Sciences, Chemistry, Biotechnology and Food Science, Ås, Norway

Natural variation is widespread in Swedish aspen (*Populus tremula*), and much of this variation is highly heritable. The genetic variation is also high due to *P. tremula* being an obligate outcrossing species, but how the genetic variation translates into phenotypic variation is far from obvious. Using traditional GWAS, only a very small portion of the total phenotypic variance can be explained. When instead associating genetic variation with molecular phenotypes, such as gene expression or metabolite abundance, the signal is much stronger. To better understand how genetic variation contributes to variation in higher order, complex, phenotypes such as leaf shape or overall growth, we integrate genetic variation with variation in gene expression and metabolite abundance. The advances in quantification of expression of individual transcripts in recent years has also made it possible to identify transcript eQTLs. We find genes that on the gene level are not significantly associated with genetic variation, but have significant associations on the transcript level. We compared these associations with results from genome wide associations with metabolites and saw that SNPs associated with the abundance of some metabolites also were associated with the expression of some genes. These results can help to build an understanding of what components are involved in synthesising these metabolites and, in the long run, help understand the genetic basis of biomass variation. One of our goals is to make it possible for anyone to make complex interrogations of these kind of data, something that right now is reserved for specialists, and work is currently ongoing to build a public, interactive tool for this purpose.

# POOLHAP2: DE NOVO HAPLOTYPE RECONSTRUCTION FROM POOLED PATHOGEN NEXT-GENERATION SEQUENCING DATA.

Lauren Mak*[1], Jia Wang*[2], Chen Cao[1], Kai Ye[3], Daniel Jeffares[4], Quan Long[1]

[1]Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, Canada, [2]Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, [3]Department of Automation, Xi'an Jiaotong University, Xi'an, China, [4]Department of Biology, University of York, York, United Kingdom

Pooled sequencing (Pool-seq) is a next-generation sequencing (NGS) strategy where the genomes of several individuals from a population are grouped together and bulk-sequenced. Pool-seq provides an efficient and cost-effective alternative to genome sequencing of individuals or single cells, especially in contexts where pathogen genomes are inherently mixed. To determine the frequencies of individual-level polymorphisms and linkage disequilibrium (LD) from a population, the aggregated variation data must be de-convoluted in silico, an even more difficult task when haplotypes are not previously known and must be assembled de novo. Our proposed program, PoolHap, approximates the genotypic resolution of single-cell sequencing using only Pool-seq data by integrating population genetics models with genomics algorithms to reconstruct haplotypes.

PoolHap runs rounds of two steps iteratively until the solution converges, or when the global distribution of inter-pool haplotypes fully explains intra-pool variant data. The first step infers the initial haplotypes within each pool, and the second step takes the initial haplotype estimates, and approximates a global inter-pool distribution of haplotypes. This global distribution is provided as a priori information to direct the first step of the next round to improve its solution for intra-pool haplotypes. These refined intra-pool haplotypes in turn lead to a refined inter-pool distribution of haplotypes. We analyzed the population dynamics of 10X-coverage Plasmodium falciparum NGS data. This analysis yielded estimates of haplotype identities and frequencies that were in excellent agreement ($R = 0.91$) with expectations without any information about population's haplotypic composition. L. Mak and J. Wang contributed equally (* indicates co-first authorship).

# TRANSCRIPTOMICS ANALYSIS OF *CANDIDA GLABRATA* TREATED WITH THYMOQUINONE USING RNA-SEQ

Naveen Malik, Cynthia M Anderson

Black Hills State University, MSIG, Spearfish, SD

The goal of this project is to identify whether Thymoquinone may be of biomedical use in the treatment of fungal infections, particularly those caused by *Candida* species. We hypothesized that an active compound from the extract of *Nigella sativa* – Thymoquinone - would inhibit growth of the human fungal pathogens *Candida albicans*, *Candida krusei*,*Candida tropicalis*, and *Candida glabrata*. Therefore, Thymoquinone, as well as three other organic compounds known to be present in extracts of *Nigella sativa* - Carvacrol, Thymol, and Phenol - were tested against each species at concentrations ranging from 0.5µg/mL to 8.192 mg/mL to determine the minimum inhibitory concentration (MIC), while using antifungal standard Amphotericin B as a positive control. MIC experiments were performed using standards for antifungal susceptibility testing of fermentative yeasts published by The National Committee for Clinical Laboratories Standards (NCCLS - document M27A).

Results showed that Thymoquinone was the most effective of the four compounds against Candida species tested. Another round of susceptibility testing focusing exclusively on Thymoquinone was performed where concentrations from 0.5 µg/mL to 256 µg/mL were tested on the four species. Results showed MIC concentrations in between 16 µg/mL and 32 µg/mL for *C. albicans, C. krusei*, and *C. tropicalis*. However, *Candida glabrata* appeared to be the most resistant with an MIC between 32 µg/mL – 64 µg/mL.

*Candida glabrata* was then selected for a transcriptome study to identify the effect of Thymoquinone treatment on gene expression. Control and treatment group mRNA was isolated and extracted, and then bioinformatic analysis was performed in order to identify differentially expressed genes between the two groups. Preliminary results show 40 up-regulated genes and 30 down-regulated genes in the treatment group. Of particular note is the downregulation of genes that code for peroxiredoxin in *Candida glabrata*. Peroxiredoxins are ubiquitous thiol-specific proteins that have multiple functions including protection against oxidative stress. We hypothesize that Thymoquinone's effectiveness is due in part to the inhibition of *Candida glabrata's* defenses against Reactive Oxygen Species (ROS) thus causing oxidative stress.

# THE DYNAMICS OF mtDNA CHROMATIN-LIKE ORGANIZATION DURING METAZOAN EMBRYOGENESIS

Shani Marom[1], Amit Blumberg[1], Tal Cohen[1], Irene Kaplow[2], Anshul Kundaje[2], Dan Mishmar[1]

[1]Ben-Gurion University of the Negev, Life Sciences, Beer Sheva, Israel, [2]Stanford University, Genetics, Stanford, CA

mtDNA higher order organization is currently thought to be governed by a single coating protein – the transcription factor TFAM, which coats the mtDNA in a dose-dependent, sequence nonspecific manner. Nevertheless, recent work from our lab showed that human and mouse mtDNA display a conserved protein-DNA organization pattern in adult tissues. As a first step towards testing this possibility we have analyzed the dynamics of mtDNA gene expression (RNA-seq) and footprinting (ATAC-seq) during embryogenesis of several Metazoans. Analysis of RNA-seq data from mouse pre-implanted embryos revealed mtDNA genes expression is dramatically elevated during two cell embryos, and descent from 4 cell-ICM. Since mtDNA gene expression followed the nuclear pattern of gene expression during development, we asked whether higher order organization of the mtDNA had also similar dynamics, as in the nucleus. Our ATAC-seq data analysis of mouse, c. elegance and drosophila embryos strongly support a dynamic footprinting pattern during embryogenesis. While analyzing the mouse ATAC-seq mtDNA footprinting landscape, we noticed progressive increase in the density of the mtDNA footprinting sites during the course of embryogenesis with a notable increase in the density of sites in the inner cell mast (ICM), and an even increase during E6 and E7.2 stages. Notably, some ATAC-seq footprinting sites occurred in multiple stages, such as transcription termination site, ORI-L, pausing sites and others where stage-specific. Such dynamics was different among the tested organisms. Taken together, our study reveals, for the first time, a dynamic chromatin-like mtDNA organization during the course of early metazoan embryonic development.

# QUANTIFYING THE CONTRIBUTION OF RECESSIVE CODING VARIATION TO DEVELOPMENTAL DISORDERS

Hilary C Martin[1], Wendy D Jones[2], James D Stephenson[1,3], Juliet Handsaker[1], Giuseppe Gallone[1], Rebecca McIntyre[1], Michaela Bruntraeger[1], Matthew E Hurles[1], Jeffrey C Barrett[1], on behalf of the DDD study[1]

[1]Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]Great Ormond Street Hospital for Children, Clinical Genetics, London, United Kingdom, [3]European Molecular Biology Laboratory–European Bioinformatics Institute, EBI, Hinxton, United Kingdom

Large-scale exome sequencing studies have demonstrated the power of unbiased, genotype-first discovery of new disease genes, and can also be used to characterize the overall genetic architecture of rare disorders. We analyzed 7,446 exome-sequenced families from the Deciphering Developmental Disorders study, and estimated the genome-wide contribution of recessive coding variation in both known and as-yet-undiscovered genes. Our approach is the first to allow a properly calibrated estimate of recessive burden.

We found that the proportion of cases attributable to recessive coding variants was 3.6% in patients of European ancestry, compared to 50% explained by *de novo* coding mutations. It was higher (31%) in patients with Pakistani ancestry, due to elevated autozygosity, and similar to the contribution from de novos (30%). Half of the recessive burden was attributable to known genes. We were surprised that the recessive coding contribution was so low, since it has previously been argued that hundreds or thousands of recessive intellectual disability (ID) genes are yet to be discovered, which could imply that these undiscovered genes account for a large proportion of undiagnosed patients. Through simulations, we showed that the number of undiscovered fully penetrant recessive ID genes is less than 500 for all plausible parameters tested, and is more likely to be closer to 100.

Three genes were significantly enriched for biallelic damaging genotypes after stringent Bonferroni correction. One of these, *EIF3F*, is a new disease gene, and the signal ($p=1.2\times10^{-10}$) is driven by a single missense variant (frequency ~0.1%) that was homozygous in 9 DDD patients. *EIF3F* is a translation initiation factor, and we are currently measuring the variant's effect on translation in induced pluripotent cell lines edited with CRISPR. Another significant gene, *KDM5B* ($p=1.1\times10^{-7}$), has previously been reported as a dominant disease gene, but appears to follow a complex mode of inheritance, in which heterozygous loss-of-function (LoF) variants show incomplete penetrance and homozygous LoFs are fully penetrant.

Our results suggest that recessive coding variants only account for a small fraction of currently undiagnosed individuals in populations with European ancestry, and that future studies should focus on noncoding variants and polygenic risk.

# DETECTION OF DNA OF A LOW ABUNDANCE BY A HANDY SEQUENCER AND A PALM-SIZE COMPUTER

Bansho Masutani, Shinichi Morishita

Bioinformatics and Systems Biology, Science, Tokyo, Japan

**Background**
Detection of DNA sequence of a low abundance with respect to whole DNA sample is an important problem in the fields such as epidemiology, field research and virome. This is because the sample is highly contaminated with non-target DNA.
To solve this difficulty, a lot of methods have been invented so far but all of these methods require an additional time-consuming and costly procedure. Meanwhile, MinION sequencer developed by Oxford Nanopore Technology is considered as a powerful tool to tackle this problem because it gives a way to selectively sequence the target DNA strand. The main technology employed here is to reject an undesirable read from the specific pore in a sequencer by inverting the voltage of that pore, which is called "Read Until" by the community. Despite its usefulness, serious drawbacks exist in "Read Until". First, relatively small computational resources are available in field research and epidemiological application such as rapid detection of pathogenic bacteria. Second, a high-speed online classification algorithm is needed to make a prompt decision. Lastly, the lack of theoretical approach makes it difficult to justify a given algorithm.
**Result**
In this paper, we invent a classifier for any background DNA profile and prove its optimality with respect to precision. Precisely, for a pre-recorded mock sample, this proposed method can selectively sequence 100Kbp region, which consists of 0.1 % of entire read pool, and achieved about 500 times amplification. Furthermore, the algorithm is proven to be fast enough to be used in selective sequencing without any extensional computer resource such as GPU or high performance computing. It can run hand in hand with nanopore MinION, and can process 26 queries per second with a $500 palm-top next unit of computing (NUC) box with Intel®Core™i7 CPU.
Next, we prepared a mixed DNA pool composed of Saccharomyces cerevisiae and lambda phage where any 100Kbp region of S.cerevisiae consists of 0.1% of the whole sample. After "proof-of-concept" experiment to amplify 20 Kbp region of lamda pahge genome from that sample, we are currently under verification step of our algorithm.
**Conclusions**
Here we show that selective sequencing might amplify the target DNA sequence without any additional preparation steps by MinION sequencer. This method is likely to give us a deeper understanding of the field such as virome, plasmidome as well as to detect pathogens in medical application. Future research includes elongation of reference sequence up to mega-bp size.

# CHARACTERIZATION OF HLA ALLELES FROM TARGETED AND WHOLE GENOME SEQUENCES OF ETHNICALLY DIVERSE AFRICAN POPULATIONS

Eric Mbunwe[1], Jamie Duke[2], Alessia Alessia [1], Gurja Belay[3], Martin Maiers[4], Dimitri Monos[2], Sarah Tishkoff*
[1]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]Children's Hospital of Philadelphia, Immunogenetics Laboratory, Philadelphia, PA, [3]Addis Ababa University, Biology, Addis Ababa, Ethiopia, [4]National Marrow Donor Program/Be The Match, Bioinformatics Research, Minneapolis, MN

Multiple next-generation sequencing (NGS)-based systems have been developed and validated for human leukocyte antigen (HLA) allele typing using targeted approaches. However, there have been only sporadic reports as to whether credible HLA typing (two-field level) can be generated utilizing whole genome sequencing (WGS) data. Considering the ever-growing availability of WGS data and the importance of characterizing the HLAs in the context of many diseases and human population studies, it is relevant to explore the potential of some computational tools for deriving HLA typing from WGS data. For this purpose, we compared the accuracy of recently developed WGS-based approaches with conventional targeted HLA-NGS typing results. This comparison was performed using WGS data from sub-Saharan African populations and more specifically from Botswana, Cameroon, Ethiopia, and Tanzania. Whole genome sequences of unrelated individuals from 12 ethnically diverse populations were sequenced at 30x coverage on the Illumina HiSeq platform. The HLA genes A, B, C, DRB1, DRB3, DRB4, DRB5, DQA1, DQB1, DPA1, and DPB1 of three selected individuals were characterized using the HLA Holotype system by Omixon. The HLA typing of the three respective samples from the WGS data, was derived using the HLA Explore software (v 1.2) by Omixon. We found that at standard 2-field HLA typing resolution, HLA Holotype identified all possible alleles of all tested HLA loci with 5.6% of alleles having an ambiguous genotype due to incomplete characterization of exon 1 (1 DRB1 allele) and a lack of phase between exons (2 DPB1 alleles). The WGS data analysis using HLA Explore generated HLA genotyping results that were 100% concordant with the Holotype genotyping, however the ambiguity rate was increased compared to the Holotype approach at 11.1% of all allele calls, and failed to produce a genotype for 14.8% of allele calls (DRB1, DRB3). The ambiguities identified by HLA Explore were found in classical loci commonly genotyped for tissue transplantation (2 alleles each for A, DPB1 and DRB1), were due to an inability to phase within and between the exons of the HLA genes, and the ambiguous result from HLA Explore included the unambiguous pair of alleles determined by the Holotype approach. Two potentially new alleles were detected at the DPA1 locus by both methods. In conclusion, our findings suggest that 2-field HLA typing from WGS reads by HLA Explore approach needs further optimization. We plan to test the accuracy of other WGS based approaches including x-HLA and HLA Scan on the same dataset and report on the frequencies of HLA types present in this African cohort with implications for bone-marrow transplantation services for populations of African ancestry living in the USA.
*, authors contributed equally

# INVERSIONS HELP MAINTAIN SEXUALLY ANTAGONISTIC BALANCED POLYMORPHISM

Christopher McAllester, John Pool

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

Inversion polymorphisms and fixed differences are well documented across life, despite the unfit generation of unbalanced gametes from heterozygotes. Inversions may fix in linkage with beneficial alleles, but many inversions maintain intermediate, balanced frequencies, potentially by linking alleles that share conditional benefit, such as in ecotypes of Mimulus guttatus. In Drosophila melanogaster, paracentric inversions are surprisingly common. We hypothesize that balanced sexually antagonistic selection may be a cause, as many inversions maintain stable intermediate frequencies across broad and diverse African lowland habitats, inconsistent with ecological clines. We wrote a forward population simulation with D. melanogaster life history to model inversion evolution in a population under sexually antagonistic selection at infinite loci and with male reproductive skew. The model represents female choice on males by a representative quality score with a normal noise parameter. Alleles carry additive reproductive quality to males and multiplicative survival cost to both sexes. We present results demonstrating balancing selection on alleles with a range of antagonistic effects, the persistence of sets of such alleles only under linkage due to competitive effects, and finally the rise in frequency and stable persistence of inversions that establish such linkage associations between sets of sexually antagonistic alleles. These segregating antagonistic and non-antagonistic haplotypes benefit from linkage to the sex determining locus and so can facilitate neo-sex chromosome formation either by chromosomal fusion or novel sex determination loci. In an empirical extension, we are currently tracking inversion frequency changes between a pooled Zambian parental population and their embryo and aged adult offspring to detect correlation between the inversions and viability and mating fitness. Balancing selection, particularly sexual or ecological antagonistic selection, may be prominent in and relevant to contemporary evolution and local adaptation across species. It is a dynamic that needs more consideration in conservation genetics, adaptation, and human disease.

# STUDYING CLONAL CELL POPULATIONS WITH BULK EXOME AND SINGLE-CELL RNA-SEQUENCING DATA

Davis J McCarthy[1,3], Raghd Rostom[1,2], Yuanhua Huang[1], Daniel Kunz[2], Sarah Teichmann[2], Oliver Stegle[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute(EMBL-EBI), Statistical Genomics, Hinxton, United Kingdom, [2]Wellcome Sanger Institute, Cellular Genetics, Hinxton, United Kingdom, [3]St Vincent's Institute of Medical Research, Stem Cell Regulation, Fitzroy, Australia

The advance in single cell sequencing technologies has allowed an unprecedented resolution of the structure within cell populations. Single-cell DNA sequencing (scDNA-seq) has been used to infer the clonal structure of cells, adapting techniques from the cancer field using mutations acquired over time. Increasingly, however, single-cell RNA sequencing (scRNA-seq) data is being generated to understand gene expression in biological systems. The ability to determine the clonal structure of cells from scRNA-seq data would allow a greater understanding of structure within tissues, along with the mutational processes underpinning this.

Here, we demonstrate approaches for inferring clonal structure in a cell population by combining bulk exome and scRNA-seq data and using scRNA-seq data alone. We benchmark a two-step model for bulk and scRNA-seq data, a joint model for bulk and scRNA-seq data and a model for use only with scRNA-seq data. We present a two-step statistical method for the assignment of cells to a clonal tree structure, utilising detected mutations in sparse single-cell expression data. First, our approach applies established clonal-tree reconstruction methods from the cancer field to infer clonal structure in a cell population from bulk exome-sequence data. Next, we use a hierarchical Bayesian model incorporating false-positive and false-negative error terms to assign cells to clones using somatic mutations detected in sparse scRNA-seq read data.

We apply these three approaches to inferring clones and assignment cells to clones with bulk exome and scRNA-seq data from 62 healthy human fibroblast lines derived from distinct human donors. Even with great variation in clonal structure across populations, we show that we are able to robustly identify clonal populations and assign a majority of cells to clones, even from a limited number of somatic mutations sparsely detected in scRNA-seq data.

Assigning cells to clones using scRNA-seq data instead of scDNA-seq data reduces the cost-per-cell of studies such as this, while also enabling interrogation of the underpinning biology. We can analyse mutations across clones as when using scDNA-seq data, but the scRNA-seq data additionally provides a readout of the transcriptional state of each cell, allowing comparison of gene expression and transcriptional behaviour between cells assigned to different clonal populations. Variance component and differential expression analyses across our 62 fibroblast lines reveal genes that differ between cells in different clones within an individual, and allows identification of sets of genes and pathways with clonal differences observed in multiple individuals. This approach has the potential to enlighten on mutational processes at single-cell resolution, applicable in both cancerous and healthy tissues.

# TESTING ROBUSTNESS IN THE 'TT METHOD' – A SIMPLE AND ANALYTICAL METHOD TO INFER MODEL PARAMETERS UNDER A SPLIT MODEL

James McKenna, Per Sjödin, Mattias Jakobsson

Uppsala University, Organismal Biology, Uppsala, Sweden

The TT method estimates split times and effective population sizes under a simple split model assuming an infinite sites model and independence between sites. The split model assumes a constant population size before the split and no migration between populations after the split, but makes no assumptions about the population size processes more recent than the split time. The estimated parameters are: the expected number of generations to the split in branch 1 ($T_1$), the expected number of generations to the split in branch 2 ($T_2$), the probability of not coalescing before the split in branch 1 ($\alpha_1$), the probability of not coalescing before the split in branch 2 ($\alpha_2$), the expected number of generations to coalescence of lineages in branch 1, given that they coalesce before the split ($V_1$), the expected number of generations to coalescence of lineages in branch 2, given that they coalesce before the split ($V_2$) and the size of the ancestral population ($\theta$). All parameters except $\alpha_1$ and $\alpha_2$ are scaled by the mutation rate. By sampling pairs of gene copies from two populations, it is possible to derive closed analytical formulas for the probabilities of all the possible sampling configuration in terms of $\alpha_1$, $\alpha_2$, $T_1$, $T_2$, $\theta$, $V_1$ and $V_2$. Given sequence data from a pair of individuals, and assuming the above model, there are only 8 possible polymorphism patterns. The TT method calculates the counts of each polymorphism pattern and uses them for parameter estimation. The TT method estimates the population split time twice, ($T_1$ & $T_2$), once for each branch in the split model, which can be particularly suitable when applied to temporally structured DNA samples. Robustness of TT method parameter estimation is investigated through the simulation of polymorphic datasets under a variety of demographic scenarios that violate its basic model assumptions, and its performance compared against that of an alternative method of parameter inference, GPhoCS (Gronau *et al.* 2011). The TT method is shown to be a robust and computationally efficient method of parameter inference.

# TRANSCRIPTOME DIVERSITY AND ALTERNATIVE SPLICING OF BRAIN-EXPRESSED TRANSCRIPTS IN ADULT PSYCHIATRIC DISORDERS

Nirmala Akula[1], Robin Kramer[2], Qing Xu[2], Kory Johnson[3], Stefano Marenco[2], Jose Apud[2], Brent Harris[2], Pavan Auluck[2], Barbara K Lipska[2], Francis J McMahon[1]

[1]Human Genetics Branch, National Institute of Mental Health, Bethesda, MD, [2]Human Brain Collection Core, National Institute of Mental Health, Bethesda, MD, [3]Bioinformatics Section, National Institute of Neurological Disorders & Stroke, Bethesda, MD

Gene expression studies in post-mortem brain have provided valuable clues to the biological basis of psychiatric disorders, but few such studies have embraced the transcriptional complexity of the brain, where numerous transcript variants are generated by alternative splicing and other mechanisms.

We performed high-depth sequencing of RNA (RNA-seq) on ribosome-depleted libraries derived from subgenual anterior cingulate cortex, a brain region that has been implicated in psychiatric disorders. A total of 200 samples were studied, 39 from people with bipolar disorder (BD), 46 with schizophrenia (SCZ), 54 with major depression (MDD), and 61 without a known psychiatric disorder (controls). Stranded, paired-end sequencing of high-quality RNA (RIN $\geq$6) was performed on the Illumina HiSeq 2500. Of 54 billion 125 bp reads, 137M properly-paired reads/sample mapped to the reference genome (hg38) by HISAT2. StringTie identified 21K ENSEMBL genes with at least 10 reads each. These harbored over 85K transcripts, of which 44% were abundant (>100 reads). After quality control and quantile normalization, differential expression was estimated at the gene and transcript levels using DESeq2, with correction for RIN, race, and GC content. Implicated genes were further investigated for patterns of alternative splicing using Leafcutter.

Many of the same genes were differentially expressed in multiple disorders (mean gene overlap 42%). Compared to controls, the overlapping genes showed a consistent direction of differential expression and a significant positive correlation in fold-change values across all 3 disorders. In the transcript-level analysis, some of the same transcripts were differentially expressed in multiple disorders, but the mean overlap of differentially-expressed transcripts across disorders was only 18%. About 10% of differentially-expressed transcripts contained clusters that were alternatively spliced by Leafcutter.

To our knowledge this is the deepest RNA-seq study in a large sample of human postmortem brain tissue. The results illustrate the enormous diversity of brain-expressed transcripts, demonstrate that major psychiatric disorders can be distinguished at the transcript level, and suggest that alternative splicing is one mechanism through which transcript-level differences may arise.

# TRIO WHOLE GENOME SEQUENCING AS A TOOL FOR GENE DISCOVERY AND GENETIC DIAGNOSIS IN CHILDREN WITH MEDICAL COMPLEXITY

Gregory Costain*[1], Robin Z Hayeems*[7], Meaghan Snell[5], Maria Marano[1], Miriam Reuter[3], Danielle Veenma[3], Susan Walker[3], Raveen Basran[8], Eyal Cohen[1], Ronald D Cohn[1,2,6], Christian R Marshall[3,5,8], Stephen W Scherer[2,3,6], Cheryl Shuman[1,2], D James Stavropoulos[5,8], Julia Orkin[1], Stephen Meyn[1,2,4,6]

[1]Hospital for Sick Children, Paediatrics, Toronto, Canada, [2]University of Toronto, Molecular Genetics, Toronto, Canada, [3]Hospital for Sick Children, The Centre for Applied Genomics, Toronto, Canada, [4]University of Wisconsin, Center for Human Genomics and Precision Medicine, Madison, WI, [5]Hospital for Sick Children, Centre for Genetic Medicine, Toronto, Canada, [6]Hospital for Sick Children, Genetics and Genome Biology, Toronto, Canada, [7]Hospital for Sick Children, Child Health Evaluative Sciences, Toronto, Canada, [8]Hospital for Sick Children, Laboratory Medicine, Toronto, Canada

Children with medical complexity have ≥1 chronic condition(s), functional limitations, multiple subspecialist involvement, and high healthcare utilization. We hypothesized that whole-genome sequencing (WGS) has the potential to efficiently and effectively establish genetic diagnoses for children with medical complexity, and that cohorts of these children are enriched for novel genetic disorders.

Screening 542 children with medical complexity from a single Complex Care Program yielded 126 children (23%) suspected of having an undiagnosed genetic condition despite previous genetic testing. Eligible participants were evaluated through a clinical genetic assessment; WGS was performed in parallel with any outstanding conventional genetic testing. WGS data were analyzed for primary diagnostic variants as well as predictive variants for paediatric onset disorders.

In the first 21 probands with WGS data (including 13 trios, 1 dyad, and 7 singletons), 10 have reportable primary diagnostic variants. We identified six pathogenic/likely pathogenic variants in known disease genes, including one intronic variant predicted to affect splicing and one that was mosaic in the proband, as well as a promising variant of uncertain significance. De novo missense variants were also identified in three genes without published disease associations: *FBXW7*, *H3F3B*, and *RAC3*. Importantly, using public databases (DECIPHER, Matchmaker Exchange, and ClinVar), similarly affected individuals were rapidly identified around the globe and collaborations have been established to publish our joint findings.

Our initial experience applying WGS to children with medical complexity suggests that trio-based genome-wide sequencing is a high yield testing strategy for this patient population, which appears to be enriched for de novo mutations and novel genetic disorders. We are using RNA-Seq to analyze cases not solved by trio WGS alone.

# COMPREHENSIVE QUALITY CONTROL OF MANY SAMPLES USING IOBIO

Chase A Miller[1,2], Alistair Ward[1,2], Nielson Phu[2], Yi Qiao[1,2], Gabor Marth[1,2]

[1]University of Utah, Center for Genetic Discovery, 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.

# 250,000 INDEPENDENT GENETIC INFLUENCES ON DNA METHYLATION AND THE CONSEQUENCES OF THESE PERTURBATIONS: THE GODMC CONSORTIUM

Josine L Min[1], Gibran Hemani[1], Eilis Hannon[2], Kimberley Burrows[1], René Luijk[3], Koen F Dekkers[3], Elena Carnero Montoro[4,5], Juan Castillo-Fernandez[4], Johanna Klughammer[6], Christoph Bock[6], Jordana Bell[4], Bastiaan T Heijmans[3], Jonathan Mill[2,7], Caroline Relton[1]

[1]University of Bristol, MRC Integrative Epidemiology Unit, Bristol, United Kingdom, [2]University of Exeter, Medical School, Exeter, United Kingdom, [3]Leiden University Medical Center, Molecular Epidemiology, Leiden, Netherlands, [4]King's College London, Department of Twin Research and Genetic Epidemiology, London, United Kingdom, [5]Pfizer - University of Granada, Andalusian Government Center for Genomics and Oncological Research, Granada, Spain, [6]Austrian Academy of Sciences, CeMM, Vienna, Austria, [7]King's College London, IoPPN, London, United Kingdom

Natural inter-individual variation in DNA methylation levels is hypothesized to mediate genetic influences on disease. The Genetics of DNA Methylation (GoDMC) consortium (www.godmc.org.uk) comprises 50+ cohorts worldwide with 45,000 samples and has been established to perform a systematic and wide-ranging set of analyses using methylation and genetic data jointly, with a focus on identifying *trans*-mQTL for which large sample sizes are a necessity. Here we develop the largest mQTL catalog to date and illustrate its utility. In a meta-analysis of 36 European cohorts (N=27,750), after clumping we identified 248,607 genetic associations in *cis* ($p<1e-8$) and 23,117 in *trans* ($p<1e-14$). Often SNPs that influenced a CpG in *cis* also influenced other CpGs in *trans*. We used genetic co-localization analysis to construct 1331 distinct CpG communities (mean size = 8.59, max = 419) of *cis-trans* CpGs that share causal variants, which highlighted a role for architectural proteins.
We examined whether mQTL SNPs are a target of selection. We found that *cis* mQTL SNPs are enriched for SNPs with extreme $F_{st}$, singleton density scores and extended haplotype homozygosity. Using bi-directional Mendelian randomization to evaluate the causal relationships between DNA methylation levels and 680 complex traits, we identified 1176 unique CpGs having putative causal influences on 220 complex traits, including associations with lipid (n=658), anthropometric (n=356), auto-immune (n=336), psychiatric (n=110) and hematological (n=101) traits. In addition, we identified 23 phenotypes causally influencing DNA methylation levels of which auto-immune (n=4812), cancer (n=3267) and lipid (n=271) traits show widespread effects. Our results make a major contribution to emerging evidence about the genetic architecture of DNA methylation variation and shed light on how inter-individual variation in DNA methylation might be implicated in pathways to, and consequences of human disease.

# A COMMON PATTERN OF DNase-I FOOTPRINTING THROUGHOUT THE HUMAN MTDNA UNVEILS CLUES FOR A CHROMATIN-LIKE ORGANIZATION

Amit Blumberg[1], Charles G Danko[2], Anshul Kundaje[3], Dan Mishmar[1]

[1]Ben-Gurion University of the Negev, Life Sciences, Beer Sheva, Israel,
[2]Cornell University, Baker Institute for Animal Health, Ithaca, NY,
[3]Stanford University, Department of Genetics, Stanford, CA

Human mitochondrial DNA (mtDNA) is believed to lack chromatin and histones. Instead, it is coated solely by the transcription-factor TFAM, which lacks binding sequence specificity and packs the mtDNA into a bacterial-like nucleoid in a dose-dependent fashion. We asked whether mtDNA packaging is more regulated than once thought. To address this, we analyzed DNase-I-seq experiments in 324 human cell types and found, for the first time, a pattern of 29 mtDNA Genomic footprinting (DGF) sites shared by ~90% of the samples. Low SNP density at the DGF sites, and their conservation in mouse DNase-seq experiments, reflect strong selective constraints. Co-localization with known mtDNA regulatory elements and transcription pausing sites, suggest a transcriptional role for such DGFs. Altered mtDNA DGF pattern in IL-3 treated CD+34 cells offer first clue to their physiological importance. Taken together, human mtDNA has a conserved protein-DNA organization, which is likely involved in mtDNA gene expression regulation.

# HUMAN PRIMITIVE BRAIN DISPLAYS NEGATIVE MITOCHONDRIAL-NUCLEAR EXPRESSION CORRELATION OF RESPIRATORY GENES

Gilad Barshad, Amit Blumberg, Tal Cohen, Dan Mishmar

Ben-Gurion University of the Negev, Life Sciences, Beer Sheva, Israel

Oxidative phosphorylation (OXPHOS), a fundamental energy source in all human tissues, requires interactions between mitochondrial (mtDNA) and nuclear (nDNA)-encoded protein subunits. Although such interactions are fundamental to OXPHOS, bi-genomic co-regulation is poorly understood. To address this question, we analyzed ~8,500 RNA-seq experiments from 48 human body sites. Despite well-known variation in mitochondrial activity, quantity and morphology, we found overall positive mtDNA-nDNA OXPHOS genes' co-expression across human tissues. Nevertheless, negative mtDNA-nDNA gene expression was identified in the hypothalamus, basal ganglia and amygdala (sub-cortical brain regions, collectively termed the 'primitive' brain). Single cell RNA-seq analysis of mouse and human brains, revealed that this phenomenon is evolutionarily conserved, and both associate with brain cell types (involving excitatory/inhibitory neurons and non-neuronal cells) and by their spatial brain location. As the 'primitive' brain is highly oxidative, we hypothesized that such negative mtDNA-nDNA coexpression likely controls for the high mtDNA transcript levels, which enforce tight OXPHOS regulation, rather than rewiring towards glycolysis. Accordingly, we found 'primitive' brain-specific upregulation of lactate dehydrogenase B (*LDHB*), which associates with high OXPHOS activity, on the expense of *LDHA*, which promotes glycolysis. Analyses of co-expression, DNase-seq and ChIP-seq experiments revealed candidate RNA-binding genes and CEBPβ as best regulatory candidates to explain these phenomena. Finally, cross-tissue expression analysis unearthed tissue dependent splice variants and OXPHOS subunit paralogs, and offered revising the list of canonical OXPHOS transcripts. Taken together, our analysis provides a comprehensive view of mito-nuclear gene co-expression across human tissues and provides overall insights into the bi-genomic regulation of mitochondrial activities.

# TRANSPOSABLE ELEMENTS, STRUCTURAL VARIATION AND THE GRAPEVINE PAN-GENOME

Gabriele Magris[1], Rachel Schwope[1], Sara Pinosio[2], Emanuele De Paoli[1], Mirko Celii[1], Gabriele Di Gaspero[2], Michele Morgante[1,2]

[1]Università di Udine, Dipartimento di Scienze AgroAlimentari, Ambientali e Animali, Udine, Italy, [2]IGA, Istituto di Genomica Applicata, Udine, Italy

The analysis of variation in plants has revealed that their genomes are characterised by high levels of structural variation, consisting of both smaller insertion/deletions, mostly due to recent insertions of transposable elements, and of larger insertion/deletion similar to those termed in humans Copy Number Variants (CNVs). These observations indicate that a single genome sequence might not reflect the entire genomic complement of a species, and prompted us to introduce the concept of the plant pan-genome, including core genomic features common to all individuals and a Dispensable Genome (DG) composed of partially shared and/or non shared DNA sequence elements. The very active transposable element systems present in many plant genomes may account for a large fraction of the DG. Uncovering the intriguing nature of the DG, i.e. its composition, origin and function, represents a step forward towards an understanding of the processes generating genetic diversity and phenotypic variation. Additionally, since the DG clearly appears to be for the most part the youngest and most dynamic component of the pan genome, it is of great interest to understand whether it is a major contributor to the creation of new genetic variation in plant evolution as well as in the artificial selection processes of plant breeding. We have resequenced to high coverage more than 120 grapevine accessions and used a variety of approaches to detect structural variants of different size and origin, including de novo assembly of a selected set of genotypes. Additionally we have analysed the transcriptome using RNASeq, DNA methylation using BSSeq, histone modifications using ChIPSeq, chromatin accessibility using ATACSeq and chromatin conformation using HiC in order to perform allele specific analysis of gene expression, DNA methylation and chromatin structure to identify the genetic and epigenetic effects of the presence/absence polymorphism due to TE insertions. We will discuss the extent and composition of the pan genome in grapevine, the different mechanisms that generate and maintain the dispensable portion and the genetic and epigenetic effects of the structural variants caused by TE movement.

# NEUROSYSTEMATICS AND PERIODIC SYSTEM OF NEURONS: INSIGHTS FROM MILLIONS OF NEURONS SEQUENCED ACROSS PHYLA

Leonid L Moroz[1,2], Andrea B Kohn[1]

[1]University of Florida, Whitney Laboratory for Marine Bioscience, St. Augustine, FL, [2]University of Florida, Neuroscience, Gainesville, FL

There is more than one way to make a brain, and animals frequently use different molecular toolkits to achieve similar functional outcomes (=convergent evolution). However, the genomic bases of convergent evolution are mostly unknown. Here, we used single-cell sequencing (scRNA-seq, 10x Genomics) to identify hundreds of novel neuronal types in representatives of 12 phyla (>1.2 million neurons). The selected species illuminate significant transitions in the formation of neural and synaptic organizations as well as memory and elementary cognitions: from diffuse neural nets in basal metazoans to composite brains in cephalopods, insects, and chordates. 1) We revisited the animal phylogeny and developed a metric to quantitate each cell's type transcriptional relationships as well as criteria for neuronal homologies. 2) We showed that for many species tested, virtually all neurons are unique in their RNA modifications, non-coding RNAs, and secretory molecules, providing the foundation for the natural evolutionary neuronal classification. The discovered molecular complexity of the numerically "simpler" neural systems and the emerging single-cell data suggest the hypothesis: neurons are different not only because they have different functions, but also because neurons and circuits have different genealogies, and perhaps independent origins at the broadest evolutionary scale. 3) Origins of neurons (and synapses) from various types of ancestral secretory cells might have occurred at least three times during animal evolution. It also appeared that muscles evolved independently. 4) Our reconstructions suggest 9-12 independent events of nervous system centralizations (i.e., formations of composite brains) from a common bilaterian/cnidarian ancestor with diffuse-like neural systems. Thus, we set up a foundation for natural genealogical classification of cell/neuronal classes across phyla over one billion years of divergent evolution. From such evolutionary standpoint, (i) a neuron should be viewed as a functional rather than a genetic character, and (ii) any given neural system might be composed of different cell lineages with distinct origins and evolutionary histories. The identification of distant neural homologies or examples of convergent evolution among phyla will not only allow the reconstruction of neural systems' evolution but together with single-cell 'omic' approaches the proposed synthesis would lead to the 'Periodic System of Neurons' with predictive power for neuronal phenotypes and plasticity. Such phylogenetic framework of Neuronal Systematics might be a conceptual analog of the Periodic System of Chemical Elements.

# AN ANCIENT INTEGRATION IN A PLANT NLR IS MAINTAINED AS A *TRANS*-SPECIES POLYMORPHISM

Helen J Brabham[1], Inmaculada Hernández-Pinzón[1], Samuel Holden[1], Jennifer Lorang[2], Matthew J Moscou[1]

[1]The Sainsbury Laboratory, Norwich, United Kingdom, [2]Oregon State University, Department of Botany and Plant Pathology, Corvallis, OR

The origin of genetic diversity in a species is a product of novel mutation and alleles maintained during speciation. Identification of alleles that are derived from a common ancestor is critical for understanding the selective processes experienced during speciation events. Here, we report on the inter- and intraspecific diversity of plant immune receptors in the *Mla* locus across Poaceae species. These immune receptors belong to the class of cytoplasmic-nuclear localized nucleotide-binding leucine-rich repeat (NLR) proteins. Plant immune receptors are under constant selective pressure to maintain resistance to plant pathogens, and novel mutations leading to improved immune responses are selected. We discovered a fusion of the plant immune receptor RGH2 with a component of the exocyst complex (Exo70). Phylogenetic analysis of *Exo70* gene families from Poaceae species found that the integrated *Exo70* was derived from *Exo70F1*. To identify the origin of this gene fusion, we assembled leaf transcriptomes from diverse Poaceae species. Interspecific conservation in the *RGH2-Exo70F1* gene fusion was found in several Pooideae species, whereas *Brachypodium distachyon RGH2* is fused to a receptor-like kinase. To establish the time of gene fusion, we performed phylogenetic analysis using non-integrated and integrated *Exo70F1*. We found a diphyletic tree composed of non-integrated *Exo70F1* and integrated *Exo70F1*, with *Exo70F1* from *B. distachyon* and *B. stacei* forming an outgroup. This indicates that a single gene fusion event occurred after the speciation of Brachypodieae, but prior to Poeae-Triticeae radiation. These results, coupled with the observation of inter- and intraspecific variation in *RGH2*, demonstrate the maintenance of *RGH2-Exo70F1* as a *trans*-species polymorphism over 24 My.

# THE EXPANSION AND RECONFIGURATION OF THE GENCODE lncRNA CATALOGS.

Jonathan M Mudge, Jose M Gonzalez, Paul Flicek, Adam Frankish

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

On behalf of the GENCODE consortium.

GENCODE produces reference gene annotation for the human and mouse genomes; these annotations also form the Ensembl genesets for both species. Our remit is to annotate all gene features, and the GENCODE genesets thus include long non-coding RNAs (lncRNAs), small RNAs and pseudogenes alongside protein-coding genes. The number of transcript models for both genesets continue to rise: human v27 contains 200,401 models; mouse M16 133,849. However, our initial goal to describe the entire transcriptomes of these species is hindered by the deluge of RNA data produced by modern transcriptomics projects, including the GENCODE Capture Long Seq (CLS) pipeline. In fact, our extrapolations suggest that human GENCODE currently contains less than 10% of all introns found in short-read datasets. Even so, important question remain as to the biological relevance of this 'transcriptional complexity'. Here, we discuss our current drive to expand and reconfigure the GENCODE lncRNA catalogs in both species. These efforts leverage our bespoke capture-seq methodology based on PacBio sequencing, integrating existing short read datasets into a semi-automated workflow. We anticipate that our lncRNA transcript count will more than double in both species, while the number of existing models classed as 'incomplete' will significantly fall. In fact, our data frequently demonstrate interconnectivity between what were previously considered separate lncRNA genes. We also propose an entirely new system of lncRNA transcript classification based on Sequence Ontology, and are keen to receive community feedback on these ideas. A key goal is to capture the relationship between non-coding transcription and regulatory elements, especially enhancers. This work links into efforts to further integrate our gene annotation with the Ensembl Regulatory build, a strategy that includes the usage of '3D genome' datasets such as promoter-capture Hi-C to infer promoter-enhancer relationships. Altogether, these endeavors promise a major step forward in providing a consolidated view of genomic organization in Ensembl.

# CHARACTERIZING TRANSCRIPTOMIC VARIATION ACROSS HUMAN PHENOTYPES BY INTEGRATING RNASEQ DATA WITH HISTOPATHOLOGY IMAGES AND ANNOTATIONS

Manuel Muñoz-Aguirre[1,2,5], Marc Combalia[3], Ferran Reverter[4], Alessandra Breschi[6], Verónica Vilaplana[3], Ferran Marques[3], Roderic Guigó[1,5]

[1]Centre for Genomic Regulation, Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat Politécnica de Catalunya, Statistics and Operations Research, Barcelona, Spain, [3]Universitat Politécnica de Catalunya, Signal Theory and Communications, Barcelona, Spain, [4]Universitat de Barcelona, Statistics, Barcelona, Spain, [5]Universitat Pompeu Fabra, Experimental and Health Sciences, Barcelona, Spain, [6]Stanford University, Genetics, Stanford, CA

The Genotype Tissue Expression (GTEx) project has generated RNAseq data for more than 900 human individuals across 53 different tissues. Using this unique resource, we aim to relate transcriptomic variation with changes in human phenotypes. To this end, we have gathered phenotype data from two different perspectives.

First, a text processing pipeline has been developed in order to extract relevant terms from pathology review comments associated to histology images of GTEx tissue samples, observing that these annotations recapitulate the biology of several phenotypes. We detect alterations in the cellular composition deconvoluted from RNAseq data in samples tagged with histological phenotypes associated to diseases such as gynecomastia and atherosclerosis.

Second, we explored image processing approaches to extract phenotype data from the GTEx histology images. As a case study, we have developed a methodology for feature extraction in skin images from two different angles. (1) Handcrafted features: using traditional image processing techniques we segment three skin layers (stratum corneum, epidermis, and dermis) and characterize their width. (2) Deep learning approach: automatic feature extraction with a Multiple Instance Learning Neural Network. From both sets of learned features, we infer that the stratum corneum is the most important skin layer to differentiate between skin exposed and not exposed to sun, and that it is also the most related layer with global changes in gene expression.

With these integrative approaches that encompass RNAseq data, pathology text processing, and *image phenotypes*, we aim to open up the path to a deeper understanding of how a pathological condition affects the human body, through either systemic (affecting multiple tissues) or local effects (specific tissues).

# INCIDENCE OF UNIPARENTAL DISOMY IN 2 MILLION INDIVIDUALS FROM THE 23ANDME DATABASE

Priyanka Nakka[1,2], Kimberly McManus[3], 23andMe Research Team [3], Anne O'Donnell-Luria[4,5], Uta Francke[6], Sohini Ramachandran[1,2], Joanna Mountain[3], Fah Sathirapongsasuti[3]
[1]Brown University, Center for Computational Molecular Biology, Providence, RI, [2]Brown University, Ecology and Evolutionary Biology, Providence, RI, [3]23andMe, Inc., Mountain View, CA, [4]Boston Children's Hospital, Boston, MA, [5]Broad Institute of MIT and Harvard, Cambridge, MA, [6]Stanford University, Palo Alto, CA

Uniparental Disomy (UPD) is the inheritance of both homologs of a chromosome from one parent with no representative copy from the other. UPD can cause clinical phenotypes by disrupting parent-specific genomic imprinting and causing an imprinting disorder, or by unmasking recessive alleles in large blocks of homozygosity on the affected chromosome. Though over 3,300 clinical cases of UPD have been collected to date, rates of UPD and its subtypes (maternal UPD, paternal UPD, uniparental heterodisomy, uniparental isodisomy, and partial isodisomy) are very poorly characterized in the general population. To address this gap, we analyze instances of UPD in consented research participants from the personal genetics company 23andMe, Inc., whose database consists of SNP data from over 2 million individuals from across the United States at the time of analysis. We estimate identity-by-descent (IBD) between 290,927 parent-child duos and identify 49 instances of UPD, corresponding to an overall incidence rate of 1 in 4000 births for UPD in the general population. In the 23andMe database, UPD cases occur most frequently on chromosomes 16, 21 and 22. In contrast, existing clinical cases cluster on chromosomes 6, 7, 11, 14, and 15; these chromosomes contain groups of imprinted genes that cause disease phenotypes. Taken together, our results suggest that clinical UPD cases are not representative of the true chromosomal distribution of UPD in the general population. We also find that per-chromosomes rates of UPD in 23andMe are significantly correlated ($R^2 = 0.67$; $p = 2.9 \times 10^{-6}$) with per-chromosomes rates of aneuploidy in spontaneous abortions, suggesting that UPD cases on chromosomes 16, 21 and 22 may be caused by trisomy rescue and monosomy rescue following meiotic nondisjunction. Certain subtypes of UPD (complete isodisomy and some partial isodisomies) can also be detected using runs of homozygosity (ROH) in individuals who do not have genotyped relatives. We calculate ROH for 8 populations in the 23andme database and use a logistic regression-based model to classify cases of complete and partial isodisomy. Lastly, we test for associations between 5 categories of phenotypes (cognitive, personality, morphology, obesity and metabolic traits) and UPD status. In addition to phenotypes known to be associated with Prader-Willi Syndrome, which is caused by maternal UPD of chromosome 15, we find novel associations between UPD of chromosome 16, the most common UPD in the 23andme database, and shorter height (OR: -1.12 (-1.93, -0.31); $p = 0.007$) and increased risk for high plasma glucose (OR: 3.63 (1.03, 6.23); $p = 0.006$). Our study presents the first estimates of uniparental disomy rates in the general population, which likely occur due to trisomic or monosomic rescue, as well as previously unrecognized phenotypes associated with UPD of chromosome 16.

# STEPWISE EVOLUTION OF SEX-BIASED GENE EXPRESSION IN MAMMALIAN TISSUES

Sahin Naqvi[1,2], Alexander K Godfrey*[1,2], Jennifer F Hughes*[1], Mary L Goodheart[1,3], Richard N Mitchell[4,5], David C Page[1,2,3]

[1]Whitehead Institute, Cambridge, MA, [2]Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, [3]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, [4]Harvard Medical School, Department of Pathology, Boston, MA, [5]Brigham and Women's Hospital, Department of Pathology, Boston, MA

*These authors contributed equally

Sex differences are widespread in human health and disease, and are extensively studied in non-human, mammalian model organisms. However, the extent to which molecular sex differences are conserved across both tissues and mammalian species remains unclear. We conducted a twelve-tissue, five-species survey of sex bias in gene expression using both publically available (human) and newly generated (cynomolgus macaque, mouse, rat, and dog) RNA sequencing data. Using a flexible modeling approach, we found greater sharing of sex bias between species than between tissues, indicating evolutionary conservation of tissue-specific sex biases. Genes with a widely conserved sex bias show signatures of regulation by both sex hormones and sex chromosome complement. Lineage-specific gains or losses in sex bias are correlated with turnover of DNA sequence motifs corresponding to transcription factors that themselves show widely conserved sex bias. We propose a stepwise model for the evolution of sex bias in mammalian gene expression wherein the initial acquisition of sex bias across the genome, due to both hormones and sex chromosome complement, allowed for subsequent lineage-specific changes in sex-biased regulatory networks that lead to present-day differences in sex-biased gene expression between species.

# COMPREHENSIVE SURVEY OF LINE-1 TRANSCRIPTIONAL ACTIVITY IN HUMAN CELL LINES, HEALTHY TISSUE, AND TUMORS

Fabio Navarro[1,2], Jacob Hoops[1,2], Lauren Bellfy[4], Eliza Cerveira[4], Qihui Zhu[4], Chengsheng Zhang[4], Charles Lee[4,5], Mark B 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]The Jackson Laboratory for Genomic Medicine, Farmington, CT, [5]Ewha Womans University, Department of Life Sciences, Seoul, South Korea

Gauging Transposable Elements (TEs) activity in cells is a major and fundamental step to elucidate their impact in genomes. TEs constitute approximately half of the human genome and, due to their repetitive nature, they remain one of the most elusive genomic elements. Despite these challenges, there is building evidence that transposable sequences play a significant role in genomes by influencing gene regulation and creating variability across individuals and species. TEs were previously thought to be completely silent after embryogenesis, however, recent reports show that these elements are active in pathogenic and healthy cells such as neuro-progenitor cells.

Due to their high copy number and broad distribution across the genome, gauging TE expression is challenging for most existent RNA-seq pipelines. In particular, transcription quantification from TEs is highly affected by RNA fragments originating from pervasive transcription. We developed a new approach that distinguishes pervasive transcription signal from autonomous transcription using cross-mappability maps. Our new method, called TeXP, gauges TE transcription at subfamily transcription level with high concordance to ddPCR. We used RNA-seq experiments from GTEx and TCGA to evaluate the activity pattern of LINE-1s in human somatic tissues. We observe that the majority of the reads mapped to transposable elements are derived from pervasive transcription. However, removing this signal reveals consistent activity L1Hs in somatic tissue. Different from expected, we found that the adult human brain does not show high levels of L1Hs transcription, while other tissues such as Skin, Testis and Peripheral nerves have remarkable levels of L1Hs transcription. Moreover, we observed a direct correlation between LINE-1 activity and INDELs in tumors. We suggest that LINE-1 aborted retrotranposition can be a major player in tumor development contributing to the mutation load in tumors, specifically, creating INDELs close it ORF2p endonuclease target sites. All together our method and results demonstrate that LINE-1 is active during the whole development of human tissues and contribute to tissue genetic diversity by creating not only somatic insertions but also INDELs.

# IDENTIFICATION OF LOCI ASSOCIATED WITH EMBRYONIC AND FETAL LOSS IN HOLSTEIN HEIFERS

Holly L Neibergs[1], Jennifer N Kiser[1], Joseph Dalton[2], Joao G Moraes[3], Thomas E Spencer[3]

[1]Washington State University, Animal Sciences, Pullman, WA, [2]University of Idaho, Animal and Veterinary Science, Moscow, ID, [3]University of Missouri, Division of Animal Sciences, Columbia, MO

Holstein heifers are commonly bred at up to five consecutive estrus cycles to achieve a pregnancy. Although 90-100% of oocytes will be fertilized at each breeding, embryonic loss results in only 55-60% of heifers remaining pregnant at day 35 after the first breeding and an additional 7-10% of fetuses will be lost before term. Embryonic losses are even greater for heifers that require more than one breeding to establish a successful pregnancy. The objective of this study was to investigate if loci associated with embryonic loss, repeated embryonic loss and fetal loss were shared in Holstein heifers. One thousand twenty-two heifers were bred by artificial insemination during observed estrus for up to 5 consecutive estrus cycles, and pregnancy was determined via palpation on day 35 for heifers not returning to estrus. Heifers with fetal loss (n = 120) were identified by failure to calve or an observed miscarriage. Heifers were genotyped using the Illumina BovineHD BeadChip (777,962 SNPs; 902 heifers) or the GeneSeek GGP50 BeadChip (48,268 SNPs; 120 heifers). GeneSeek GGP50 BeadChip SNPs were imputed to 777,962 SNPs using Beagle 4.1 and a reference population of BovineHD genotypes from 4,838 Holsteins. A genome-wide associated analysis (GWAA) was conducted with a significance threshold of $P < 5 \times 10^{-8}$ and $P < 1 \times 10^{-5}$ to identify individual and shared associations, respectively. The GWAA identified 66 SNPs associated with one embryonic loss, 11 SNPs associated with presumed repeated embryonic losses (conceived at the fourth or fifth service) and 174 SNPs associated with fetal loss. Thirteen SNPs associated with embryonic and repeated embryonic loss were shared, 5 SNPs were shared with embryonic and fetal loss, 2 SNPs were shared with repeated embryonic and fetal loss and a SNP within *RGS6* was shared across all pregnancy losses. These results support that embryonic, repeated embryonic and fetal loss are mostly independent but do share a limited number of loci that could be selected for to reduce pregnancy loss in Holstein heifers.

# AN EXTENDED *MSPRIME* COALESCENT SIMULATION FRAMEWORK AVOIDS BIASES FROM LARGE SAMPLE SIZES, AND INCREASES PERFORMANCE WHEN SIMULATING LONG REGIONS.

Dominic Nelson[1], Jerome Kelleher[2], Gil McVean[2], Simon Gravel[1]

[1]McGill University, Human Genetics, Montreal, Canada, [2]University of Oxford, Big Data Institute, Oxford, United Kingdom

Coalescent simulators provide a very efficient way of generating sequence data for large sample sizes and long genomic regions. However, the scale at which simulations can now be performed has begun to go beyond the limits of the underlying model of coalescence, leading to unrealistic patterns of relatedness in the simulated samples. The ability to simulate at this scale is thanks largely to the recently-released *msprime* software package. While it re-implements exactly the algorithm of the ms software from 2002, it uses a new data structure, the succinct tree sequence, to improve the efficiency of the simulation by several orders of magnitude.

We present here a new hybrid simulator, an extension of the *msprime* simulation package, which generates sensible genealogies regardless of samples size or sequence length, and offers significant performance improvements over the original msprime simulations at large scales.

In order to achieve these improvements, we first implemented a discrete-time Wright Fisher simulation model within the *msprime* framework, using the same underlying data structures. Conceptually the modification to implement such a model is simple - allow for back-and-forth recombination, and draw events on a per-generation basis, rather than drawing the time to the next event. Integrating such a model into the *msprime* framework allows for highly-efficient hybrid simulations to be performed, which consist of Wright Fisher simulations to avoid biases in the recent past due to large sample sizes or long regions, before switching to much faster coalescent simulations when the assumptions of the underlying model are once again satisfied.

A somewhat surprising result is that for long regions, this hybrid approach is significantly faster than coalescent simulations alone. This is due to the fact that the number of new lineages created by recombination is limited to two per generation in the Wright Fisher (one for each parent), whereas the coalescent model creates a new lineage for each recombination event, which will happen many times per generation for long regions. This means that recent generations can be simulated more quickly (and in this case more accurately) with the Wright Fisher model.

To emphasize the importance of Wright Fisher genealogies in the recent past, we then use our hybrid simulator to show significant distortions in patterns of IBD sharing in large-scale simulations under the full coalescent, and how these can be addressed with our new simulation strategy while improving computational efficiency as well.

# DECODING PATIENT GENOMES THROUGH THE HIERARCHICAL PATHWAY ARCHITECTURE OF THE CANCER CELL

Trey Ideker

University of California-San Diego, La Jolla, CA

Although cancer is governed by complex molecular systems, the composition and modular organization of these systems remains poorly understood. I will describe efforts by the Cancer Cell Map Initiative (CCMI) to generate large protein interaction maps of tumor cells, which we are integrating with existing molecular and structural data to assemble a comprehensive multiscale map of cancer cell biology. The current draft map contains a hierarchy of ~250 systems covering both known hallmarks and unexpected components. Integration with tumor mutation profiles suggests that the modules under strongest selective pressure in cancer are often not genes, but extend upwards in scale to include protein complexes, broad cellular processes, and organelles. This observation suggests a classification of cancer based on convergence of mutations at key bottlenecks in the hierarchy of cellular systems.

# STOCHASTIC OR DETERMINISTIC? DECODING THE REGULATORY ROLE AND EPICLONAL DYNAMICS OF DNA METHYLATION IN 1482 BREAST TUMOURS.

Rajbir N Batra[1], Ana V Tufedgzic[2], Suet-Feung Chin[1], Ankita S Batra[1], Maurizio Callari[1], Oscar Rueda[1], Aviezer Lifshitz[3], Amos Tanay[3], Carlos Caldas[1]

[1]University of Cambridge, CRUK - Cambridge Institute, Cambridge, United Kingdom, [2]The Francis Crick Institute, Genome Integrity, London, United Kingdom, [3]Weizmann Institute of Science, Applied Mathematics, Rehovot, Israel

**INTRODUCTION:** Breast cancer is a clinically and molecularly heterogeneous disease displaying distinct therapeutic responses. Although recent studies have explored the genomic and transcriptomic landscapes, the epigenetic architecture has received less attention.
**METHODS:** An optimised Reduced Representation Bisulfite Sequencing (RRBS) protocol was used to profile the methylomes of 1482 primary breast tumours (and 237 matched normal tissues).
**RESULTS AND DISCUSSION:** Noticeable epigenetic drift (both gain and loss of homogeneous DNA methylation patterns) was observed in breast tumours when compared to normal tissues, with markedly higher differences in late replicating genomic regions. The extent of epigenetic drift was also found to be highly heterogeneous between the breast tumours and was sharply correlated with the tumour's mitotic index, indicating that epigenetic drift is largely a consequence of the accumulation of passive cell division related errors.
A novel algorithm called DMARC (Directed Methylation Altered Regions in Cancer) was developed that utilised the tumour-specific drift rates to discriminate between methylation alterations attained as a consequence of stochastic cell division errors (background) and those reflecting a more deterministic biological process (directed). Directed methylation alterations were significantly enriched for gene expression changes in breast cancer, compared to background alterations. By integrating with existing copy number and mutational profiles for these tumours, mutually exclusive patterns between DNA methylation and genomic aberrations were detected. This led to the identification of potential tumour suppressors and oncogenes, and revealed the deterministic nature of these epigenetic alterations.
Finally, intra-tumoural methylation content was analysed to identify tumours associated with low epigenetic polymorphism implying a deterministic process versus those with high polymorphism reflecting a stochastic process. Intra-tumoural heterogeneity as well as the extent of epiallelic burden were found to be prognostic, and revealed a distinction in the role of epiclonal dynamics in different breast cancer subtypes.
**CONCLUSION:** The data presented here constitutes the largest sequencing-based cancer methylome study. Stochastic and deterministic DNA methylation patterns have been identified with distinct roles in tumour evolution.

# THE CONTRIBUTION OF *DE NOVO* MUTATIONS IN ULTRA-CONSERVED REGULATORY ELEMENTS TO NEURODEVELOPMENTAL DISORDERS AND AUTISM SPECTRUM DISORDER

Patrick J Short[1], Sebastian Gerety[1], Holly Ironfield[1], Giuseppe Gallone[1], Caroline F Wright[2], Helen V Firth[3], David R FitzPatrick[4], Jeffrey C Barrett[1], Matthew E Hurles[1]

[1]Wellcome Trust Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, United Kingdom, [3]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom, [4]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Edinburgh, United Kingdom

We previously sequenced more than 6,000 ultra-conserved non-coding elements and enhancers covering 5 megabase of sequence in almost 8,000 families with a child with a severe developmental disorder and identified a significant enrichment for *de novo* mutations in ultra-conserved non-coding elements. We estimated that mutations in these elements contribute to up to 3% of severe developmental disorders. We also identified a significant two-fold enrichment of elements recurrently mutated in independent families. We have expanded this analysis to include nearly 10,000 individuals with severe developmental disorders, more than 2,000 families with autism spectrum disorder, and 2,000 unaffected parent-child trios. This expanded analysis identifies more than seventy recurrently mutated elements non-coding elements, which we show act primarily as enhancers or are involved in alternative splicing. We also detect a significant enrichment of *de novo* mutations in clusters of ultra-conserved elements, which have been shown to be enriched near important developmental genes.

Population genetics analyses based on deep whole genome sequencing data suggests that in regulatory elements, nucleotide-level conservation metrics may be more predictive of purifying selection, and hence likely disease relevance, than element-wide measures of evolutionary conservation. However, lack of understanding of the 'enhancer code' to distinguish benign from damaging variation has made determining the precise impact of mutations in these elements a challenge. To address this challenge, we synthesised more than 50,000 putative regulatory sequences to test in a series of massively parallel reporter assay (MPRA) experiments and intend to present preliminary data from these MPRA experiments.

# PREDICTABLE AND PRECISE TEMPLATE-FREE CRISPR REPAIR OF DISEASE MUTATIONS

Max W Shen[1], Mandana Arbab[2], Jonathan Hsu[3], Daniel Worstell[4], Olga Krabbe[4], Christopher Cassa[4], David Liu[2], Richard I Sherwood[4], David K Gifford[1]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, MA, [4]Brigham and Women's Hospital and Harvard Medical School, Division of Genetics, Department of Medicine, Cambridge, MA

We find that template-free DNA repair of Cas9-cleaved and Cpf1-cleaved DNA produces a predictable set of repair genotypes that can result in the gain-of-function repair of human disease mutations. Contrary to the assumption that end-joining following double-strand breaks is random and difficult to harness for applications beyond gene disruption, here we show that template-free end-joining repair of DNA cleaved by CRISPR-associated nucleases produces a predictable set of repair genotypes. We constructed a library of 2000 guide RNAs paired with target DNA sites, integrated them into mouse and human genomes, applied Cas9, and performed high-throughput sequencing of repair genotypes. Data from this assay are consistent with results from 98 endogenous loci. Building upon prior work, we show that the majority of repair genotypes in cells with saturated exposure to both CRISPR-Cas9 and Cpf1 are deletions associated with sequence microhomology. Using 1,588 sequence contexts from our data, we trained CRISPR-Texture, a machine learning method that accurately predicted the frequencies of template-free Cas9-mediated microhomology-associated deletions as well as 1 bp insertions. On 282 held-out sequence contexts, CRISPR-Texture predicted frameshift rates more accurately than published methods and accurately predicted the statistical entropy of repair product distributions. Applied to the human genome, CRISPR-Texture identified an appreciable fraction of Cas9 target sites supporting high-precision repair distributions that are dominated by a single genotype. Further, we find that a class of human disease-associated micro-duplication mutations can be repaired to wildtype at high frequency by template-free Cas9 nuclease editing and used our assay to validate hundreds of such alleles. We also validated template-free Cas9 nuclease-mediated rescue of pathogenic LDLR alleles to wildtype phenotype in cellular models. This work establishes a strategy for predicting the outcomes of template-free end-joining and demonstrates that CRISPR editing can also mediate efficient gain-of-function editing at certain disease alleles without homology-directed repair.

# CANCER RISK AMONG CHILDREN WITH NON-CHROMOSOMAL BIRTH DEFECTS IN THE GENETIC OVERLAP BETWEEN CONGENITAL ANOMALIES AND CANCER IN KIDS (GOBACK) STUDY: A POPULATION-BASED ASSESSMENT IN 10 MILLION LIVE BIRTHS

J.M. Schraw[1], T.A. Desrosiers [2], W.N. Nembhard [3], G. Copeland[4], R.E. Meyer[5], A.B. Brown[6], T.M. Chambers[6], H.E. Danysh[6], S. Sisoudiya[7], C. Luo[3], A. Mian[3], M.E. Scheurer[6], A. Sabo[7], S.E. Plon[6], Lupo P.J.[6]

[1]Baylor College of Medicine, Dept. of Medicine, Houston, TX, [2]University of North Carolina, Dept. of Epidemiology, Chapel Hill, NC, [3]University of Arkansas for Medical Sciences and Arkansas Children's Research Institute, Dept. of Epidemiology, Little Rock, AR, [4]Michigan Department of Health and Human Services, Lansing, MI, [5]North Carolina Division of Public Health, Birth Defects Monitoring Program, Raleigh, NC, [6]Baylor College of Medicine, Dept. of Pediatrics, Houston, TX, [7]Baylor College of Medicine, Dept. of Molecular and Human Genetics, Houston, TX

**Purpose**: Little is known about associations between specific non-chromosomal structural birth defects and specific childhood cancers. In the Genetic Overlap Between Congenital Anomalies and Cancer in Kids (GOBACK) Study we have a two-pronged approach to this question: (1) A population-based registry linkage study to identify novel birth defect-childhood cancer (BD-CC) associations, and (2) a family-based sequencing study to evaluate underlying genetic mechanisms.
**Methods**: A retrospective cohort of >10 million children, was established by pooling statewide registry data from four U.S. states (Texas, Michigan, North Carolina, and Arkansas) from 1992-2013. We used Cox proportional hazards models to evaluate associations between 60 birth defects and 31 childhood cancers (≥5 comorbid cases) with a hazard ratio (HR) and 95% confidence interval (CI). The false discovery rate (FDR) was computed via the Benjamini-Hochberg procedure. Based on these associations, subjects and their parents are being recruited with a focus on proband/parent trios. Salivary samples are undergoing whole-genome sequencing (WGS-40X coverage). Variants are being called using Platypus and multiple callers to identify structural rearrangements.
**Results**: Across the four states we identified 517,548 children with non-chromosomal structural birth defects and 14,774 children with cancer. The risk of any cancer was increased among children with any non-chromosomal structural defect compared to children without any birth defect (HR=2.6, 95% CI 2.4-2.7) particularly at young ages. We identified 496 of 606 BD-CC associations with significantly elevated HRs at a 5% FDR. Notably, hepatoblastoma, astrocytoma, ependymoma, and extracranial germ cell tumors were each strongly associated with several birth defects. For example, the risk of hepatoblastoma was increased among children with craniosynostosis (HR=15.4, 95% CI 7.6-31.3). Elevated risk of extracranial germ cell tumors was observed among children with CNS defects (HR=22.5, 95% CI 10.9-46.4). To date, 65 families are enrolled in GOBACK and WGS completed on 16 trios. From our initial analysis of structural rearrangements, we have identified a de novo deletion in *USP9X* in a female patient with acute lymphoblastic leukemia (ALL) and several major congenital anomalies. De novo loss-of-function mutations in *USP9X* have recently been implicated in a related female-specific syndrome with overlapping features to our patient including some patients with malignancy. We are now working to further identify the role of *USP9X* related disorders with cancer susceptibility.
**Conclusions**: By pooling registry data across four U.S. states, we find that children with non-chromosomal birth defects have a significantly elevated risk of several childhood cancers and we report on several novel associations. Early evaluation of the WGS data are beginning to reveal novel molecular mechanism of these new childhood cancer syndromes.

# AFFAIRS OF THE HEART: HOW eccDNA-MEDIATED SCARS IN THE TTN GENE MAY CONTRIBUTE TO MYOFIBER DIVERSITY

Massa J Shoura[1], Victoria N Parikh[2], Alexandra Dainis[2], Stephen D Levene[3], Euan A Ashley[2,4], Andrew Z Fire[1,4]

[1]Stanford University, Pathology, Stanford, CA, [2]Stanford University, Cardiovascular Medicine, Stanford, CA, [3]University of Texas at Dallas, Bioengineering, Richardson, TX, [4]Stanford University, Genetics, Stanford, CA

Truncating variants in TTN segregate with disease in familial dilated cardiomyopathy and are also associated with decreased ventricular wall thickness in the general population. Studies of variant TTN transcriptional isoforms have attributed the diversity of Titin protein structures to germline sequence variants, alternative RNA splicing, and potentially the translation of circular RNA transcripts. Diversity in the Titin protein pool is thought to be clinically significant in that shorter isoforms confer increased stiffness. Here we present evidence for a novel diversification mechanism involving a class of tissue-specific rearrangements at the DNA level. We show that a subset of unique extra-chromosomal-circular DNA molecules (eccDNAs), originate from the Titin locus in heart tissues and reprogrammed cardiomyocytes. Our data suggest a model for somatic TTN allelic diversity in which these circular DNA molecules are products of genomic recombination resulting from lineage-specific somatic alterations in gene length and sequence. Based on known correlations between TTN truncation and isoform length with cardiac muscle stiffness, our findings indicate a possible new mechanism for regulation of myocardial stiffness, an important feature of cardiomyopathy.

# TOWARDS MAPPING FUNCTIONAL CANCER GENOME ATLASES

Sidi Chen

Yale University, Genetics & Systems Biology Institute, New Haven, CT

International consortia such as TCGA mapped a comprehensive catalog of molecular alterations in the cancer genome. However, for many novel molecular alterations in patients, it is still unclear which of them, or their combinations, are necessary, sufficient, additive, antagonistic or synergistic in causing cancer and subsequent tumor progression. With genome editing, it is now feasible to systematically and quantitatively assess the contribution of each gene and various combinations to cancer evolution directly in vivo. This leads to the concept of mapping the functional cancer genome atlas (FCGA) for each cancer type. I will discuss the initial work towards mapping of most frequently mutated genes in cancer genomes in two highly lethal cancer types.

Our initial effort to map a provisional FCGA started from hepatocellular carcinoma (HCC). We developed a direct in vivo AAV-CRISPR screen approach, which generated highly complex, yet genetically defined, autochthonous liver tumors, for which we devised a novel readout strategy by directly sequencing variants at predicted sgRNA cut sites using molecular inversion probe sequencing (MIPS). The combination of AAV-CRISPR autochthonous mutagenesis and MIPS readout illuminated the mutational landscape of tumors, demonstrating quantitative maps of a large collection of variants. This screen revealed a functional map of drivers in liver tumorigenesis in fully immunocompetent mice, identifying novel functional tumor suppressors such as Setd2, B2m, Kansl1, Arid2, Kdm5c, Zc3h13 and Cic.

We extended this approach for mapping of FCGA for glioblastoma (GBM), a disease with single digit five-year survival. We performed an AAV mediated autochthonous CRISPR screen in GBM. Stereotaxic delivery of an AAV library targeting genes commonly mutated in human cancers into the brains of conditional Cas9 mice resulted in tumors that recapitulate human GBM. Capture sequencing revealed diverse mutational profiles across tumors. Notably, across all genes represented in the mTSG library, the experimental mutational frequencies observed in these AAV-CRISPR pooled mutagenized mouse tumors correlated with the clinical mutational frequencies in human patients tumors in two independent patient cohorts. Co-mutation analysis identified co-occurring driver combinations such as Mll2, B2m-Nf1, Mll3-Nf1 and Zc3h13-Rb1, which were subsequently validated using AAV minipools. This study provides a functional landscape of gliomagenesis suppressors in vivo.

The concept of FCGA and the AAV-CRISPR-MIPS approach can be applied to virtually the whole genome in any cancer types, or be focused down to the level of personalized mutations in each patient's cancer. Such studies can be performed in combination with many pre-clinical or co-clinical settings, providing new and powerful avenues for therapeutic discovery.

# SIGNATURES OF COMPLEX STRUCTURAL VARIATION ACROSS THOUSANDS OF CANCER WHOLE GENOMES

Kevin M Hadi[1,2], Xiaotong Yao[1,2], Evan Biederstedt[1,2], Mahmoud Ghandi[3], Marcin Imielinski[1,2]

[1]Weill Cornell Medicine, Pathology and Laboratory Medicine, New York, NY, [2]New York Genome Center, Cancer Program, New York, NY, [3]Broad Institute, Cancer Program, Cambridge, MA

Though the signatures of single nucleotide variants (SNVs) in cancer genomes are well defined, the mutational processes that give rise to somatic structural variation (copy number and rearrangements) patterns are largely unknown. The key challenge is conceptual: while substitutions are easily counted and categorized on the basis of their local nucleotide context, the definition of a somatic structural variant events are usually elusive. The simple taxonomy of structural variation inherited from constitutional genetics (duplications, deletions, translocations, and inversions) fails to unambiguously describe the many cancer loci with two or more rearrangement junctions in proximity. Furthermore, state-of-the-art approaches for somatic structural variant signature detection consider rearrangement junctions in isolation, thus ignoring the complex cis structure of highly rearranged loci or the close coupling of copy number and rearrangement. Such features are likely critical for identifying mutational processes generating complex somatic variants.

We have applied a novel algorithm for cancer graph genome inference (JaBbA, https://github.com/mskilab/JaBbA) to characterize mutational processes governing somatic structural variation across nearly 3000 whole genome sequences (WGS) from the Pan Cancer Analysis of Whole Genomes (PCAWG) and the Cancer Cell Line Encyclopedia (CCLE) projects. Through analysis of linear alleles (i.e. walks) in cancer graph genomes, we identify over 100 walk motifs that are statistically enriched above background. We also develop graph-based criteria to detect complex rearrangement events like chromoplexy, chromothripsis, and breakage-fusion-bridge cycles across 30 cancer types. Supplementation of standard junction taxonomy with these integrative rearrangement event definitions allows detection of novel mutational processes through Latent Dirichlet Allocation of event counts. Our results demonstrate enrichment of lineage-specific mutational processes (e.g. AID in lymphoma and CLL, tandem duplicator in ovarian / breast cancer) and correlation of mutational process burden with somatic driver genotype (e.g. SPOP mutations). We identify novel correlations of SNV and structural variant mutational processes, including a novel correlation of APOBEC-driven GC-strand coordinated clusters with specific rearrangement processes in sarcoma. Our results demonstrate a fundamental similarity between rearrangement signatures observed in CCLE cell lines and PCAWG primary cancer WGS. The proposed signatures provide a novel basis with which to categorize cancers, and may serve as novel biomarkers of specific mutagen exposures or therapeutic sensitivity to DNA damage-inducing anticancer drugs (e.g. chemotherapy, radiotherapy, PARP inhibitors).

# METHODS FOR THE JOINT ANALYSIS OF HIGH-DIMENSIONAL TRAITS AND SAMPLE SUBSTRUCTURE IN HUMAN COHORTS

Oliver Stegle[1,2]

[1]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany, [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Generating “omic” profiles with multiple assays and data types on the same set of biological samples, is a fundamental experimental design pattern in biomedical research and basic biology. In addition to classical omics, typical data modalities that are being collected in human cohorts increasingly include environmental variables and other types of side information. In this talk I will describe a statistical framework for modelling these data that builds on generalisations of factor analysis. These approaches are applicable to a broad range data types and allow for elucidating shared and dataset specific axes of variation. In a second step we will show applications of these principles to model and account for sample substructure due to different environmental exposures in human populations. We illustrate these methods in an application to data from UK biobank, where we report interactions between genetic effects on body mass index and environmental exposure in the population.

# DYNAMIC EFFECTS OF GENETIC VARIATION ON GENE EXPRESSION DURING CELLULAR DIFFERENTIATION.

Benjamin Strober*[1], Reem Elorbany*[2], Katherine Rhodes*[2], Nirmal Krishnan[3], David Knowles[4], Jonathan Pritchard[4,5], Yoav Gilad^[2,6], Alexis Battle^[1,3]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]University of Chicago, Human Genetics, Chicago, IL, [3]Johns Hopkins University, Computer Science, Baltimore, MD, [4]Stanford University, Genetics, Stanford, CA, [5]Stanford University, Biology, Stanford, CA, [6]University of Chicago, Medicine, Chicago, IL

*,^ equal contribution

A comprehensive model of the genetic regulation of gene expression would provide important insights into human development and complex disease. This remains challenging, in part because the effects of genetic variation have been shown to vary considerably between cell types and conditions, and can have dynamic effects during processes such as differentiation. The vast majority of studies have focused on adult tissue sampled at a single time point, but dynamic effects likely also contribute to the genetics of disease. To investigate cell-type-specific and dynamic effects of genetic variation, we generated time-series data capturing the differentiation progression from induced pluripotent stem cells (iPSCs) to cardiomyocytes in 19 human Yoruba HapMap cell lines. For each cell line, RNA was extracted and sequenced every 24 hours at 16 time points to capture the entire differentiation process. Combined with existing genotype information, these data provide a novel resource to understand how genetic regulation changes throughout cardiomyocyte differentiation, and to identify genetic variants that alter the temporal pattern of gene expression. First, we identified cis-eQTLs in each time step independently; identifying a median of 168 significant genes (FDR < .05) per time step. We found that eQTL summary statistics from proximal time points are more strongly correlated than distal time points, indicating that cardiomyocyte differentiation is contributing to genetic regulatory changes. In fact, temporal structure was the dominant axis of variation in both gene expression and eQTL effects. Next, we developed a novel probabilistic model, FALCON, to directly identify cis-eQTLs whose effects change during differentiation. FALCON gains power by modeling temporal ordering of samples, along with both total and allele-specific expression. This approach yielded 381 significant genes with “dynamic eQTLs” (FDR < .05). Finally, we characterized the variants and genes that underlie dynamic eQTLs by investigating genomic and epigenomic context as well as gene function. Overall, these data shed light on how the genetic regulation of gene expression changes during cardiomyocyte differentiation and we develop novel statistical tools that can be used to identify dynamic eQTLs from any time series RNA-seq data.

# POPULATION SCALE SINGLE CELL SEQUENCING TO REVEAL CONTEXT SPECIFIC EFFECTS OF SLE VARIANTS

Meena Subramaniam[1], Lenka Maliskova[1], Nadav Rappoport[2], Cristina Lanata[3], Lindsey Criswell[3], Noah Zaitlen[4], Jimmie Ye[1]

[1]Institute for Human Genetics, UCSF, San Francisco, CA, [2]Institute for Computational Health Sciences, UCSF, San Francisco, CA, [3]Rosalind Russell/Ephraim P Engleman Rheumatology Research Center, Department of Medicine, UCSF, San Francisco, CA, [4]Lung Biology Center, Department of Medicine, UCSF, San Francisco, CA

Recent advances in microfluidics and next generation sequencing have enabled the profiling of thousands of cells in a single experiment. Here we apply multiplexed single cell RNA-Seq (Mux-seq) with our previously developed method *demuxlet* to profile over 1 million PBMCs from 120 patients with systemic lupus erythematosus (SLE) and 21 healthy controls. Across all individuals, we detected 12 major cell types, including 3 subgroups of CD4+ T cells, 4 subgroups of CD14+ Monocytes, CD8+ T cells, Dendritic Cells, Natural Killer cells, B cells, and Megakaryocytes. We identified an increased proportion of CD14+ Monocytes in the SLE patients ($\beta$=0.11, $p<6.5x10^{-6}$), and a decreased proportion of CD4+ T cells ($\beta$=-0.14, $p<9.05x10^{-10}$). We validated the increased monocyte proportion using Complete Blood Counts from over 2,000 controls and cases in the UCSF Electronic Health Record database, with a similar effect ($\beta$=0.10, $p<4.78x10^{-53}$). We also showed that monocyte proportion is positively correlated with type I interferon response, a known indicator of SLE activity ($p<1.77x10^{-6}$). To further assess variability within the SLE patients and identify context-specific effects of gene expression, we performed a cell type specific eQTL analysis across all cell types. We detected 237 eQTLs that were present in at least one cell type, and found high enrichment for previously published eQTLs. Among the eQTLs were *LYZ* ($p<3.56x10^{-23}$) and *LILRA3* ($p<1.89x10^{-27}$), with high effects in monocytes. Single cell profiling allows for the estimation of the variance of gene expression across single cells, and we can assess the variability in this parameter across individuals. Within our cohort, we detected 39 variance-QTLs in B cells 38 variance-QTLs in CD4+ T cells. Notably, we found *TYK2*, a region of susceptibility for SLE that influences several cytokine signaling pathways, as a variance-QTL in B cells ($p<1.77x10^{-10}$). We also identified the transcription factor *YY1*, a regulator of Th2 cytokine production and GWAS locus, to be a variance-QTL in CD4+ T cells ($p<4.13x10^{-7}$). These results demonstrate that Mux-seq can be used not only to ascribe novel mechanisms to GWAS loci, but also to identify cell type specific context for disease heterogeneity. These factors could lead to better diagnostics as well as targeted therapies for SLE patients.

# INFORMATION THEORY ANALYSIS OF ATAC-SEQ DATA PREDICTS LOCAL CHROMATIN KINETICS AND REVEALS NOVEL ASPECTS OF GENE REGULATION AND GENOME ORGANIZATION.

Ricardo D'Oliveira Albanus[1], John Hensley[1], Yasuhiro Kyono[1,2], Jacob Kitzman[1,2], Stephen Parker[1,2]

[1]University of Michigan, Department of Computational Medicine & Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Department of Human Genetics, Ann Arbor, MI

Understanding how non-coding genetic variants influence human traits and diseases is a central problem in biomedical research. Identification of the target regulatory elements and upstream transcription factors (TFs) that mediate these genetic effects are critical steps in dissecting the causal biological pathways. Chromatin accessibility assays, like ATAC-seq, reveal diverse information about these *in vivo* regulatory landscapes. For example, regional properties of chromatin architecture are captured in ATAC-seq fragment lengths (short TF-flanking or longer nucleosomal-sized units). We reasoned that the positional information encoded in the pattern of fragment lengths surrounding specific TFs is a reflection of how they interact with DNA and nucleosomes. Here, we introduce a novel information-theory approach to quantify such positional fragment length patterns in ATAC-seq data. This method allows us to quantify the impact of a TF on local chromatin architecture through nucleosome phasing. We show that our measure of local information content is a proxy of TF-DNA binding kinetics, as measured by previous fluorescence recovery after photobleaching (FRAP) studies. In one example, we predict that CTCF co-bound with cohesin has much slower binding rates compared to CTCF bound without cohesin, which supports a recent FRAP / single molecule tracking study. We validate these findings using a modified ATAC-seq protocol where we introduce random shearing to remove fragment size information. Next, we show that footprinting-based methods for inferring TF binding are highly sensitive to our kinetics measure and perform poorly on most TFs, which have fast inferred binding rates. To improve footprinting performance for such TFs, we develop a prediction approach based on TF-specific negative binomial models of the local ATAC signal and the number of neighboring binding sites. We show that this approach is more robust to predicted TF-DNA binding kinetics and outperforms current methods. Finally, using a motif-agnostic *k-mer* scanning approach and our new local chromatin architecture information content measures, we found that genome-wide chromatin architecture information is significantly associated with diverse aspects of gene regulation. We report both in human islet tissue samples and lymphoblastoid cells that highly organized regions are enriched for active enhancers, while disorganized regions are enriched for expression quantitative trait loci (eQTLs) and active promoters. We find that ATAC-seq allelic imbalance is significantly shifted towards alleles that are predicted to increase local chromatin organization, providing a framework to understand basic genetic principles underlying chromatin architecture. Collectively, our work provides novel insights on how regulatory information is encoded in the human genome.

# USING ALLELIC EXPRESSION DATA FOR STUDYING RARE DISEASE BIOLOGY

Pejman Mohammadi[1,2,3,4], Stephane Castel[3,4], Beryl Cummings[5,6], Daniel MacArthur[5,6], Tuuli Lappalainen[3,4]

[1]The Scripps Translational Science Institute, -, La Jolla, CA, [2]The Scripps Research Institute, Department of Integrative Structural and Computational Biology, La Jolla, CA, [3]New York Genome Center, -, New York, NY, [4]Columbia University, Department of Systems Biology, New York, NY, [5]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [6]Broad Institute of MIT and Harvard, -, Cambridge, MA

Despite the abundance of data from sequencing-based assays from individuals with severe disease, discovering potential pathogenic regulatory variants from these data has been challenging. Allelic expression (AE) data is a potentially valuable resource to address this problem, specifically due to the fact AE data captures the net effect of all cis-regulatory variants in an individual and it is minimally obscured by other confounding factors. However, widespread use of this valuable data source has been hampered by its conceptual complexity and lack of theoretical frameworks to enable advanced analyses.

In this work we describe a mathematical model of cis-regulation. Using this model we demonstrate that AE data distribution emerges as a constrained form of Binomial Logistic Normal (BLN) distribution function. We further show how this model can be used for quantifying regulatory variation in human population using our new method called ANalysis of Expression VAriance (ANEVA). Building upon this we provide methods for testing extreme allelic imbalance and dosage outliers.

We apply our model to estimate per-gene natural variation in cis-regulation for 13,399 genes in the Genotype-Tissue Expression (GTEx) project v6 data. We then use these estimates to identify genes with potentially pathogenic regulatory mutations in rare genetic disease data. Applying dosage outlier test to AE data from 26 patients with rare muscle disorders from Cummings et al. 2017, we found on average 24 such dosage outlier genes at 5% FDR. The genes which were outliers in at least two patients were ~2.5 folds enriched for known neuromuscular disease genes ($p=10^{-6}$). Furthermore, known neuromuscular disease genes appeared in the top three dosage outliers in 11 out of the 26 samples. This demonstrates how the modeling of genetic regulatory variation in the general population can be used to interpret potential disease-causing variation in patient transcriptomes.

# VARIATION GRAPHS FOR EFFICIENT UNBIASED PANGENOMIC SEQUENCE INTERPRETATION

Erik Garrison[1], Jouni Sirén[2], Adam M Novak[2], Glenn Hickey[2], Jordan M Eizenga[2], Eric T Dawson[1,3], Eppie Jones[5,6], William Jones[1], Michael F Lin[4], Benedict Paten[2], Richard Durbin[1,5]

[1]Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [2]UC Santa Cruz, Santa Cruz, CA, [3]National Cancer Institute, Rockville, MD, [4]DNAnexus, Mountain View, CA, [5]University of Cambridge, Cambridge, United Kingdom, [6]Trinity College Dublin, Dublin, Ireland

Variation graphs are bidirected DNA sequence graphs that compactly represent genetic variation, including large scale structural variation such as inversions and duplications. They extend the concept of linear reference genomes, which are fundamentally limited in that they represent only one version of each locus, whereas the population may contain multiple variants. Equivalent structures are produced by *de novo* genome assemblers. When a reference represents an individual's genome poorly, either because it is missing significant variation or because it is incompletely assembled, it can impact read mapping and introduce bias. Here we present vg, a toolkit of computational methods for creating, manipulating, and utilizing variation graphs as references at the scale of the human genome. vg provides an efficient approach to mapping reads onto arbitrary variation graphs using generalized compressed suffix arrays, with improved accuracy over alignment to a linear reference, creating data structures to support downstream variant calling and genotyping.

To demonstrate the utility of vg, we apply it to a number of resequencing contexts where extensive work has been required to correct for reference bias, or where reference genomes are not available. We present a number of our findings: Alignment to a pangenome graph made from the human 1000 Genomes Project data resolves reference bias at known indels, regardless of allele length. A graph constructed of seven whole *de novo* assemblies of *S. cerevisiae* strains provides a better match for sequencing data from other strains than the linear reference. Aligning ChIP-seq reads from ENCODE to a human pangenome yields a dramatic reduction in reference bias at known alleles. Studies of real and simulated ancient DNA show that our method resolves reference bias at known alleles even in the context of high rates of damage and short reads. While linear alignment methods fail to match a large portion of sequencing data to the contigs of an Arctic viral metagenome assembly, vg is able to achieve near perfect alignment of the same data to the assembly graph from which the contigs were extracted. Through comparison with standard tools we show that vg is capable of accurately aligning long reads from single molecule sequencing, and furthermore that it can align them against any kind of sequence graph.

# TISSUE-SPECIFIC ENHANCER AND PROMOTER EVOLUTION IN MAMMALS

Maša Roller*[1], Ericca Stamper*[1,2], Louise Harewood[2], Diego Villar[2], Aisling Redmond[2], Duncan T Odom[2,3], Paul Flicek[1,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom, [2]University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom, [3]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

*Equal contributors

Gene expression is established through spatiotemporal coordination of regulatory elements including enhancers and promoters. The evolution of enhancers and promoters has been widely studied within single mammalian cell types or tissues and has established, among other results, that enhancers evolve more rapidly than promoters. Thus far, however, the evolution of regulatory elements across mammalian tissues has largely been limited to comparisons between human and mouse. To more comprehensively understand the tissue specific components of regulatory evolution across the mammalian lineage, we have established extensive regulatory profiles of four tissues in ten species of mammals. Specifically, we mapped *in vivo* occupancy of key histone modifications histone 3 lysine 27 acetylation (H3K27ac), histone 3 lysine 4 trimethylation (H3K4me3) and histone 3 lysine 4 monomethylation (H3K4me1) as a means to annotate active promoters and enhancers. The four tissues - liver, muscle, brain and testes - display the diverse regulatory patterns that are active in somatic and reproductive tissues of different functions and developmental origins.

The somatic tissues showed a consistent regulatory profile – the majority of somatic promoters are active across multiple tissues, while enhancers mostly have tissue specific activity. The testes had a markedly different, but consistent across species, regulatory profile that is in line with increased transcriptional activity of reproductive tissue. Within this framework we explored the relationship between tissue specific regulatory activity and evolutionary turnover. We confirmed previous findings that promoters on average have a slower evolutionary turnover than enhancers. However, tissue-specific regulatory elements turnover more quickly than those active across tissues. We were also able to stratify enhancers of different types, as evidenced by distinct histone enrichment profiles, and discovered that they evolve at different rates. These results provide important insights into the evolution of regulatory elements in mammals, especially elements with tissue-specific activity.

# EXPANDING GEMINI TO ANNOTATE AND PRIORITIZE SUBCLONAL MUTATIONS IN HETEROGENEOUS TUMORS

Thomas J Nicholas[1,2], Brent S Pedersen[1,2], Yi Qiao[1,2], Xiaomeng Huang[1,2], Gabor Marth[1,2], Aaron R Quinlan[1,2]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT,
[2]University of Utah, Center for Genetic Discovery, Salt Lake City, UT

DNA sequencing has unveiled extensive tumor heterogeneity in several different cancer types, with many exhibiting substantial clonal substructures. Considerable genetic variation exists within individual tumors and between metastatic sites from shared phylogenies, highlighting ongoing tumor growth and evolution. These differences can be further magnified over time by selective pressures introduced by different treatments. Properly identifying and tracing this variation throughout the expansion and progression of a tumor represents a significant challenge. Furthermore, being able to prioritize and focus on mutations most likely to contribute to tumor evolution or that could serve as potential therapeutic targets represents an ongoing problem, but is crucial towards determining possible clinical care.

GEMINI is an effective tool for exploring genetic variation by enabling the querying of multiple genome annotations into a single database. Here we describe our progress towards an adaptation to the GEMINI framework that is more ideally suited for the complex patterns of genetic variation inherent to the study of changes in tumor heterogeneity across multiple biopsies over the course of treatment. This is accomplished by enhancing GEMINI variant annotation to include tumor clonal specifications, allowing for filtering methods that reflect specific tumor properties. Additionally, by incorporating existing tools and databases that facilitate the interpretation of cancer mutations (e.g., CIViC, DGIdb, CRAVAT) into the GEMINI framework, the identification of mutations that may be driving clonal evolution is simplified.

The advances to the GEMINI framework that we are developing will better enable cancer scientists and research oncologists to discern genetic details, at a clonal level, of how a tumor evolves over time, and better interpret these mutations in the context of potential patient care.

# EPIDEMIOLOGICAL EXPLORATION OF FACTORS CONTRIBUTING TO THE POPULATION-LEVEL DIVERSITY IN THE HUMAN GUT MICROBIOME

Suguru Nishijima[1,2,3], Wataru Suda[2,3,5], Kenshiro Oshima[3], Masahira Hattori[2,3,4]

[1]National Institute of Advanced Industrial Science and Technology, Computational Bio-Big Data Open Innovation Lab., Tokyo, Japan, [2]Waseda University, Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan, [3]The University of Tokyo, Graduate School of Frontier Sciences, Tokyo, Japan, [4]RIKEN, Center for Integrative Medical Sciences, Tokyo, Japan, [5]Keio University, Department of Microbiology and Immunology, Tokyo, Japan

Human gut metagenomes shows inter-country variability among populations. However, little is known about factors that profoundly contribute to this variability in the human gut microbiome.

Publically available datasets of metagenomic and 16S rRNA sequences of human gut microbiomes from approximately 2,000 and 4,000 individuals, respectively, across 27 countries were collected. Epidemiological data on dietary intake and antibiotic usage in humans and farms in each country were obtained from the Food and Agriculture Organization Corporate Statistical Database (FAOSTAT) and the scientific literatures, respectively. We examined the association between the microbial composition and the epidemiological data of the 27 countries.

The comparative analysis found that the abundance of several species in the human gut microbiome was strongly correlated with antibiotic usage as well as dietary intake. Notably, several major species showed significant positive correlations with several antibiotics used in humans. We also identified particular antibiotic resistant genes that might be involved in this positive correlation.

Our findings suggest that antibiotics is one of the primary factors responsible for the population-level diversity in human gut microbiomes, and the influence of antibiotics is mediated not only by its individual practice but also global and indirect experience in the population.

# PREDICTION OF B-CELL ACUTE LYMPHOBLASTIC LEUKEMIA SUBTYPES FROM THE EXPRESSION OF LONG NONCODING RNAS AND PROTEIN CODING GENES WITHIN COMMON TOPOLOGICALLY ASSOCIATED DOMAINS

Conor Nodzak[1], Gabrielle Centoducatte[2], J. Andrés 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, Sao Paulo, Brazil

Extensive research has shown how the three dimensional structure of the genome contributes to regulation of gene expression through proximal contacts between regulatory elements and targets throughout cellular development. Long noncoding RNA (lncRNA) transcripts have been shown to facilitate some aspects of chromatin remodeling at nearby protein coding genes loci by the recruitment and binding of proteins capable of modulating the epigenetic state along gene and promoter regions. In addition, lncRNAs may act as microRNA sponges or form complexes with proteins that can have the net effect of both up and down regulation of coding genes. The role of long noncoding RNAs in gene regulation and involvement in disease continues to be an active area of research. Here, we explore the relationships of lncRNAs and protein coding genes within intrachromosomal topologically associated domains (TADs) using expression profiles of B-cell acute lymphoblastic leukemia (ALL) patients. These common TADs represent the genomic regions that form long range physical interactions along individual chromosomes and are comprised of genes and associated regulatory elements. Leukemias are well characterized by a collection of recurrent breakpoints and large genomic rearrangements that can produce disparate gene fusions, as well as non-cognate enhancer mediated gene expression events. We begin by learning the variations among co-expression of lncRNAs with protein coding genes that exist for various subtypes of ALL, and use sets of proximally associated lncRNAs and protein coding genes to construct a statistical model to predict acute lymphoblastic leukemia subtypes based on their expression levels alone. The low misclassification error rate achieved by the multinomial model demonstrates the usefulness of integrating topologically associated domains into expression studies of hematological cancer, and highlights the great degree of coding and noncoding variation across acute lymphoblastic leukemias.

# MERGING GENE ANNOTATIONS ENABLES HIGH-RESOLUTION CELL TYPE IDENTIFICATION

Jim Notwell, Thomas Portmann

Circuit Therapeutics, Neurobiology and Transcriptomics, Menlo Park, CA

Single-cell RNA sequencing (scRNA-seq) has enabled novel cell type identification. Many projects are using this technology to catalog every cell type in different tissues or even the entire body. Cell types identified in this way are determined by the genes they express, making detecting these marker genes crucial. Our study shows that gene detection is complicated by the combination of the 3' sequencing bias of many scRNA-seq technologies and poorly annotated 3' untranslated regions (UTRs), the main portion of the gene being measured. We show that many of the marker genes defining cell populations are missed by common workflows, using D1/D2 medium spiny neurons as an example, two cell types that control movement and whose dysregulation contributes to Parkinson's disease. We address this problem by integrating alternative transcripts from different gene annotations, including across different species. This simple trick allows us to accurately differentiate between these two neuron subtypes and corrects for what appeared to be drop-outs, traditionally thought of as active genes that remain undetected due to stochasticity of gene expression or lack of technical sensitivity. In conclusion, our findings show the combination of methodological 3' sequencing bias and poor UTR annotation is a major source of data loss in current workflows and therefore, is a major obstacle to high-resolution cell type identification.

# ADmiRE: ANNOTATION OF microRNA SEQUENCE VARIATION ACROSS HUMAN POPULATION AND ADULT CANCER DATASETS

Ninad Oak[1,2], Rajarshi Ghosh[2], Kuan-lin Huang[3], Deborah I Ritter[1,2], Li Ding[3], Sharon E Plon[1,2]

[1]Baylor College of Medicine, Molecular and Human Genetics, Houston, TX, [2]Texas Children's Hospital, Pediatrics, Houston, TX, [3]Washington University in St. Louis, McDonnell Genome Institute, Saint Louis, MO

MicroRNAs (miRNAs) are the most abundant class of non-coding RNAs, regulating expression of >60% genes and deregulated in many human diseases. Sequence variation in miRNAs has been shown to alter their expression or targeting ability. However, lack of miRNA variant annotation specific tools and prioritization strategies has hampered global analysis from large-scale sequencing studies.

We developed a set of comprehensive miRNA annotations, Annotative Database of miRNA Elements (ADmiRE), and applied it to the genome aggregation dataset (gnomAD) and subsequently to adult cancer exomes from TCGA. First, we describe the landscape of miRNA variation through re-annotation of gnomAD (15,596 genomes and 123,136 exomes). ADmiRE reveals approximately twice the number of mature miRNA variants compared with other annotation tools that prioritize protein-coding annotations for a total of 23,869 precursor miRNA variants. The allele frequency distribution of intragenic and intergenic located miRNAs closely resembles ($p=0.62$) the exonic regions but is significantly different ($p<0.001$) than intronic and intergenic regions. Expectedly, the seed and mature domains harbor fewer variants compared to the stem-loop sequences or flanking 100bps. We identified 33% (940/2815) highly conserved mature miRNAs across 100 vertebrates. A set of expression-curated 'high-confidence' miRNAs are enriched within this conserved group ($p<0.001$) and gnomAD allele frequency is negatively correlated with the conservation scores across all miRNAs ($p<0.001$). We then applied ADmiRE to 10,568 adult cancer exomes (n=33 cancer types) from TCGA PanCanAtlas to identify rare germline and somatic miRNA mutations. In addition to finding a known miRNA-cancer association (miR-142 mutations in 30% of lymphoma (DLBC) and acute myelogenous leukemias), we identified novel miRNAs somatically mutated in uveal melanoma and cholangiocarcinomas. We also found rare germline variation in miRNAs frequently deregulated in cancers (eg. Let-7 and miR-181 in Breast cancer). We utilized available expression data to identify the impact of these mutations on miRNAs and their target genes' expression.

In conclusion, using ADmiRE annotations we developed a reference for miRNA sequence variation and conservation, which in turn aids the identification of known and novel miRNA variation across TCGA dataset. Availability of ADmiRE to the community will help the analysis of miRNA variation from additional human disease datasets.

# ARTIFICIAL GENOME REARRANGEMENT SYSTEM USING A RESTRICTION ENZYME

Arisa Oda[1], Takahiro Nakamura[1], Nobuhiko Muramoto[2], Hidenori Tanaka[2], Kazuto Kugou[1], Kunihiro Ohta[1]

[1]The University of Tokyo, Graduate School of Arts and Sciences, Tokyo, Japan, [2]Toyota Central R&L labs, Genome Engineering Program, Aichi, Japan

Chromosomal rearrangements which are triggered by DNA double strand breaks followed by DNA repair are expected to affect genome diversification and evolution. In this study, we propose an artificial genome rearrangement system by double strand DNA break induction using a heat activated endonuclease. This genome rearrangement system which can be introduced by transient *in vivo* enzyme activation was applied to *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. Since we succeeded to acquire strains with various phenotypes in both yeast and plant, precise genomic changes of these mutants were analyzed. Whole genome sequencing and tiling array analysis revealed that the chromosomes of these strains have several features such as gene conversions, chromosomal rearrangements, SNPs and InDels. Interestingly, whole genome duplication and large copy number variation events were observed more often in the higher ploidy organisms, diploid yeast and tetraploid Arabidopsis than in haploid yeast and diploid Arabidopsis. This genome rearrangement system would be a useful tool for genome engineering and for the analysis of genome rearrangement.

# THE RARE GENOMES PROJECT: IMPROVING OUR ABILITY TO DIAGNOSE RARE GENETIC CONDITIONS THROUGH A NATIONWIDE PARTNERSHIP WITH FAMILIES

Anne O'Donnell-Luria[1,2,3,4], Melanie O'Leary[1,3], Mekdes Getaneh[1,3], Idara Ndon[1,3], Clara Williamson[1,3], Jaime Chang[1,3], Katherine Blakeslee[1], Julia Goodrich[1,3,4], Monica Wojcik[1,2,4], Nadya Lopez Zalba[1], Anzu Hakone[1], Jennifer Hendry Lapan[1], Esme Baker[1], Moran Cabili[1], Samantha Baxter[1,3], Ben Weisburd[1,3], Heidi Rehm[1,3,4,5], Daniel MacArthur[1,3,4]

[1]Broad Institute of MIT and Harvard, Medical and Population Genomics, Cambridge, MA, [2]Boston Children's Hospital, Genetics and Genomics, Boston, MA, [3]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [4]Harvard Medical School, Boston, MA, [5]Laboratory for Molecular Medicine, Cambridge, MA

A majority of patients in the US with a suspected rare genetic disease do not yet have a molecular genetic diagnosis for their disorder. A diagnosis can have many medical implications including management strategies, prognoses, and family planning using preimplantation genetic diagnosis and in rare instances, can even identify effective treatments. Many families reach an end to their clinical diagnostic odyssey without answers, typically after exome sequencing and analysis, or when their insurance will not cover further genetic testing options. The Rare Genomes Project (RGP; raregenomes.org), launched in June 2017, is a direct-to-families research study leveraging online participation and partnerships with patient advocacy groups to dramatically increase the reach of rare disease genomic research by engaging families as stakeholders at each step. In the first 6 months of our pilot, over 250 families applied to participate in RGP, and about half met inclusion criteria. We enrolled 121 families through video consent, mailed kits for blood draw, and began whole genome sequencing of their samples. Initial analysis includes variant calling with the GATK pipeline and structural variant analysis with a suite of algorithms (Manta, DELLY, MELT, and cn.MOPS). In cases where we identify the genetic cause of the patient's condition, we will perform clinical (CLIA) validation in coordination with a current treating physician identified by the family and provide a genetic testing report to the patient's family and their doctor. Data is rapidly shared through the Broad's Data Use Oversight System (DUOS, duos.broadinstitute.org) and other means, to empower the rare disease research community. Our goal is to provide a proof of principle of the value of a patient-engaged approach to diagnosis and gene discovery in rare disease.

# BIOLOGICAL FACTORS STRONGLY IMPACT CELL TYPE ABUNDANCES IN HUMAN TISSUES

Meritxell Oliva[1], Sarah Kim-Hellmuth[2], François Aguet[3], GTEx Consortium[3], Barbara E Stranger[1]

[1]U Chicago, -, Chicago, IL, [2]New York Genome Center, -, NY, NY, [3]The Broad Institute, -, Cambridge, MA

Human tissues are complex and composed of multiple cell types. Accurate tissue cell-type heterogeneity profiles and knowledge of the extent to which they vary with biological contexts (e.g. sex and age) are lacking for low abundant cell types and non-blood tissues. Tissue cell-type composition can be characterized by single-cell RNA sequencing, but is challenging in many aspects. Alternatively, computational cell-type deconvolution methods applied to the comprehensive catalogue of tissue expression profiles generated by the Genotype-Tissue Expression (GTEx) Consortium (17,382 RNA-seq samples from 838 individuals, v8 release) provides a feasible approach to investigate tissue cellular composition in a high-throughput manner.

Here, we apply cell-type enrichment analysis to quantify the inter-sample variability of 64 immune and non-immune cell types in 49 GTEx tissues and identify numerous tissue-specific associations between cellular abundances and biological factors (donor sex, age, self-reported ethnicity, body mass index (BMI)). Donor sex is a major determinant of breast cell type composition. Stromal cells (smooth muscle, endothelial) are 7-27% more abundant in males and epithelial cells are 500% more abundant in females. Donor age is associated with cellular composition of several tissues; in tibial artery, osteoblasts decrease and chondrocytes increase with age reflecting bone-vascular interactions that shape the artery structure. The immune profile is also widely affected and the most significant age trends are observed in T-helper and dendritic cells. Aging also shapes adipose tissue: adipocyte content is negatively correlated with age even after adjusting for BMI, although this observation might also be compatible with low adipocyte transcriptional activity or larger adipocytes in older tissue. We find an independent impact of donor BMI on adipose tissue composition, particularly on the immune profile which include previously-described and novel obesity-associated effects. Interestingly, several of the age-related trends are non-linear, especially in blood, where they weaken (e.g. basophils), strengthen (e.g. B-cells) or invert their effect (e.g. CD4+ memory T-cells) after the 50-60 years timepoint. Moreover, some trends manifest in a sex-specific manner: neutrophils decrease with age significantly (FDR 5%) more strongly in females. Finally, we detect an effect of donor ancestry shaping tissue heterogeneity predominantly in skin and blood. Importantly, many of the trends we observe replicate using an orthogonal method and in additional datasets.

This work characterizes biological factors associated with tissue cell-type heterogeneity and provides a resource of age-, sex-, and tissue-specific cell-type abundance profiles for the interpretation of tissue composition changes in the context of health and disease.

# A SEARCHABLE CATALOGUE OF VALIDATED ANTIBODIES USED IN THE ENCODE PROJECT

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

Stanford University, Genetics, Stanford, CA

Antibodies are common reagents used in biological sciences to selectively target and isolate specific molecules of interest for various downstream applications. While they are undoubtedly powerful tools, they can add complications when they have not been carefully vetted prior to use. Common problems can include: lot-to-lot variation in efficacy and quality, cross-reactivity resulting in off-target binding, improper use in unintended applications, and poor implementation due to poor training. Indeed, the idea of establishing a standard validation framework to guard against antibody-related data reproducibility issues is gaining momentum in the community (Uhlen et al., 2016; Baker, 2015), as variations in antibody quality and lack of proper vetting have been implicated as major drivers behind the so-called "reproducibility crisis" facing biological research.

ENCODE (ENCyclopedia Of DNA Elements) is an ongoing NHGRI-funded project aimed at cataloguing functional sequence elements in the genome that may act to regulate activity under different cell type and condition contexts. Over half of the ~8,500 experiments done in the project involve the use of antibodies for identifying protein-DNA (ChIP-seq) and protein-RNA (eCLIP-, RIP-, and Bind-and-seq) interactions to delineate candidate regulatory elements. Each of the ~3,200 antibody lots considered for use to date in ENCODE is catalogued by a number of key attributes, including vendor, product number, lot id and are also searchable by gene name via the ENCODE Portal (https://www.encodeproject.org/search/?type=AntibodyLot) to retrieve the experimental data produced by aforementioned antibody-based assays. Moreover, the ENCODE consortium has developed a set of standards (https://www.encodeproject.org/about/experiment-guidelines/#antibody) for the characterization of antibodies to evaluate their sensitivity, specificity and reproducibility. These standards require antibody lots approved for use in ENCODE to be backed by at least two supporting characterizations by different methods (e.g. immunoprecipitations, mass-spectrometry, and knockdowns). Each characterization is then assessed by a panel of reviewers in a transparent manner against those standards to determine their eligibility for use in binding assays. Outcomes for each tested antibody lot are reported on the portal for users to disseminate. All data, metadata, documentation and standards are freely available at the ENCODE portal (https://www.encodeproject.org).

# CONCURRENT ANALYSIS OF GENE EXPRESSION OF COMPLEXED SAMPLES WHICH CONSIST OF DIFFERENT SPECIES: ANALYSIS OF HUMAN AND MOUSE GENE EXPRESSION IN THE LIVER FROM CHIMERIC MOUSE TREATED WITH KMTR2

Yoko Ono[1], Kaito Nihira[1], Ken-ichiro Nan-ya[1], Masakazu Kakuni[2], Toshio Ota[1]

[1]Kyowa Hakko Kirin Co., Ltd., Translational Research Unit, Shizuoka, Japan, [2]PhoenixBio Co., Ltd., Hiroshima, Japan

Some therapeutic antibodies targeting tumor necrosis factor (TNF)-related apoptosis-inducing ligand receptor 2 (TRAIL-R2) have experienced severe hepatotoxicity in clinical trials, but the involvement of TRAIL-R2 in the toxicity still has not been fully elucidated. To analyze the molecular events that could be related to the onset and progression of hepatotoxicity by TRAIL-R2 activation, we implemented RNA-seq for the anti-human TRAIL-R2 monoclonal antibody (KMTR2)-administered chimeric mice with humanized liver (PXB-Mouse ®). Sequence reads were mapped to human- and mouse-reference genome under the stringent condition, and quantified independently.
Following functional analysis of different expressed mouse genes (DEMGs) showed the activation of “chemotaxis of cells". On the other hand, different expressed human genes (DEHGs) characterized as the activation of functions related to cell cycle. The activation of the recruitment of immune cells should reflect the infiltration of immune cells into the liver of PXB-mouse. Given that this approach made it possible to obtain different characteristics of biological function of each species, the combination of PXB-mice and RNA-seq was the suitable way to investigate molecular mechanisms of TRAIL-R2-induced human liver toxicity. RNA-seq and the following analysis we introduce here would also be applicable to the studies that investigate the interaction between tumor cells and stromal cells in xenograft samples.

# PATTERNS OF ROBUSTNESS AND DEREGULATION IN GENE EXPRESSION NETWORKS UNDER DIETARY STRESS

Luisa F Pallares, Anett Schmittfull, Serge Picard, Julien F Ayroles

Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ

At the functional level, differences in sensitivity to environmental stress likely emerge from the disruption of regulatory systems, where more sensitive individuals have decreased transcriptional robustness. However, the degree to which robustness is under genetic control remains unclear. To overcome the limitations of previous studies addressing this question, we developed two resources: First, a synthetic outbred population of *Drosophila melanogaster* that allows us to move away from inbred lines and skewed allele frequency spectrums, and allow us to assay individual outbred flies. Second, we have developed a new method for 3' TagSeq to generate gene expression data from single fly heads at a fraction of the cost of existing methods; this allowed us to screen an unprecedented number of flies. Here, we explore how gene expression in the brain responds to dietary stress by exposing flies to a high-sugar diet compared to the standard diet. We obtained RNA-seq data for thousands of individual fly heads in each environmental condition. The scale of this study allowed us to identify not only condition-dependent eQTL for individual genes, but also condition-dependent co-expression networks. Interestingly, the response of the head's transcriptional profile to high-sugar diet, points towards the regulation of lipid metabolism. This suggests that the fly brain might be controlling the processing and storage of lipids in the body when exposed to stress. By using a systems genetics approach, we explored the general landscape of stress response providing a comprehensive picture of transcriptional robustness in *Drosophila melanogaster*.

# WHAT DID WE MISS? QUANTIFYING AND VISUALIZING VARIANT DETECTION POWER IN SEQUENCING STUDIES.

Brent S Pedersen, Aaron R Quinlan

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

Given a “negative” exome, or whole genome from a patient with a Mendelian disease phenotype, it could be that no causal variant was found for a number of reasons. One reason could be that the pathogenic variant was missed because of a lack of coverage due to poor capture in genes known to be relevant to a phenotype. Using the features in our tool *mosdepth*, we can quickly generate per-base and per-exon coverage information. Building upon the speed of *mosdepth*, we have developed new tools to summarize depth of coverage, and in turn, variant discovery power, in one or more samples, both genome-wide and for each gene and exon. To convey this information, we have developed a viewer that allows quick, interactive visualization of genes and regions. This type of rapid overview provides a new level of insight into the power to detect causal variants in exome, gene-capture, and whole-genome sequencing projects.

We share our findings after analyzing hundreds of exomes and whole genomes from disease cohorts with this tool.

# NCBI SERVICES FOR VARIANT NORMALIZATION, REMAPPING, AND ANNOTATION

Lon Phan, Eric Moyer, Evgeny Ivanchenko, Damon Revoe, Hua Zhang, Wang Qiang, Eugene Shekhtman, David Shao, Ming Ward, Anna Glodek, Brad Holmes

National Center for Biotechnology Information, NIH, Bethesda, MD

Variant normalization and annotation are crucial steps to understanding the genetic basis of their effects on disease and biological functions. Normalizing variants and matching them with data in public variation databases, such as dbSNP and ClinVar, are central to the discovery efforts to identify known and novel variants. However, this process can be difficult because identical variants can be represented differently by variation resources and tools leading to duplications and complicating downstream analysis and annotation. Various tools have been proposed to address variant normalization but they are not robust handling variants in different formats and reporting on different reference sequence coordinates.
Here we present a suite of NCBI variation tools (https://www.ncbi.nlm.nih.gov/projects/variation/services/v0/) that 1) normalizes representation of genetic variants from dbSNP rs IDs, HGVS expressions, and VCF formats for matching existing variants in dbSNP and other public resources and to identify novel variants, 2) remapping of variants between assembly versions and between coordinates on mRNA, protein, and genomic sequences to a normalized sequence coordinate, and 3) provide annotations from dbSNP and ClinVar. The tools employ a new algorithm and data model called Sequence Position Deletion Insertion (SPDI) to describe sequence changes and variant normalization which allow for:

- Producing a contextual allele, a new feature of our SPDI format that corrects left/right-shifting or shuffling for insertions or deletions
- Transforming data into right-shifted HGVS
- Transforming data into left-shifted VCF fields
- Remap (or lift-over) a variant to all available locations based on the alignment dataset used by ClinVar and dbSNP
- Remap (or lift-over) to a canonical representative at a single location that you can use to group identical variants

Acknowledgments:
Work at NCBI is supported by the NIH Intramural Research Program and the National Library of Medicine

# COMPLETE ASSEMBLY OF PARENTAL HAPLOTYPES WITH TRIO BINNING

Sergey Koren[1], Arang Rhie[1], Brian P Walenz[1], Alexander T Dilthey[1,2], Derek M Bickhart[3], Sarah B Kingan[4], Stefan Hiendleder[5,6], John L Williams[5], Timothy P Smith[7], Adam M Phillippy[1]
[1]National Human Genome Research Institute, Genome Informatics Section, Bethesda, MD, [2]Heinrich-Heine-University Düsseldorf, Institute of Medical Microbiology, Düsseldorf, Germany, [3]ARS USDA, Cell Wall Biology and Utilization Laboratory, Madison, WI, [4]Pacific Biosciences, Informatics, Menlo Park, CA, [5]The University of Adelaide, Davies Research Centre, Adelaide, Australia, [6]The University of Adelaide, Robinson Research Institute, Adelaide, Australia, [7]ARS USDA, US Meat Animal Research Center, Clay Center, NE

Reference genome projects have historically selected inbred individuals to minimize heterozygosity and simplify assembly. We challenge this dogma and present a new approach designed specifically for heterozygous genomes. Prior approaches for assembling heterozygous diploid genomes only phase small variants or partially reconstruct the haplotypes. Our "trio binning" method uses short reads from two parental genomes to partition long reads from an offspring into haplotype-specific sets. Each haplotype is then assembled independently. The output of this process is a complete genome for each parental haplotype, containing all classes of haplotype variation assembled from the long reads, including single nucleotide, structural, and copy number variants.
To demonstrate the effectiveness of trio binning on a heterozygous genome, we sequenced an F1 cross between cattle subspecies *Bos taurus taurus* and *Bos taurus indicus*, and assembled both parental haplotypes with NG50 haplotig sizes >20 Mbp each, surpassing the quality of current cattle reference genomes. In fact, these haplotype-specific contigs (haplotigs) are larger than the *scaffolds* of many prior inbred or haploid reference genome assemblies. In addition to their high continuity, both haplotypes approach 99.999% accuracy at the base level using PacBio data alone. Further application of this method to a benchmark human trio (NA12878 and parents) also achieved high accuracy and recovered complex structural variants missed by alternative approaches such as 10x Genomics linked-read sequencing.
Trio binning of both the human and cattle haplotypes successfully reconstructed highly heterozygous loci important for immunity and adaptation. For example, in human, both parental haplotypes of the Major Histocompatibility Complex (MHC) were accurately assembled and showed perfect human leukocyte antigen (HLA) gene typing accuracy. For cattle, many heterozygous regions between the newly assembled Angus and Brahman haplotypes intersected with previously identified quantitative trait loci (QTL). For example, in one structurally diverse region relative to the Brahman haplotype, the Angus haplotype is missing a ~140 kbp duplication containing *GBP2*, while containing its own duplicated *GBP6*-like sequence. This region intersects with previous QTLs linked to muscularity and visual conformation score, making it a suggestive candidate for adaptation among the cattle breeds.
Given the quality of the assemblies we were able to achieve with this approach, we propose trio binning as a new best practice for diploid genome assembly that will enable platinum-quality reference genomes and new studies of haplotype variation and inheritance.

# ANTIVIRAL ENZYME APOBEC3G INTRODUCES CLUSTERED INHERITED MUTATIONS THAT FUEL ADAPTATION IN HUMAN POPULATIONS

Yishay Pinto[1,2], Edward Li[2], Erez Y Levanon[1], Alon Keinan[2,3,4,5]

[1]Bar-Ilan University, Mina and Everard Goodman Faculty of Life Sciences, Ramat-Gan, Israel, [2]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY, [3]Cornell University, Center for Comparative and Population Genomics, Ithaca, NY, [4]Cornell University, Center for Vertebrate Genomics, Ithaca, NY, [5]Cornell University, Center for Enervating Neuroimmune Disease, Ithaca, NY

The molecular clock is based on the assumption that mutations mostly occur randomly and independent of each other, thereby accumulating at a constant rate over time. We recently discovered a mutational process driven by the activity of a member of the APOBEC family of deaminases. APOBECs can concurrently introduce clusters of mutations in single-stranded DNA (ssDNA), which enables diversifying immunoglobulin genes and inactivating viral DNA. Recently, this gene family has been shown to introduce clusters of somatic mutations caused by cytosine deamination in multiple types of cancers. We hypothesized that APOBEC3 members, following their rapid expansion in primates, may have introduced similar clusters of mutations in the germline that can be inherited. We tested this hypothesis by relying on the unique mutagenesis pattern of APOBEC3s: In the presence of one of several sequence motifs that include a C, it can mutate this together with many nearby C nucleotides. We traced such activity of APOBEC3G in a comparative genetic analysis of the human, archaic humans (Neanderthal and Denisovan) and chimpanzee genomes. We conservatively identified tens of thousands of such clustered nucleotide-specific mutations. We hypothesized that these mutation clusters are less likely to be neutral, as compared to single-nucleotide mutations, with many not being observed as inherited due to being lethal or very deleterious early in conception and development. Combined with them occurring more often in functional elements due to APOBECs acting on ssDNA, the mutation clusters that segregate are more likely to be targeted by positive selection. Hence, we applied several population genetic tests of adaptation for a subset of mutation clusters that are also polymorphic in human populations. Our results point to positive selection on several such clusters of genome-wide significance. Of interest is that most instances of adaptation are population-specific, which seem to correspond to adaptation to local environments, and act on cis-regulatory eQTLs. These results suggest that APOBEC3G-introduced mutations are more likely to have functional consequences, than is already expected by APOBECs' tendency to mutate ssDNA, which are maintained by natural selection. Combined, we provide evidence for exaptation of an antiviral mechanism as a source of genomic variation with population-specific functional consequences in humans.

# CentiSNPs: A COMPENDIUM ANNOTATION TO ANALYZE EQTLS IN HETEROGENEOUS TISSUES

Andrew Freiman[1], Anthony Findley[1], Xiaoquan Wen[2], Francesca Luca[1,3], Roger Pique-Regi[1,3]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [3]Wayne State University, Department Obstetrics and Gynecology, Detroit, MI

A large fraction of genetic loci identified by genome wide association studies (GWAS) are located in non-coding regions of the genome and likely act by affecting gene regulation. Expression quantitative trait loci (eQTLs) analysis has emerged as a powerful tool to link genetic variants with changes in gene expression, but as in GWAS, eQTL resolution is limited by linkage disequilibrium. High resolution annotations of genetic variants are useful in locating and predicting the effect of the causal variant in transcription factor (TF) binding sites, but the available data are often on specific cell-types and may not exactly match the eQTL tissue of interest. Here, we propose to use a new annotation (centiSNPs) that aggregates CENTIPEDE chromatin footprints across hundreds of cell-types and retains the TF identity as a categorical variable (1117 factors) to be used for QTL enrichment analysis. This is a useful simplification that assumes that when a TF is active it may bind to similar locations across different cell-types. Using centiSNPs, we analyzed GTEx eQTLs in 44 different tissues with the deterministic approximation of posteriors (DAP) approach we recently developed. By combining multiple cell-types, we obtain a more complete genome-wide annotation for each TF, with tighter estimates for the enrichment parameters. For homogeneous tissues with obvious matching cell-types, we obtain a similar set of enriched TFs (e.g. HNF4A in liver), as compared to cell-type matched annotations. The signature of enriched TFs is highly representative of the underlying eQTL tissue, even if the relevant cell types were not available in building the centiSNPs compendium. Additionally, for many tissues (e.g. fat), we discover that the most enriched categorical annotations correspond to TFs that can be modulated by the environment (e.g. glucocorticoid receptor). These results support the presence of latent GxE interactions in the GTEx dataset, which are likely the consequence of non-homogeneous environmental conditions.

# NETREX: NETWORK REPROGRAMMING USING EXPRESSION - UNCOVERING SEX-SPECIFIC GENE REGULATION IN DROSOPHILA

Yijie Wang[1], Dong-Yeon Cho[1], Hangnoh Lee[2], Justin Fear[2], Brian Oliver[2], Teresa M Przytycka[1]

[1] NCBI/NLM, NIH, Bethesda, MD, [2]NIDDK, NIH, Bethesda, MD

Maintenance of cell type specific states, response to stress, sexual dimorphism, and other cell functions are controlled by gene regulatory programs. In particular, Gene Regulatory Networks (GRNs) capture the regulatory relationships between transcription factors (TFs) and their target genes. Since GRNs provide information that is essential for a global understanding of the logic of gene-gene interactions, inference of these networks is one of the key challenges in system biology. Methods to infer GRNs typically combine computational approaches and experimental data collected from different sample types, different conditions, different techniques, and different labs. Such data integration leverages dependencies that can be confidently uncovered thanks to the multitude of surveyed conditions, but leads to context-agnostic wiring diagrams. These context-agnostic networks do not accommodate regulatory program reality, which is specific to tissue types, developmental stages, sex, and other factors. Here we introduce, NetREX, a novel method to construct GRNs by iterative reprogramming of a prior network, given a prior network and expression data. The main idea of NetREX is to reprogram the prior network by adding and removing edges to obtain a network that provides the best explanation of the observed gene expression. Simultaneously, NetREX optimizes several other objectives to ensure that the resulting network is biologically relevant. NetREX is the first approach that systematically explores the landscape of possible GRN topologies to generate context-specific GRNs. We applied NetREX and PriorBoost to construct the first sex-specific Drosophila GRNs. We validated the predicted GRNs computationally and experimentally and compared the results obtained from NetREX with those obtained from alternative methods. NetREX constructed sex-specific Drosophila GRNs that, on all applied measures, strikingly outperformed networks obtained from other methods indicating that NetREX is an important milestone towards building more accurate GRNs.

# IDENTIFICATION OF SUBCLONE-SPECIFIC TUMOR CELLULAR PHENOTYPES BY SINGLE-CELL ASSIGNMENT TO BULK DNA-DERIVED SUBCLONES.

Yi Qiao[1,2], Xiaomeng Huang[1,2], Samuel Brady[3], Andrea Bild[4], William Johnson[5], Gabor Marth[1,2]

[1]USTAR, Center for Genetic Discovery, Salt Lake City, UT, [2]University of Utah, Eccles Institute of Human Genetics, Salt Lake City, UT, [3]University of Utah, Pharmacology and Toxicology, Salt Lake City, UT, [4]City of Hope, Medical Oncology & Therapeutics Research, Duarte, CA, [5]Boston University, Computational Biomedicine, Boston, MA

Several approaches are now available for subclonal reconstruction of heterogeneous tumor biopsies from somatic variant allele frequencies measured in bulk DNA sequencing datasets, including our own SubcloneSeeker program. However, to understand the role of each subclone in the context of chemoresistance, relapse, or metastasis, it would be highly desirable to understand subclone-specific cell phenotype. Single-cell RNA-Seq technology now allows one to study the transcriptomic characteristics and phenotypic behavior of individual cells. Methods also exist that cluster individual cells based on similar expression profiles, but no algorithm exits at present that can link individual cells to bulk DNA sequence-derived subclones.

Here we present a computational approach to make such assignments. Computationally derived subclones from bulk DNA sequencing data are defined by specific combinations of somatic mutations (SNVs or CNVs). scRNA-Seq data can be used to assess the presence or absence of these, subclone-defining, mutations in individual cells, which can then be used for subclone-identity assignment. This process is complicated by the sparseness of scRNA-Seq data i.e. the fact that the majority of somatic mutations in a tissue are not covered by sufficiently deep sequencing data in any given cell. We utilize a Bayesian-probabilistic framework to deal with missing single-cell data, using bulk-derived subclone frequencies as priors for subclone assignment.

We have successfully applied this method to patient datasets, consisting of both deep bulk whole-genome DNA sequencing data and scRNA-Seq data collected using 10X technology. To our knowledge, this study represents the first attempt to unify genomic and transcriptomic subclonal analysis in cancer.

# EFFICIENT AND EXACT COMPUTATION OF LINKAGE STATISTICS FOR INFERENCE

Aaron P Ragsdale, Simon Gravel

McGill University, Human Genetics, Montreal, QC, Canada

Patterns of genetic polymorphism within and across populations encode information about demographic and evolutionary history. Statistical approaches to infer such histories have often relied on simple summaries of the data, such as the distribution of allele frequencies (called the allele frequency spectrum, or AFS). While the AFS has proven to be a powerful tool for inference, it assumes independence between sites and thus does not take advantage of information contained in patterns of correlation of allele frequencies at linked loci. Statistics on linkage disequilibrium (LD) are themselves informative: two-locus haplotype frequencies are commonly used to infer recombination maps and are sensitive to demography. However, computing expected values for two-locus statistics has been numerically challenging beyond the simplest situations. Here we present an exact approach to efficiently compute LD statistics with arbitrary recombination rate, a flexible mutation model, and complex multi-population demography with continuous migration. We integrate this method into an inference approach for reconstructing demography from joint AFS and LD statistics.

Technically our method avoids computing the full two-locus sampling probabilities, and instead solves a recursion on a drastically reduced set of statistics that includes moments on the covariance of allele frequencies ($D$). This approach follows a set of equations studied by Hill and Robertson in the late 1960s, who considered the evolution of low order moments of $D$. We show that this recursion can be extended to efficiently compute arbitrarily high moments of $D$. We similarly extend the recursion to multi-population demographies with splits and continuous migration. This allows for the direct computation of expected correlation of LD statistics across populations under realistic and complex demographic scenarios. We couple our computational approach for linkage statistics with our existing moments-based software for computing the AFS in order to perform likelihood-based demographic inferences from the joint distribution of allele frequencies and LD statistics. Finally, we revisit classical models of human demographic history and compare inferences using joint AFS and LD statistics to those using the AFS alone.

# USING NEURAL NETWORKS TO PREDICT GENE EXPRESSION: APPLICATION IN GENOMIC SELECTION FOR FIELD TRAITS IN MAIZE

Guillaume P Ramstein[1], Edward S Buckler[1,2]

[1]Cornell University, Institute for Genomic Diversity, Ithaca, NY, [2]USA Department of Agriculture, Agricultural Research Service, Ithaca, NY

Genomic selection consists of selecting individuals based on their breeding values estimated by genomic markers. In plant breeding, such prediction models have often achieved satisfactory accuracy for traits of agronomic interest. However, the functional basis for the relationship between genomic markers and traits of interests is typically not reliably captured by such models. In order to develop more robust prediction models which may shed light onto the functional relationships between genes and traits of interest, we are developing a framework to predict field traits based on gene expression levels.

In this study, we investigate the performance of neural networks to predict gene expression in maize. Expression levels were measured by 3'-RNA-seq in a diverse panel of 282 maize lines. Expression levels were then imputed based on recombination events in the 25 families of the Nested Association Mapping panel (NAM panel), each comprising up to 200 inbred lines. Here we develop neural network models to predict expression based on gene sequences. First, we determine an optimal set of genotypes to use for fitting neural networks, considering reference lines from both maize and sorghum. We then compare predictions to observed and imputed levels. Finally, we perform validation on selected NAM families to assess whether incorporating expression information into genomic selection models results in improved accuracy, compared to standard models which directly model the relationship between genomic markers and traits of interest.

Our study assays the usefulness of expression information for increasing the genetic gains achieved by genomic selection. Models based on prediction information are particularly promising in that they allow for the transfer of information across species, thereby transferring the information gained in well-studied crop species, e.g., maize, to species with relatively few genetic resources, e.g., sorghum.

# COMPREHENSIVE ALTERNATIVE SPLICING ANALYSIS OF 8,512 TCGA DONORS

André Kahles[1,2], Kjong-Van Lehmann[1,2], Nora C Toussaint[3], Matthias Hüser[1], Stefan Stark[1], Timo Sachsenberg[5], Oliver Stegle[4], Oliver Kohlbacher[5], Chris Sander[6], TCGA PanCanAtlas Network[1], Gunnar Rätsch[1,2]

[1]ETH Zurich, Department of Computer Science, Zurich, Switzerland, [2]Memorial Sloan Kettering Cancer Center, Computational Biology Department, New York, NY, [3]ETH Zurich, NEXUS Personalized Health Technologies, Zurich, Switzerland, [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton Cambridge, United Kingdom, [5]University of Tübingen, Department of Computer Science, Tübingen, Germany, [6]Dana-Farber Cancer Institute, cBio Center, Boston, MA

Cancer genome analysis commonly focuses on somatic non-synonymous protein-altering mutations and their potentially pathogenic impact. The Cancer Genome Atlas (TCGA) has built up a unique resource of complementary molecular data, that also allows for an in-depth integrative analysis, a direction that has received much less attention so far.
We analyze RNA and whole exome sequencing data of tumors from 8,512 donors spanning a range of 32 cancer types. We focus on a) the underlying genetic changes leading to splicing variability in tumors, b) a comprehensive analysis of quantitative and qualitative changes of alternative splicing (AS) in tumors, and c) determining to which extent splicing aberrations can be exploited for immunotherapy. To our knowledge, this study presents the first *comprehensive analysis of known as well as novel alternative splicing events* across all suitable TCGA samples.
Depending on the cancer type, we find an up to 40% increase in AS in tumor samples relative to normal and detect an average of 740 "neo-junctions"- exon-exon-junctions that are not normally observed in normal samples. The integration of splicing data over all samples and genes allows us to uncover strong splicing signatures for individual cancer types and subtypes, often even overpowering the respective tissue-specific effects. Our analysis of splicing phenotypes combined with variants obtained from re-analysis of WXS data for a genome-wide association analysis is the largest reported splicing quantitative trait loci (sQTL) study with respect to number of donors thus far. For the first time, the available data provides sufficient statistical power to detect somatic *trans*-sQTL at scale. Aside from confirming sQTL-variants in splicing factors U2AF1 and SF3B1, we also detect novel *trans*-sQTL, such as IDH1 and ANAPC1. Finally, our study is the first to comprehensively analyze to which extent alternative splicing in tumors leads to new RNA transcripts that are translated into tumor-specific peptides that may be targeted by immunotherapy. Through an integrative analysis of GTEx, TCGA, and CPTAC protein mass spectra, we can show for three tumor types that the resulting mRNAs are indeed translated into tumor-specific peptides. From all peptides predicted to be MHC-I binders, we are able to confirm on average 16 splicing derived neo-epitopes per sample - an over 10-fold increase over the number of neo-epitopes predicted with classic approaches taking into account only SNVs.

# PAN-CANCER STUDY OF HETEROGENEOUS RNA ABERRATIONS AND ASSOCIATION WITH WHOLE-GENOME VARIANTS

Claudia Calabrese[1], Natalie Davidson[2], Nuno Fonseca[1], Yao He[3], Andre Kahles[2], Kjong Lehmann[2], Fenglin Liu[3], Yuichi Shiraishi[4], Cameron Soulette[5], Lara Urban[1], ICGC PCAWG Transcriptome Analysis Group[2,1,5], Alvis Brazma[1], Angela Brooks[5], Jonathan Göcke[6], Gunnar Rätsch[2], Roland Schwarz[7], Oliver Stegle[1], Zemin Zhang[3]

[1]EBI, EMBL, Hinxton, United Kingdom, [2]ETH, CS, Zürich, Switzerland, [3]Peking University, Peking, China, [4]University of Tokyo, Tokyo, Japan, [5]UCSC, Santa Cruz, CA, [6]Genome Institute, Singapore, [7]MDC, Berlin, Germany

We present the most comprehensive catalogue of cancer-associated gene alterations through characterization of tumor transcriptomes from 1,188 donors across 27 tumor types of the Pan-Cancer Analysis of Whole Genomes project. This study provides an extensive catalog of RNA alterations and insights into how cancer augments the molecular mechanisms. We produced catalogues of: alternative promoters, splicing, fusions, and allele specific expression (ASE) and performed association studies across RNA and alteration types. We also analyse RNA and DNA alteration using an integrative approach. We re-analysed RNA-seq libraries and assessed data quality, alignment strategy and library size normalization. We map expression quantitative trait loci (eQTL) of germline variants identifying 2,509 genes with a germline effect (FDR < 5%) and 426 cancer-specific eQTL. We find that somatic copy number alteration (SCNA) is a major driver of expression variation followed by somatic mutation in flanking regions of the individual genes and nearby germline variants. We assessed multiple strategies for association between somatic mutational burden and gene expression identifying 649 somatic eQTL (FDR<5%) including genes with known roles in the pathogenesis of specific cancers. ASE analysis of variant effects showed that somatic protein truncating mutations are a major driver of ASE. Alternative splicing variation showed a strong association between splicing outliers and nearby splice site mutations with most of these variants having a negative effect on splicing. We also found that Alu sequence exonization is frequently observed. The landscape of RNA-fusions showed that structural variants are the major cause of gene fusions, but does not explain all events. We identified a novel class of fusions called “bridged” fusions, where a third genomic location is used as a bridge. To identify cancer relevant genes, we created a new method for performing recurrence analysis on all transcriptomic features. Our analysis yielded 1,012 genes (emp. p-Value<0.05) which show enrichment for cancer census genes and driver genes. Among the top 5% of our ranked genes is CDK12, which is impacted by multiple, non-overlapping, alterations within a protein kinase domain associated with dysregulation of DNA repair in cancer. In our cohort, we found 87 samples that have an alteration in this domain, with 64 samples having only a RNA alteration in the domain. This is the first study of the transcriptional landscape of 1,188 cancer patients with both whole genome and transcriptomic data. This work represents the first large-scale assessment of the effects of a wide range of transcriptome alteration in context to DNA-alterations.

# THE RELATIONSHIP BETWEEN EVOLUTIONARY RATES, EXPRESSION AND GENE DUPLICATION IN THE SEROTONERGIC SYSTEM

Guillermo Reales[1], Vanessa R Paixão-Côrtes[2], Maria Cátira Bortolini[1]

[1]Universidade Federal do Rio Grande do Sul (UFRGS), Departamento de Genética, Porto Alegre, Brazil, [2]Universidade Federal da Bahia (UFBA), Departamento de Biologia Geral, Salvador, Brazil

Understanding how selection shapes gene evolution is a major aim of molecular biology. Multiple variables influence the evolutionary history of protein-coding genes, such as evolutionary rates, gene and genome duplication, gene function, position in genetic networks, and expression. In particular, a negative correlation between levels of gene expression and protein evolutionary rates appears at a genomic scale in a broad range of lineages, from unicellular organisms to animal and plants. Expression breadth has also been suggested to be a strong predictor of protein evolutionary rates in mammals. In this work, we explore the relationship among evolutionary rates, expression and gene duplication in the serotonergic system. This network has two types of receptors that went through several duplication events (18 genes in total), and which together with two tryptophan hydroxylases (TPH1 and TPH2), the serotonin transporter (SLC6A4), and a monoamine oxidase (MAOA) comprise the serotonergic system. Serotonin (5-HT) has a fundamental regulatory role in numerous physiological processes across most life forms. The genes of the system show overall low evolutionary rates and heterogeneous expression distribution among tissues and species.

Here we analyze the expression and evolutionary rates in these 22 serotonergic genes of 20 chordates. For it, we retrieved their coding sequences from Ensembl, and estimated the evolutionary rates in PAML v4.8, and we recovered RNA-seq expression data from the Bgee v14 server. Subsequently, we calculated scores for expression breadth across tissues and species, and studied the relationship between expression levels and breadth, and evolutionary variables, such as average dN/dS (ω), and proportion of sites under purifying, neutral, and positive selection.

Our results show signals of positive selection for six genes (HTR1E, HTR2C, HTR3C, HTR3D, HTR3E, and HTR7; $\omega = 3.43, 1.77, 2.08, 13.10, 2.25$ and $31.42$, respectively). Three of them (HTR3C-E) arose from tandem duplications, and show extraordinarily high mean ω (0.38, 0.46 and 0.53, respectively), while having low expression levels. HTR3C, HTR3D and HTR3E duplication seems to be more recent than the other genes, and they appear in less species. For most genes, on which purifying selection is the major force we observe a slight positive correlation ($\rho = 0.18$, $P=0.48$) between ω and expression, which might indicate a different behavior of the serotonergic system with what is observed at genomic level.

# NEXTGEN SEQUENCE PROFILING OF EUKARYOTIC DIVERSITY IN DEEP SUBSURFACE BIOFILMS

Bethany Reman, Oxana Gorbatenko, Cynthia Anderson, Shane Sarver

Black Hills State University, Biology, Spearfish, SD

For the last decade, the underground area of the former Homestake Gold Mine in Lead, South Dakota has been operating as the Sanford Underground Research Facility (SURF). The use of this subterranean space for science has allowed biologists access to some of the more remote parts of the old mine. The level 1478 meters below the surface in a peripheral tunnel provides a unique opportunity to study microbial life that has become established in a unique ecosystem. No sunlight nor cosmic radiation penetrate to this depth. The temperature of our remote study site reaches 33$^{o}$C and humidity is one-hundred percent. It is at this area that biofilms develop where deep subsurface (non-meteoric) water interfaces with the rock wall of the tunnels.

Six biofilms were harvested and examined by compound light and scanning electron microscopy to detect eukaryotic groups within the samples. One sample was chosen for custom amplicon ribosomal DNA (rDNA) library preparation for NextGen sequencing to assess the diversity of the eukaryotic community. Libraries were constructed for the 18S variable regions 4 and 8-9, and ITS2.

Sequence data revealed that several metazoan taxa reside within the biofilms as well as many protist and fungal species. Molecular data was complemented by microscopy, which verified that members of phyla Nematoda, Platyhelminthes, Rotifera and Gastrotricha are present in the biofilms.

The organisms that live in the deep subsurface habitats of SURF may provide insight into evolutionary processes, ideas about astrobiology, or even provide potential opportunity for medical applications. Future work includes NextGen sequencing of additional biofilm samples to further assess diversity and to understand what types of eukaryotes dwell deep in the crust of the earth. Prokaryotic diversity will also be profiled with 16S rDNA custom amplicon sequencing of the same biofilms.

# DNA BINDING PREFERENCES OF HELICONIUS OPTIX TRANSCRIPTION FACTOR

Jose A Rodriguez-Martinez

University of Puerto Rico - Rio Piedras, Biology, San Juan, PR

Sequence-specific DNA-binding proteins such as transcription factors are key determinants of cellular state. Transcription factors serve as the ultimate link between genome and its diverse phenomes. Evaluating the protein-DNA interactome of transcription factors is a nontrivial challenge that limits our ability to decipher gene regulatory networks. Optix is a homeodomain transcription factors that has been identified as a master regulator of red color patterns in Heliconius butterfly wings. However, the direct gene targets of optix and their effects in final wing coloration remain to be determined. Using SELEX-seq, we have determined the intrinsic DNA-binding preferences of optix. Comprehensive protein-DNA interactome are used to identify optix's putative gene targets throughout the Heliconius genome. These results will enable us to decipher the gene regulatory network that controls wing coloration in Heliconius butterflies.

# DEEP TAXON SAMPLING REVEALS THE EVOLUTIONARY DYNAMICS OF NOVEL GENE FAMILIES IN *PRISTIONCHUS* NEMATODES

Neel Prabh, Christian Roedelsperger

Max Planck Institute for Developmental Biology, Evolutionary Biology, Tuebingen, Germany

The widespread identification of genes without detectable homology in related taxa is a hallmark of genome sequencing projects in animals, together with the abundance of gene duplications. Such genes have been called novel, young, taxon-restricted, or orphans, but little is known about the mechanisms accounting for their origin, age and mode of evolution. Phylogenomic studies relying on deep and systematic taxon sampling and employing the comparative method can provide insight into the evolutionary dynamics acting on novel genes. We used a phylogenomic approach for the nematode model organism *Pristionchus pacificus* and sequenced six additional *Pristionchus* and two outgroup species. This resulted in 10 genomes with a ladder-like phylogeny, sequenced in one laboratory using the same platform and analyzed by the same bioinformatic procedures. Our analysis revealed that 68-81% of genes are assignable to orthologous gene families, the majority of which defined nine Age classes with presence/absence patterns that can be explained by single evolutionary events. Contrasting different Age classes, we find that older Age classes are concentrated at chromosome centers whereas novel gene families preferentially arise at the periphery, are lowly expressed, evolve rapidly, and have a high propensity of being lost. Over time, they increase expression and become more constrained. Thus, the unprecedented phylogenetic resolution allowed a comprehensive characterization of the evolutionary dynamics of *Pristionchus* genomes indicating that distribution of Age classes and their associated differences shape chromosomal divergence. This study establishes the *Pristionchus* system for future research on the mechanisms that drive the formation of novel genes.

# FINE-SCALE MAPPING REVEALS DRAMATICALLY LOWER RATES OF RECOMBINATION IN RHESUS MACAQUES THAN IN HUMANS OR GREAT APES

Jeffrey Rogers[1,2], Cheng Xue[1,2], Navin Rustagi[1,2], Xiaoming Liu[3], Muthuswamy Raveendran[1,2], R.Alan Harris[1,2], Manjunath G Venkata[4], Fuli Yu[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Baylor College of Medicine, Dept. of Mol. and Human Genetics, Houston, TX, [3]Univ. of Texas Health Science Ctr., School of Public Health, Houston, TX, [4]Oak Ridge National Laboratory, Oak Ridge, TN

Meiotic recombination is a fundamental genetic process that serves essential cellular functions, determines genome haplotype structure and affects the process and outcome of natural selection. Rates of recombination per megabase of DNA differ across species, and among populations within species. In most mammals, recombination occurs primarily in "hotspots" that experience crossover events much more frequently than the genome-wide average. Rhesus macaques (Macaca mulatta) are the most commonly used nonhuman primate in biomedical research, and are extensively analyzed for fundamental aspects of primate biology. The rate and distribution of recombination in the rhesus genome will have influenced their current population genetics, the past effects of selection on macaque populations and other species-level characteristics. We used genotypes for 18 million SNPs scored across the autosomes of 123 Indian-origin rhesus macaques, and LDhat software, to calculate a fine-scale recombination map for this species. We used the same analytical pipeline, and 1000 Genomes OMNI data to estimate recombination rates for Yoruban, European and Chinese populations as the control experiments and validated our pipelines. Comparisons show that rhesus macaques have a significantly lower genome-wide rate of recombination (0.45 $\pm$ 0.29 cM/Mb) than equivalent estimates for humans, ranging from 1.092 $\pm$ 0.79 to 1.094 $\pm$ 0.83 cm/Mb depending on the population. The chimpanzee and gorilla rates are similar to human. In addition, the standard deviation of recombination rate across genomic segments (either 1 Mb, 100 kb or 10 kb windows) is lower in rhesus than humans or chimpanzees. Human recombination shows a dramatically higher proportion of windows with >1.0 cM/Mb than does rhesus (regardless of window size). We conclude that, since the divergence of the human/chimpanzee ancestor from macaque ancestor ~26 mya, the human and chimpanzee genomes developed a larger number of genomic regions with higher rates of recombination, while rhesus recombination occurs less frequently and more homogeneously across the genome.

# GTEx RESOURCES AND NEW 'BARCHART' AND 'INTERACTION' TRACK DISPLAYS INSPIRED BY GTEx IN THE UCSC GENOME BROWSER

Kate Rosenbloom, Galt Barber, Max Haeussler, Angie Hinrichs, Christopher Lee, Jairo Navarro, Brian Raney, Cath Tyner, John Vivian, Brian Lee, Ann Zweig, Bob Kuhn, Jim Kent

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

Inspired by the comprehensive and high-quality gene expression and variant datasets released by the NIH Genotype-Tissue Expression (GTEx) project, we have created new visualization and data mining resurces in the UCSC Genome Browser.

This effort began with the GTEx Gene Expression and Transcript Expression tracks, released in 2016 which summarize gene expression by tissue across thousands of samples via a newly developed bar graph display in genomic context. GTEx gene expression by tissue has also been incorporated into the UCSC Genes track detail pages and the Gene Sorter Tool, which previously relied on microarray expression datasets.

With the accumulation of sufficient GTEx variant data to support expression QTL analysis, the GTEx consortium last year published analyses identifying variant/gene interactions, which are the basis of the browser GTEx eQTL tracks released in 2017. A GTEx Trans eQTL track is under development for release in spring 2018. The latter track features a new paired region interaction display with curved connectors drawn between variant and affected gene. Recognizing that bar graph and curved interaction displays would be useful to browser users formatting their own data for display via track hub or custom track presentation, we have generalized the data management and display for broader use, as the new 'barChart' and 'interaction' track types.

Other GTEx resources at UCSC include two public track hubs: the GTEx RNA-Seq Signal hub which hosts >7500 tracks of RNA-seq read coverage from GTEx samples and the GTEx Analysis Hub, currently hosting tracks of allele-specific expression (ASE).

As a companion to the GTEx data resources, we commissioned artwork on which to base an interactive human anatomy 'Body Map' of the 52 GTEx tissues. This graphic provides a user-friendly tissue legend for GTEx resources in the browser, and is available for tissue selection on the configuration page of several GTEx tracks.

# LONG READ SEQUENCING OF TUMOR DNA TO ANALYZE FUSIONS AND AMPLICONS

Jeffrey A Rosenfeld[1], Sara Goodwin[2], Robert Wappel[2], Shridar Ganesan[1]

[1]Cancer Institute of NJ, Pathology, New Brunswick, NJ, [2]Cold Spring Harbor Laboratory, Genome Genter, Woodbury, NY

Cancer genomes are characterized by changes in the genome on many different levels. In breast cancer, the HER2 amplicon is very important, but is not been well understood. It is known that the genes in the region of HER2 are increased in copy number, but the exact organization of the amplicon in cancer cells is not known. There is the potential for rearrangements between the genes and for different segments of the amplicon to exist and varied copy numbers. We have used long read sequencing of primary breast cancers to determine their HER2 amplicon structure.

# STABILIZING ROLE OF EARLY ALU EXPANSION VIA NON-ALLELIC HOMOLOGY DIRECTED REPAIR OF SPONTANEOUS DSBs IN GERMLINE

Tanmoy Roychowdhury, Alexej Abyzov

Mayo Clinic, Division of Biomedical Statistics and Informatics, Rochester, MN

Structural variations (SV) in human genome originate from different mechanisms related to repair and retrotransposition and are frequently associated with different diseases. We analyzed 26,927 germline SVs of different origin from phase 3 of the 1000 Genomes Project in context of low (Hi-C compartments) and high (DNase sites, nucleosome positioning, histone marks) resolution epigenomic properties. Analyses revealed differential distribution and consequence of SVs of different origin, for instance, NAHR events are more prone to disrupt TAD boundaries as well as loop anchors while processed pseudogenes can create accessible chromatin. NAHR deletions likely reflect errors either during directed recombination in meiosis or during homology-directed repair (HDR) of spontaneous DSBs that in turn was found to be associated with transcription factor binding sites in accessible chromatin. While an allelic HDR doesn't generate any deletion, a non-allelic HDR (due to unavailability of sister chromatid) may lead to gene conversion or deletion. We found that majority of NAHR deletions are of non-meiotic origin i.e., strongly associated with spontaneous DSBs. This evidence along with strong physical interaction of NAHR breakpoints suggests that majority of NAHR SVs originate as error during non-allelic HDR of spontaneous DSBs. Contrasting observations from NHEJ deletions suggest HDR as choice of DSB repair mechanism in functional, gene-rich A compartments. Additionally, a striking difference between distributions of fixed and variable Alus across genome compartments was observed. Through co-localization of fixed Alus and NAHR in A compartment, we hypothesize that early Alu expansion has a stabilizing role on human genome via promotion of high fidelity HDR.

# SIGNIFICANT SHARED HERITABILITY UNDERLIES SUICIDE ATTEMPT AND CLINICALLY PREDICTED PROBABILITY OF ATTEMPTING SUICIDE

Douglas Ruderfer[1,2], Colin Walsh[2], Matthew Aguirre[3], Yosuke Tanigawa[3], Jessica Ribeiro[4], Joseph Franklin[4], Manuel Rivas[3]

[1]Vanderbilt University Medical Center, Medicine, Nashville, TN, [2]Vanderbilt University Medical Center, Biomedical Informatics, Nashville, TN, [3]Stanford University Medical Center, Biomedical Data Science, Stanford, CA, [4]Florida State University, Psychology, Tallahassee, FL

Suicide accounts for nearly 800,000 deaths per year worldwide with rates of both deaths and attempts rising. Family studies have estimated substantial heritability of suicidal behavior; however, collecting the sample sizes necessary for successful genetic studies has remained a challenge. We utilized two different approaches in independent datasets to characterize the contribution of common genetic variation to suicide attempt. The first is a patient reported suicide attempt phenotype from genotyped samples in the UK Biobank (337,199 participants, 2,433 cases). The second leveraged electronic health record (EHR) data from the Vanderbilt University Medical Center (VUMC, 2.8 million patients, 3,250 cases) and machine learning to derive probabilities of attempting suicide in 24,546 genotyped patients. We identified significant and comparable heritability estimates of suicide attempt from both the patient reported phenotype in the UK Biobank (h2SNP = 0.035, p = 7.12x10-4) and the clinically predicted phenotype from VUMC (h2SNP = 0.046, p = 1.51x10-2). A significant genetic overlap was demonstrated between the two measures of suicide attempt in these independent samples through polygenic risk score analysis (t = 4.02, p = 5.75x10-5) and genetic correlation (rg = 1.073, SE = 0.36, p = 0.003). Finally, we show significant but incomplete genetic correlation of suicide attempt with insomnia (rg = 0.34 - 0.81) as well as several psychiatric disorders (rg = 0.26 - 0.79). This work demonstrates the contribution of common genetic variation to suicide attempt. It points to a genetic underpinning to clinically predicted risk of attempting suicide that is similar to the genetic profile from a patient reported outcome. Lastly, it presents an approach for using EHR data and clinical prediction to generate quantitative measures from binary phenotypes that improved power for our genetic study.

# COMPREHENSIVE ANALYSIS OF ANTIMALARIAL DRUG RESISTANCE IN MALARIA PARASITE USING PORTABLE DNA SEQUENCER

Lucky R Runtuwene[1], Junya Yamagishi[2], Josef S Tuda[3], Arthur E Mongan[3], Yutaka Suzuki[1]

[1]The University of Tokyo, Graduate School for Frontier Sciences, Kashiwa, Japan, [2]Hokkaido University, Center for Zoonosis Control, Hokkaido, Japan, [3]Sam Ratulangi University, Faculty of Medicine, Manado, Indonesia

Malaria persists because of the resistance to antimalarial drugs, among other factors. Since drug resistance is caused by mutation or structural variation in the genome of malaria parasites, genome sequencing might reveal the resistant trait that could cause medication failure. Malaria parasites have developed resistance to almost all of the antimalarial repertoire. The last effective drugs have shown a speed decline in eliminating resistant parasites. With the arrival of portable DNA sequencer, MinION, we have tried to genotype Plasmodium falciparum for the presence of drug-resistant mutations. Using four laboratory strains (3D7, Dd2, 7G8, and K1), we scanned for mutations in nine genes that are correlated with resistance to choloroquine, mefloquine, sulfadoxine-pyrimethamine, atovaquone, and artemisinin. We developed our own pipeline to call for mutations and despite of low MinION accuracy (74.3%), we could call for mutations confidently (13% mismatch, 9% deletion, and 4% insertion) using the obsolete R7.2 flow cell and this numbers greatly improved with the new R9.4 flow cell. We validated our results with Sanger and Illumina sequencing and found a reasonable 0.92 and 0.76 of precision and recall, respectively, even for the obsolete flow cell. We expanded this pipeline to analyze clinical samples from Indonesia, Thailand, and Vietnam and deduced the drug resistance status in all samples. We also have tried to develop a pipeline to detect copy number variation (CNV) in a gene responsible for piperaquine resistance. Because clinical samples regularly yield low DNA concentration, we first tried to sequence 3D7 laboratory strain with the low input sequencing kit to sequence the parasite whole genome. This process could generate sequencing tags that covered 82.7% of the genome with an average depth of 34 from 100 ng of high-molecular weight DNA. Using simulated reads, we tried to develop a pipeline to detect CNV of clinical samples and applied this pipeline to clinical samples from Indonesia.

# ASSESSING THE USAGE OF ALTERNATIVE TRANSCRIPTS IN HUMAN TISSUES FROM RNA-SEQ

Sergio Santos, Nuno A Fonseca, Mar Gonzalez-Porta, Alvis Brazma
University of Cambridge, EMBL-EBI, Cambridge, United Kingdom

Alternative splicing is an important step in gene expression regulation in eukaryotes, through which a single gene can lead to the expression of different proteins. With the advances in sequencing technology, we can now use RNA-seq data to identify which isoforms of each gene are being expressed in a specific condition, even though quantification of different isoforms remains a difficult problem. By quantifying the expression level of each isoform of a gene, we can better understand how prevalent this process is and how often a gene expresses different isoforms. It can also be evaluated if all isoforms of a gene are equally expressed or if there is one dominant isoform that is significantly more expressed than the others. Moreover, by applying this analysis to different tissues, it can be assessed if there are changes in splicing between different conditions and if such a change has a biological role.
The methods in this study were developed over a dataset of 32 normal human tissues. The results show that, although alternative splicing can lead to the expression of different transcripts of a gene, many genes have a n-fold dominant transcript – a transcript that is expressed at n times higher level than the second most expressed (Gonzàlez-Porta M et. al 2013). On average, 63% of the expressed genes in a given tissue have a 2-fold dominant transcript and 41% have a 5-fold dominant transcript.
It was observed that the dominant transcript of a gene tends to be the same across tissues, but there are cases where the dominant isoform switches between tissues, these cases are designated switch events. For a given pair of tissues, there are on average around thirty 2-fold switch events and just below four 5-fold switch events. The transcripts that switch share on average around 50% of the exons and the most common types of alternative splicing are alternative 3' selection (24% of the cases) and alternative 5' selection (21%).
To evaluate the conservation of the transcripts, the dominant transcripts were compared to APPRIS principal isoforms (Rodriguez JM et al. 2015). 69.2% of the 2-fold dominant transcripts and 81.1% of the 5-fold dominant transcripts are APPRIS principal isoforms. It was also observed that in 80% of the switches there are no protein domain changes.
These results show that in most cases, changes in alternative splicing do not change transcripts significantly, and respectively, the changes at the protein level are minor. This and similar observations (Tress ML et al. 2017; Reyes A et al. 2018) indicate that the main role alternative splicing may be different from generation of protein diversity.

References

Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. "Transcriptome Analysis of Human Tissues and Cell Lines Reveals One Dominant Transcript per Gene." Genome Biology 14.7 (2013).
Reyes A, Huber W. "Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues." Nucleic Acids Research 46.2 (2018).
Rodriguez JM, Carro A, Valencia A, and Tress ML. "APPRIS WebServer and WebServices." Nucleic Acids Research 43.W1 (2015).
Tress ML, Abascal F, Valencia A. "Alternative Splicing May Not Be the Key to Proteome Complexity." Trends in Biochemical Sciences 42.2 (2017).

# DIRECTLY MEASURING THE DYNAMICS OF THE HUMAN MUTATION RATE USING LARGE, MULTI-GENERATIONAL PEDIGREES

Thomas A Sasani, Brent S Pedersen, Mark F Leppert, Lisa M Baird, Aaron R Quinlan, Lynn B Jorde

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

Developing an accurate estimate of the human germline mutation rate is critical to our understanding of evolution and genetic disease. Early phylogenetic analyses inferred mutation rates from the observed sequence divergence between humans and related primate species at particular genes and pseudogenes. However, as whole-genome sequencing has become ubiquitous, these estimates have been refined using pedigree-based approaches. By identifying mutations present in offspring that are absent from their parents (*de novo* mutations, DNMs), it is possible to more accurately approximate the human germline mutation rate. Many studies have used whole genome sequencing to analyze the mutation rate in two-generation pedigrees, producing estimates nearly two-fold lower than those from phylogenetic comparison; however, both approaches may be imprecise. Phylogenetic estimates require numerous assumptions about ancestral population sizes and generation times, and two-generation pedigree approaches lack a means of biologically validating putative DNMs.

To obtain a precise, unbiased estimate of the human mutation rate, we performed whole-genome sequencing on blood-derived DNA from 34 of the original three-generation CEPH families from Utah. These families, which each contain grandparents (P0 generation), parents (F1), and their children (F2), are considerably larger than any used in prior estimates of the human mutation rate, and offer unparalleled power to detect DNM. With a median of 8 F2 individuals per pedigree, we were able to biologically validate *de novo* variants in the F1 generation by assessing their transmission to the F2. Using this dataset, we have generated a high-confidence estimate of the human mutation rate and observe a significant parental age effect on the rate of spontaneous mutation. To our knowledge, this study represents the first longitudinal analysis of the parental age effect on mutation rates within individual families. Additionally, we have identified recurrent *de novo* variants present in multiple F2 offspring, which are likely the result of mosaicism in the parental germline.

Finally, we have trained a classification model on the high-quality, transmitted *de novo* variants in our dataset, and used this model to identify DNMs in a large cohort from the Simons Foundation Autism Research Initiative (SFARI). Combining both the CEPH and SFARI callsets, we observe regional enrichment of *de novo* variants in the human genome, and will investigate the role of sequence context, as well as molecular processes like recombination and gene conversion, on the rate of human mutation.

# ANALYZING MULTI-OMIC INSTABILITY IN BREAST CANCER WITH NANOPORE SEQUENCING OF PATIENT-DERIVED ORGANOIDS

Michael C Schatz[1,2], Fritz J Sedlazeck[3], Sara Goodwin[1], Gayatri Arun[1], Isac Lee[2], Sam Kovaka[2], Michael Kirsche[2], Robert Wappel[1], Melissa Kramer[1], Karen Kostroff[4], David L Spector[1], Winston Timp[2], W Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [4]Northwell Health, Oncology, Lake Success, NY

Molecular instability is a hallmark of cancer, leading to widespread copy number variations, chromosomal fusions, aberrant methylation, and other alterations. Precisely characterizing these changes is essential for developing effective treatments, but previous technologies lacked resolution or are prohibitively expensive for routine use.

Addressing these challenges, we are sequencing multiple ER+/PR+/Her2- breast tumor+normal pairs of patient-derived organoids using the Oxford Nanopore GridION. Organoids recapitulate many aspects of the primary tumor better than cell lines, allowing for detailed molecular characterization and treatment response trials. They also enable ample collection of DNA, and by comparing tumor and normal samples from each patient we can isolate alterations specific to the cancer.

Using our optimized DNA extraction and library preparation techniques, we have sequenced more than 30x coverage of the first tumor organoid with read lengths up to 158,810bp long (n50=13,350bp), and the primary tissue to similar coverage depths. Using our read mapper NGMLR and structural variation detection algorithm Sniffles optimized for long reads, we have found 3,475 CNVs and 89 interchromosomal translocations specific to the tumor, most never previously reported. Interestingly, we also find structural variations in the patient's normal tissue within the important BRCA1 gene that cannot be detected using whole genome short read sequencing or targeted cancer panels. We also study differential methylation using the raw signal of the long reads, along with gene fusions and expression using short and long read RNA-sequencing. Jointly analyzed, we provide one of the most comprehensive analyses of cancer genomic architecture to date.

# DIRECT ESTIMATION OF MUTATIONS IN GREAT APES REVEALS SIGNIFICANT RECENT HUMAN SLOWDOWN IN THE YEARLY MUTATION RATE

Søren Besenbacher[1], Christina Hvilsom[2], Tomas Marques-Bonet[3], Thomas Mailund[4], Mikkel H Schierup[4]

[1]Aarhus University, Department of Molecular Medicine, Aarhus, Denmark, [2]Copenhagen Zoo, Genetics, Copenhagen, Denmark, [3]Universitat Pompeu Fabra, Comparative Genomics Lab, Barcelona, Denmark, [4]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

Several large studies have estimated the mutation rate in humans using whole genome sequencing of thousands of parent-offspring trios, almost exclusively of European descent. These studies have consistently estimated a yearly mutation rate of approximately 0.43x10-9. This rate is, however, markedly lower than prior estimates of ~1x10-9 per year from phylogenetic comparisons of the great apes. This discrepancy suggests either a slowdown over the past 20 million years or an inaccurate interpretation of the fossil record. We set out to estimate the current mutation rate in other great apes and used Illumina sequencing (~30-55X) of 1 chimpanzee extended trio (father, mother, child and grandchild), 2 gorilla extended trios and 1 orangutan trio. Additionally, we reanalyzed 6 chimpanzee trios from a previous study. Assuming that the relationship with maternal and paternal age is similar to humans we estimate a higher mutation rate in chimpanzee, gorilla and orangutan compared to humans by a factor of 1.6+/- 0.15, 1.5+/- 0.20 and 1.6+/- 0.21, respectively. These large differences in inferred rates contrast with the fact that the overall great apes phylogeny almost adheres to a molecular clock with e.g. the chimpanzee branch only being 2% longer than the human branch and the gorilla branch 6% longer than the human branch. We interpret this as evidence that the apparent slowdown by ~35% on the human branch since separation with the chimpanzee has occurred recently, ie. within the last half million years. Thus, human-Neanderthal divergence time may also be overestimated if using the present human mutation rate for calibration. Using our new great apes yearly mutation rate estimates we re-estimate great apes genomic and species divergence times and find a human-chimpanzee species divergence time of 6.3 my, of human-gorilla of 8.2 my and of human-orangutan of 14.1 my. We discuss these results in relation to the fossil record as well as possible reasons for the apparent recent mutation rate slowdown in humans.

# SOMETHING OLD, SOMETHING NEW: RESOURCES FOR ASSEMBLY CURATION AND EVALUATION

Valerie A Schneider[1], Kerstin Howe[2], Tina Graves-Lindsay[3], Paul Flicek[4]

[1]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, [2]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [3]The McDonnell Genome Institute, Washington University, St. Louis, MO, [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

The Genome Reference Consortium (GRC) curates the human and mouse reference genome assemblies to ensure they continue to support the diverse analysis needs of the basic and clinical research communities. These two assemblies, GRCh38 and GRCm38, remain among the highest quality vertebrate genomes ever produced. Historically, their curation involved the use of Sanger-sequenced genomic clones to close remaining gaps and assembly errors, and to add population or strain-specific diversity. The original clone resources used for such curation have largely been exhausted, however, and in today's era of next generation sequencing and whole genome assembly, the new production of such resources is scarce. Furthermore, analyses have demonstrated that many of the remaining assembly issues are recalcitrant to resolution with clone-based sequences. In response, the GRC has been evaluating and employing other resources and new techniques for reference improvement. We will present recent assembly curation efforts in both species, highlighting commonalities and differences in remaining issues, resources and consortium goals. In addition, we will discuss the impact introduction of WGS sequences into a clone-based assembly has on quality metrics. The GRC is also committed to transparency in its curation efforts and communication in matters of assembly quality. We will present the suite of publicly accessible GRC resources for assembly evaluation as well as recent updates to NCBI tools, including the Genome Data Viewer (GDV), for visualization and analysis of genome assemblies.

# HEATITUP IDENTIFIES ITD STRUCTURAL FEATURES DETERMINING AML PATIENT RESPONSE TO *FLT3* INHIBITORS.

Gregory W Schwartz[1,2], Bryan Manning[3], Yeqiao Zhou[1,2], Priya Velu[1], Ashkan Bigdeli[4], Rachel Astles[3], Anne W Lehman[3], Jennifer Morrissette[1,4], Alexander E Perl[3], Martin Carroll[3], Robert B Faryabi[1,2,4,5]

[1]University of Pennsylvania, Department of Pathology and Laboratory Medicine, Philadelphia, PA, [2]University of Pennsylvania, Abramson Family Cancer Research Institute, Philadelphia, PA, [3]University of Pennsylvania, Division of Hematology and Oncology, Philadelphia, PA, [4]University of Pennsylvania, Center for Personalized Diagnostics of the Perelman School of Medicine, Philadelphia, PA, [5]University of Pennsylvania, Department of Cancer Biology, Philadelphia, PA

Acute myeloid leukemia (AML) is a rapidly progressing cancer of myeloid cells. Approximately 25% of AML cases carry an internal tandem duplication (ITD) mutation. ITD mutations are a diverse class of mutations that results in an insertion of a duplicated DNA segment near the original segment. ITDs within the FMS-like tyrosine kinase 3 (*FLT3*) receptor are associated with poor prognosis in AML. While several *FLT3* inhibitors (FLT3i) are in clinical trials for targeted therapy of high-risk *FLT3*-ITD-positive AML, the variability in overall survival suggests that the presence of an *FLT3*-ITD is not the only determinant of response. Motivated by the heterogeneity of *FLT3*-ITD mutations, we sought to investigate the effects of *FLT3*-ITD structural features on response to FLT3i treatment. In order to accurately and efficiently identify ITD structural features, we developed the HeatITup (HEAT diffusion for Internal Tandem dUPlication) algorithm and software package. HeatITup enabled categorization of ITD mutations into two new classes of "typical" and "atypical". While the former contains only segments endogenous to the reference sequence, atypical ITDs have insertions where part of the inserted segment is composed of nucleotides exogenous to the wild-type locus. We applied HeatITup to the University of Pennsylvania cohort of de novo and relapsed AML in order to identify and classify *FLT3*-ITD mutations and investigated their correlation with overall survival of FLT3i-treated *FLT3*-ITD-positive AML patients. Our data demonstrated that patients with atypical *FLT3*-ITDs had significantly lower survival rate than those with typical mutations in both de novo and relapsed diseases, while controlling for other clinical variables. Furthermore, we recapitulated these results in the TCGA AML cohort treated with induction chemotherapy. These findings highlight the role of structural features of complex mutations such as ITDs in progressing towards personalized treatment for AML patients.

# STRUCTURAL VARIATION IN 35,600 MULTI-ETHNIC HUMAN GENOMES AND ITS IMPLICATION FOR THE GENETIC ARCHITECTURE OF HEALTH AND DISEASE.

Fritz J Sedlazeck*[1], Goo Jun*[2], Bing Yu[2], Olga Krasheninina[1], Han Chen[2], Andrew Carroll[3], Adam J Mansfield[1], Ziad M Khan[1], Vipin K Menon[1], Samantha Zarate[3], Harsha Doddapaneni[1], Ginger Metcalf[1], Donna Muzny[1], William J Salerno[1], Richard Gibbs[1], Eric Boerwinkle[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]University of Texas Health Science Center at Houston, Human Genetics Center, Houston, TX, [3]DNAnexus, Mountain View, CA

The role of structural variants (SV) (deletions, insertions, rearrangements and copy number variation) in Mendelian disease is well established, but their impact on common diseases (e.g. heart disease and diabetes) has not been elucidated because of the complexities of identifying and genotyping SVs in large numbers of well-phenotyped ethnically diverse individuals.

We characterized the underlying architecture (single nucleotide variation, SV and haplotype scaffolds) of 35,600 genomes across 50 ethnic groups. Multiple discovery and genotyping methods were combined, considering sample relatedness, ethnic diversity and sequencing data heterogeneity(e.g. platform and depth), to provide a comprehensive catalog of variation. In addition, we established a novel cost-efficient strategy to apply long-read sequencing to SV validation, complex break-point characterization and long-range haplotyping. These methods resulted in more than 500,000 putative SVs with corroborating genotype evidence. The resulting data resource is the largest comprehensive and harmonized variation set on samples of deeply-phenotyped individuals available to date. As expected, most variants were rare (<0.01 AF). There were substantial amounts of low frequency, ethnicity-specific and common variation, fueling opportunity for novel phenotype-related discovery.

Within the three major US ethnic groups, European-, African- and Hispanic-Americans, each individual was measured and analyzed for >20 traits including anthropometric measurements, blood pressure, lipid levels, kidney function, glucose metabolism, pulmonary function and prevalent disease status (e.g. cardiovascular disease, diabetes, and lung disease) along with a list of demographic and environmental factors (e.g. smoking and lipid lowering drug use). Novel analytic methods show significant bias of standard SNP-based GWAS results due to lack of inclusion of large SVs. Standard analysis methods also reveal novel discoveries attributable to the identified structural variants being related to multiple chronic disease phenotypes (e.g. blood lipid levels and hematocrit).

# BRAIN REGION-SPECIFIC DNA METHYLATION CORRELATES OF SCHIZOPHRENIA AND ITS GENETIC AND DEVELOPMENTAL RISK

Stephen A Semick, Ran Tao, Joo Heon Shin, Richard E Straub, Emily E Burke, Leonardo Collado-Torres, BrainSeq Consortium, Thomas M Hyde, Joel E Kleinman, Daniel R Weinberger, Andrew E Jaffe

Lieber Institute for Brain Development, Data Sciences, Baltimore, MD

**Introduction**: DNA methylation (DNAm) is an epigenetic mark that plays a crucial role in the development and normal function of the human brain. We have previously profiled DNAm levels in the dorsolateral prefrontal cortex (DLPFC) and found widespread associations with brain development, nearby genetic variation, and schizophrenia diagnosis. However, the effects of these factors on DNAm levels in another brain region implicated in schizophrenia— the hippocampus— and the overall brain regional specificity of these associations, are unknown.
**Methods**: We generated DNAm profiles in DNA extracted from homogenate hippocampal tissue in 238 non-psychiatric controls and 109 patients with schizophrenia using the lllumina Infinium HumanMethylation450K ("Illumina 450k") BeadChip. We combined these data with existing Illumina 450k data on the DLPFC and jointly processed the microarrays (N=694). We used linear modeling to identify associations between DNAm levels and age, genotype, and schizophrenia diagnosis, both within and across brain regions.
**Results**: The majority of tested CpGs were differentially methylated between the two adult brain regions (57.8%, N=242,581; FDR<1%) and were enriched for genes related to neurodevelopment. Epigenetic age acceleration, a measure of the rate of biological aging, was partially correlated between brain regions ($r=0.37$, $p<5.4E-13$). We found DNAm signatures of cell-type composition of hippocampus to be distinct from DLPFC; notably, hippocampus contains ~12% fewer putative neuronal cells than DLPFC ($p<2.2E-16$). Nevertheless, differences in cell-type composition cannot fully explain most DNAm differences between brain regions. We further identified DNAm differences between schizophrenia patients and non-psychiatric controls (3,510 CpGs), changes over the course of development (86,572 CpGs), and methylation trait loci effects (meQTLs; 7,676 CpGs) which were brain region dependent. Genetic risk loci from genome wide association studies were enriched for meQTLs, including those linked to schizophrenia where region-specific and interaction meQTLs were present.
**Discussion**: We have shown that human DLPFC and hippocampus have distinct DNAm profiles and that development, genetic variation, and neuropsychiatric illness can have differential effects in the hippocampus than in DLPFC. These findings reinforce the need to study multiple different brain regions and the interplay between them to better understand the role of DNAm in the physiology and pathophysiology of the human brain.

# IDENTIFICATION OF POTENTIAL REGULATORY MUTATIONS USING MULTI-OMICS ANALYSIS AND HAPLOTYPING OF LUNG ADENOCARCINOMA CELL LINES.

Sarun Sereewattanawoot[1], Ayako Suzuki[2], Masahide Seki[1], Yoshitaka Sakamoto[3], Takashi Kohno[4], Sumio Sugano[1], Katsuya Tsuchihara[2], Yutaka Suzuki[1]

[1]the University of Tokyo, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, Chiba, Japan, [2]National Cancer Center, Division of Translational Genomics, Exploratory Oncology Research and Clinical Trial Center, Chiba, Japan, [3]the University of Tokyo, Department of Bioinformatics and Systems Biology, Faculty of Science, Tokyo, Japan, [4]National Cancer Center Research Institute, Division of Genome Biology, Tokyo, Japan

The functional relevancy of mutations occurring in the regulatory regions in cancers remains mostly elusive. Here, we identified and analyzed regulatory mutations having transcriptional consequences in lung adenocarcinoma-derived cell lines. We phased the mutations in the regulatory regions to the downstream heterozygous SNPs in the coding regions and examined whether the ChIP-Seq variant tags of the regulatory SNVs and the RNA-Seq variant tags of their target transcripts showed biased frequency between the mutant and reference alleles. We identified 137 potential regulatory mutations affecting the transcriptional regulation of 146 RefSeq transcripts with at least 84 SNVs that create and/or disrupt potential transcription factor binding sites. For example, in the regulatory region of NFATC1 gene, a novel and active binding site for the ETS transcription factor family was created and experimentally validated. Further examination revealed that 31 of these disruptions were presented in clinical lung adenocarcinoma samples and were associated with prognosis of patients.

# TRANSFER LEARNING APPROACHES FOR ROBUST FINE-MAPPING OF PUTATIVE CAUSAL REGULATORY VARIANTS ASSOCIATED WITH COLORECTAL CANCER

Anna Shcherbina[1], Stephanie Bien[2], Jeroen R Huyghe[2], Alina Saiakhova[3], Stephanie Schmidt[4], David Conti[5], Tabitha Harrison[2], Flora Qu[2], Li Hsu[2], Michael Wainberg[7], Michael Bassik[8], Graham Casey[5], Steven Gruber[5], Ulrike Peters[2], Peter Scacheri[3], Anshul Kundaje[7]

[1]Stanford University, Stanford, CA, [2]Fred Hutchinson Cancer Research Center, WA, [3]Case Western Reserve University, OH, [4]Moffitt Cancer Center, FL, [5]University of Southern California, CA, [6]University of Virginia, VA, [7]Stanford University, CA, [8]Stanford University, CA

We present an integrative approach for context-specific fine mapping and interpretation of non-coding disease-associated variants based on deep learning models of regulatory DNA sequence. We used this approach to analyze over 50 genome-wide significant loci associated with colorectal cancer (CRC) based on a meta-analysis of 125,218 CRC cases and controls from three consortia: Colon Cancer Family Registry, Colorectal Cancer Transdisciplinary Study, and Genetics and Epidemiology of Colorectal Cancer Consortium. First, we trained novel multi-task, convolutional-recurrent neural networks (NN) to accurately map 1Kb bins of DNA sequence across the genome to corresponding DNase-seq (DHS) profiles across 193 diverse ENCODE+REMC cellular contexts. Next, we fine-tuned the reference NNs on DHS and H3K27ac profiles in 21 primary CRC tumor samples and mucosa from 6 healthy controls. This transfer learning approach resulted in significant accuracy gains relative to training only on the CRC relevant tissues. We used a novel feature attribution method for neural networks called DeepLIFT to efficiently infer sample-specific regulatory potential (predictive importance scores) of every nucleotide in 1Kb bins centered at 23,727 variants across all the associated loci, which include 3910 variants in the 99% credible sets of causal variants. Variants with high, robust DeepLIFT scores across multiple samples, bootstrapped models and heterogeneous genomic background sets, highlighting putative functional regulatory variants, were further interrogated using in-silico mutagenesis (ISM) to estimate the signed influence (effect size) of all alleles on local chromatin state. DeepLIFT profiles of nucleotides surrounding high scoring variants further highlighted sequence features such as transcription factor (TF) motifs harboring these variants. Our approach exhibited high specificity. Within each GWAS locus, across 100s of 99% credible set variants and 1000s of background SNPs, only 1-5 SNPs exhibited high and statistically significant functional scores. Several high scoring SNPs overlapped tissue-specific enhancers in colorectal mucosa. Further, these variants were often found at flanks of DHSs directly disrupting, creating or indirectly influencing motifs of lineage-specific TFs. High confidence candidates are being tested using high-throughput CRISPR screens and reporter assays in CRC lines. Our framework can easily generalize to other disease phenotypes and to score non-coding rare variants and somatic mutations including indels and short SVs.

# SMALL MOLECULE MODULATION OF THE D4Z4 LOCUS TRANSCRIPTIONAL ACTIVITY IN FSHD

Ning Shen, Alejandro Rojas, Pete Rahl, Owen Wallace, Angela Cacace, Aaron Chang

Fulcrum Therapeutics, Cambridge, MA

Facioscapulohumeral dystrophy (FSHD) is one of the most common muscular dystrophies that manifests typically in the facial and shoulder muscles. This genetic disease is caused by the loss of repression at the D4Z4 macrosatellite repeats in the subtelomeric region of chromosome 4 (4q35), a region that shares high homology with chromosome 10. In the last decade, great progress has been made in the understanding of the genetics behind this pathology and two distinct forms have been described. FSHD1, caused by contraction in the number of D4Z4 repeats (<8) in ch4, and FSHD2, caused by mutations in the SMCHD1 gene. Both forms of FSHD result in the aberrant expression of the homeobox transcription factor DUX4 in the muscle, whose sequence is codified in each D4Z4 repeat. Muscle expression of this transcription factor, which plays a critical role in defining stages of early development, results in cell death and muscle fiber loss. While some progress has been made towards understanding myogenic signals driving DUX4 expression, the factors involved in the transcriptional activation of DUX4 are still largely unknown. Performing a medium throughput small molecule screen using patient-derived FSHD myotubes, we have identified targets that participate in the regulation of expression of DUX4. To further understand the mechanisms of DUX4 regulation in the context of the disease locus, we performed RNA-seq and ChIP-seq profiling in cells treated with small molecules that modulate the activity of this target. This analysis demonstrated a specific effect on DUX4 expression and its downstream program. Out of ~400 differentially expressed genes after this treatment, over 200 were found to be directly regulated by DUX4 and participate in early embryo functions. These genes are also not normally expressed in the wild-type muscle. We are now characterizing the activity of DUX4 and the specific effects of small molecules by examining changes in regulatory elements genome-wide using H3K27Ac ChIP-seq data. The identification of targets and the characterization of cis-elements relevant for FSHD are key elements of our approach to develop therapeutics in a disease where no therapies are available for patients

# NEW STRATEGIES TOWARD SCALING UP EPISTASIS ANALYSIS ON LARGE-SCALE GENOMIC DATASETS

Jia Wen, Colby Ford, Daniel Janies, Xinghua Shi

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

Epistasis reflects the joint effect of more than one genes or variants on a phenotype. Epistasis is considered as an important genetic component for understanding complex phenotypes, however, it is challenging to identify epistasis in large-scale genomic data. This challenge contains two interlinked constraints. 1), the epistasis analysis on the high dimensional genomic data typically induces an over-saturated model that in turn demands efficient statistical methods to solve the model. 2) Even once the appropriate model and data are specified the solution requires intensive computing time to identify epistasis among two or more genes or variants. In this study, we present two strategies to scale up epistasis analysis using the empirical Bayesian Elastic Net (EBEN) method. First, we introduce a matrix strategy that pre-computes the correlation matrix using matrix multiplication. This strategy narrows down the search space and thus accelerate the modeling process of epistasis analysis. Next, we develop a parallelized version of the EBEN package (parEBEN) that affords multi-fold speed up in comparison with the classical EBEN method. Thus parEBEN enables application of the regression model to be applied to larger and more complex genomic datasets. As a use case, we applied these two strategies on a yeast cross dataset. Our results indicated that we can identify a set of marginal and pairwise epistatic SNPs associated with quantitative traits relevant to yeast fitness.

# iMETHYL: AN INTEGRATIVE HUMAN MULTI-OMICS QTL DATABASE FOR 3 BLOOD CELL TYPES

Shohei Komaki[1], Yuh Shiwa[1], Ryohei Furukawa[1], Tsuyoshi Hachiya[1], Hideki Ohmomo[1], Yoichi Sutoh[1], Mamoru Satoh[1], Kenji Sobue[2], Makoto Sasaki[3], Atsushi Shimizu[1]

[1]Iwate Medical University, Division of Biomedical Information Analysis, Iwate Tohoku Medical Megabank Organization, Yahaba, Japan, [2]Iwate Medical University, Dean, Morioka, Japan, [3]Iwate Medical University, Iwate Tohoku Medical Megabank Organization, Yahaba, Japan

**PURPOSE:**
In the past decade, genome- and epigenome-wide association studies on various human traits have revealed numerous interactions between genomic and epigenomic variations and the traits of interest. Presently, to understand further the biological consequences of the genomic and epigenomic variations associated with phenotypic variations, the importance of the integrative analysis of multi-omics data is demonstrated. In 2016, we established the integrative DNA methylation database, iMETHYL, comprising a comprehensive dataset on genomic variation (~9 million SNVs), gene expression (≥ 14,000 genes), and DNA methylation (~24 million CpGs) each for $CD4^+$ T lymphocytes (CD4Ts), monocytes, and neutrophils of approximately 100 Japanese subjects. Herein, we report the release of additional information in iMETHYL, and results of cell type-specific cis-eQTL, cis-eQTM, and cis-mQTL analyses based on the multi-omics datasets for ~ 100 Japanese.
**METHODS:**
For CD4Ts, monocytes, and neutrophils, genotypes of 8,951,822, 8,945,669, and 8,792,880 SNVs, expression levels of 18,894, 16,789, and 14,957 genes, and DNA methylation (DNAm) levels of 24,037,522, 23,941,826, and 25,395,185 CpG sites, respectively, are available in iMETHYL. In the eQTL analyses, pairwise associations between the expression level of each gene and genotype of its neighboring SNV (within 1 Mbp from TSS) were tested. In the eQTM and mQTL analyses, associations between expression and DNAm (within 1 Mbp) and between genotype and DNAm (within 20 kbp), respectively, were tested. In each association test, a simple linear regression model was applied.
**RESULTS:**
We performed 119,454,345, 106,067,569, and 92,860,381 eQTL analyses, 491,353,343, 439,889,100, and 388,571,061 eQTM analyses, and 3,285,524,083, 3,270,268,912, and 3,412,466,876 mQTL analyses for CD4Ts, monocytes, and neutrophils, respectively. The results of all pairwise association tests, including *P* values, coefficients, and r-squares were uploaded, and could be browsed on the iMETHYL genome browser along with allele frequencies, gene expression levels, and DNAm levels.
**CONCLUSION:**
We have upgraded our iMETHYL genome browser by adding the results of the multi-omics eQTL/eQTM/mQTL analyses. Each cell type showed different eQTL/eQTM/mQTL trends, representing the efficiency of cell type-specific multi-omics analysis. The iMETHYL multi-omics browser facilitates the understanding of the consequences of genomic/epigenomic variations affecting phenotypic variations, regarding the biological functions of each cell type.
**URL:** http://imethyl.iwate-megabank.org/

**References**

1. Hachiya T. et al. *Sci. Rep.* 2017, **7**: 16147 | DOI:10.1038/s41598-017-16493-0
2. Komaki S. et al. *Hum. Genome Var.* 2018, *in press.*

# THE PREDICTIVE CAPACITY OF REGULATORY ELEMENTS TO IMPACT MAMMALIAN PHENOTYPES

Siddharth Sethi[1], Ilya Vorontsov[2], Ivan Kulakovskiy[2], Vsevolod Makeev[2], Kenneth Condon[1], Michelle Simon[1], Ann-Marie Mallon[1]

[1]MRC Harwell Institute, Biocomputing, Didcot, United Kingdom, [2]Engelhardt Institute Of Molecular Biology, Bioinformatics, Moscow, Russia

Densely spaced clusters of active enhancers called super-enhancers have recently been found to control and maintain the cell identity in the mammalian genome. Sequence variants are enriched in super-enhancers of disease-relevant cell types compared to typical-enhancers and have been proposed to be involved in development and disease. However, the structure and function of super-enhancers is controversial and not completely understood. In addition, the extent of the relationship and mechanisms between super-enhancers and the corresponding phenotypes of genes they control remain unclear.

Using an in silico integrated approach, we have profiled genome-wide enhancer activity and produced a catalogue of tissue-specific enhancers and super-enhancers across 22 mouse tissues and cell lines, that regulate key cell identity genes. Like previous studies, we demonstrate that super-enhancers significantly upregulate tissue-specific expression, suggesting super-enhancer associated genes may have exclusive yet important roles in lineage-specific functions. However, we found gene regulation is not confined to super-enhancers, there are 14 times more typical-enhancers in the genome and they contribute to 28% of total tissue-specific expression while super-enhancers contribute only 10%. By analysing phenotype and interaction data, we identified no difference in the functional pathways regulated by super-enhancers and typical enhancers. Furthermore, we found there is no significant difference in the phenotypic expression levels between super and typical enhancer associated genes, suggesting that both super and typical enhancers equally impact mammalian phenotypes.

To further understand the relationship between regulatory elements and mammalian phenotypes, we implemented a random forest classifier to assess the capacity of tissue-specific regulatory elements, amongst other functional datasets, to predict mammalian phenotypes. We observe that the capability of regulatory elements to predict gene-phenotype associations is relatively low, while protein-protein interaction and expression data significantly improve the predictions. Overall, this study highlights the significant difference in expression between super-enhancer and typical-enhancer associated genes, while their phenotypic roles are unchanged. However, a lower predictive capacity of regulatory elements to model mammalian phenotypes suggests other molecular pathways may have a greater impact on the manifested phenotype.

# A STUDY OF ETHNIC POLYMORPHIC COPY NUMBER VARIATIONS IN THE ISRAELI POPULATION

Pola Smirin-Yosef[1,5], Sarit Kahana[3], Idit Maya[3], Shiri Yacobson[3], Doron Levi[1], Elisheva Biton[1], Danny Baranes[1], Lina Basel-Vanagaite[2,3,4,5], Mali Salmon-Divon[1]

[1]Ariel University, Genomic Bioinformatics Laboratory, Molecular Biology Department, Ariel, Israel, [2]Tel Aviv University, Sackler Faculty of Medicine, Tel Aviv, Israel, [3]Raphael Recanati Genetics Institute, Rabin Medical Center, Beilinson Campus, Petah Tikva, Israel, [4]Schneider Children's Medical Center of Israel, Pediatric Genetics Unit, Petah Tikva, Israel, [5]Felsenstein Medical Research Center, Rabin Medical Center, Petah Tikva, Israel

The Israeli population is composed of a collection of diverse ethnic groups. Each group shares specific genetic variations that passed from its common ancestors throughout the generations. Together with pathogenic events, non-pathogenic polymorphism happen to occur in ancestors, subsequently spread into the restricted genomic pool of its descendants. Chromosomal Microarray Array (CMA) has had a high impact in clinical diagnostics, leading to the discovery of new genomic disorders, and has become an indispensable tool for routine molecular and cytogenetic testing. CMA is a first line diagnostic test for individuals with developmental disabilities, dysmorphic features and congenital malformations as well as fetuses with congenital malformations and abnormal growth.

Here we apply a data mining approach on the results of CMA testing performed at the Raphael Recanati Genetic Institute, in order to characterize ethnic-specific polymorphism. The data resource contains around 10,000 tests from individuals, fetuses with clinical abnormalities, and in fetuses from low-risk pregnancies.

The use of an extracted ethnicity-based genetic information, in order to detect ethnic-specific Copy Number Variations (CNVs) polymorphism in the Israeli population will allow geneticists to distinguish between relevant pathogenic genomic aberrations from benign ethnicity-related variations.

# CHO-OMICS REVIEW: THE IMPACT OF CURRENT AND EMERGING TECHNOLOGIES ON CHINESE HAMSTER OVARY BASED BIOPRODUCTION

Gino Stolfa, Matthew T Smonskey, Ryan Boniface, Anna-Barbara Hachmann, Paul Gulde, Atul D Joshi, Anson P Pierce, Scott J Jacobia, Andrew Campbell

Thermo Fisher Scientific, Bioproduction R&D, Grand Island, NY

CHO cells are the most prevalent platform for modern bio-therapeutic production. Currently, there are several CHO cell lines used in bioproduction with distinct characteristics and unique genotypes and phenotypes. These differences limit advances in productivity and quality that can be achieved by the most common approaches to bioprocess optimization and cell line engineering. Incorporating omics-based approaches into current bioproduction processes will complement traditional methodologies to maximize gains from CHO engineering and bioprocess improvements. In order to highlight the utility of omics technologies in CHO bioproduction, the authors discuss current applications as well as limitations of genomics, transcriptomics, proteomics, metabolomics, lipidomics, fluxomics, glycomics, and multi-omics approaches and the potential they hold for the future of bioproduction. Multiple omics approaches are currently being used to improve CHO bioprocesses; however, the application of these technologies is still limited. As more CHO-omic datasets become available and integrated into systems models, the authors expect significant gains in product yield and quality. While individual omics technologies provide incremental improvements in bioproduction, the authors will likely see the most significant gains by applying multi-omics and systems biology approaches to individual CHO cell lines.

# DICHOTOMY IN REDUNDANT ENHANCERS REFLECTS DIFFERENCES IN SEQUENCE ENCRYPTION AND GENE REGULATORY MECHANISMS

Wei Song, Ivan Ovcharenko

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

The regulatory landscape of a gene locus often consists of several functionally redundant enhancers within the same locus to provide phenotypic robustness and evolutionary stability. We observed that the DNA sequence signature differs between redundant enhancers and a single enhancer in a locus maintaining the same function as redundant ones, suggesting differences in the regulatory mechanisms of these two classes of enhancers. Furthermore, there is a small subset of redundant enhancers, which contains sequence encryption similar to single enhancer (named singleton-like enhancers), but not the rest of redundant enhancers. Our results show that single and singleton-like enhancers are more evolutionary conserved and tend to be located in closer proximity to the nearest gene than regular redundant enhancers. Single and singleton-like enhancers feature more 3D chromatin contacts with transcription start sites of genes, suggesting their role as primary activators of gene expression. The fact that singleton-like enhancers form many interactions with nearby enhancers reveals their function as intermediate catalysts of gene regulation established between genes and the regular redundant enhancers.

# SEX DIFFERENCES AT THE MOLECULAR LEVEL: LESSONS FROM THE HUMAN TRANSCRIPTOME

Meritxell Oliva[1], Eric Gamazon[2], Ferran Reverter[3], Diego Garrido[3], Valentin Wucher[3], Francois Aguet[4], Bruna Balliu[5], Princy Parsana[6], GTEx Consortium[4], Stephen Montgomery[5], Alexis Battle[6], Kristin Ardlie[4], Roderic Guigo[3], Barbara Engelhardt[7], Barbara E Stranger[1]

[1]University of Chicago, Chicago, IL, [2]Vanderbilt University, Nashville, TN, [3]CRG, Barcelona, Spain, [4]Broad Institute, Cambridge, MA, [5]Stanford University, Palo Alto, CA, [6]Johns Hopkins University, Baltimore, MD, [7]Princeton University, Princeton, NJ

The factors contributing to sex-differentiated traits include genetics, hormones, and response to the exogenous and endogenous environment. Here, we focus on sex differences in gene regulation as a fundamental process by which a cell, tissue, or organism can generate phenotypic diversity from a genome that is common between males and females except for sex chromosomes. We have characterized multiple modalities of sex differences in the human transcriptome within and between tissues of the NIH Genotype-Tissue Expression (GTEx) project (17,382 RNA-seq samples from 838 individuals, v8 data release). We identify significant (FDR<0.05) sex-biased gene expression levels and splicing of autosomal and sex chromosome genes, with enrichment for genes on the X chromosome. The extent of transcriptome sex differences is highly variable across tissues, with breast exhibiting the most affected genes. Differentially expressed and spliced genes include those previously implicated in human phenotypes, many of which exhibit sex differences in prevalence or clinical disease presentation. We identify sex-differentiated regulatory networks and genetic regulation of gene expression and splicing by characterizing sex-biased expression QTLs (eQTLs), splicing QTLs, and sex differences in the heritability of gene expression within and across tissues. We identify significant (FDR<0.1) SNP-by-sex interaction cis-eQTLs for over 350 genes, 74% of which derive from a single tissue. Discovery of sex-biased eQTLs is positively correlated with sample size and negatively correlated with ratio of male to female donors. We describe functional enrichment analyses of differentially expressed, spliced, and genetically-regulated genes. We integrate our genetic analyses with large-scale Biobanks including the UK biobank and the electronic health records of BioVU, to identify genetically determined expression that varies across the sexes in these two biobanks and are associated with human phenotypes. Collectively, our integrative multi-tissue analyses provide the most comprehensive characterization to-date of human sex differences at the transcriptome level, with important implications for studies of complex traits.

# THE ENCODE ANNOTATION PIPELINE: SOFTWARE ARCHITECTURE FOR REPRODUCIBLE ANALYSES OF CHIP-SEQ, RNA-SEQ, DNASE-SEQ, ATAC-SEQ, HIC,CHIA-PET, AND WHOLE-GENOME BISULFITE EXPERIMENTS

J Seth Strattan[1], Ulugbek K Baymuradov[1], Timothy R Dreszer[1], Neva C Durand[2], Idan Gabdank[1], Ben C Hitz[1], Otto A Jolanki[1], Jin W Lee[1], Esther T Chan[1], Jason A Hilton[1], Anshul Kundaje[1], J Michael Cherry[1]

[1]Stanford University, Genetics, Stanford, CA, [2]Baylor College of Medicine, Genetics, Houston, TX

The ground-level annotations ENCODE applies to the epigenome and the transcriptome are produced by defined computational pipelines that anyone can use. 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) are deploying the pipelines to a containerized environment to achieve reproducible results across multiple platforms. Because the exact software pipelines the ENCODE DCC uses are available for anyone to run, researchers can produce analysis results and quality-control metrics from their own data that are directly comparable to ENCODE's annotations. Pipelines are available or in development for the analysis of ChIP-seq, RNA-seq, DNase-seq, ATAC-seq, HiC, ChIA-PET, and whole-genome bisulfite experiments. The pipelines are open-source, easy to use, and require minimal pre-requisites.

The ENCODE DCC codebase is at https://github.com/ENCODE-DCC
ENCODE analyses are distributed through the ENCODE Portal at https://www.encodeproject.org/

# A STATISTICAL METHOD ENABLES TO ESTIMATE *PERSONAL DIPLOID METHYLOME AND TRANSCRIPTOME* WITH LONG READS

Yuta Suzuki[1], Yunhao Wang[2], Kin Fai Au[2], Shinichi Morishita[1]

[1]The University of Tokyo, Department of Computational Biology and Medical Sciences, GSFS, Kashiwa, Japan, [2]University of Iowa, Department of Internal Medicine, Iowa city, IA

It is challenging to obtain personal diploid methylomes, CpG methylome pairs of homologous chromosomes that are distinguishable with respect to phased heterozygous variants (PHVs). A CpG site can be phased and assigned to its originating haplotype if its proximal PHV is found on the identical read. However, only 11.3%~12.3% of CpG sites lay within 100 bp from nearest PHVs whereas 72.2%~81.3% within 8 kbp in typical personal genomes, demanding long reads having both CpG sites and informative PHVs. Single molecule real-time (SMRT) sequencing likely provides a unique solution because of its ability to output long reads with CpG methylation information. For accurate phasing of CpG sites, however, a serious concern is whether reliable PHVs are available in erroneous SMRT reads with an error rate of ~15%. We here propose a statistical model that identifies reliable PHVs and reduces the error rate of phasing CpG site down to 1%, thereby calling CpG hypomethylation in each haplotype with >90% precision and sensitivity. Thus, we obtained personal diploid methylomes for two personal genomes with long reads.

Our statistical method allowed us to reconfirm that many CpG islands with allele-specific methylation (ASM) were associated with known imprinted genes and resided in the transcribed or repressed regions reflecting allele-specific expression (ASE). Combined with personal diploid transcriptomes obtained by assigning RNA-seq reads with PHVs to their alleles, we revealed complex ASM events controlling ASE of alternative isoforms within genes (e.g., ZNF331). For GNAS complex locus, which is known for a combination of maternally, paternally, or biallelically expressed isoforms, we observed ASM pattern almost perfectly reflecting their respective expression status, confirming correlation between ASM and ASE statuses. While the scarcity of exons with PHVs (only 10.9% of all exons) often hinders associating RNA-seq reads with their alleles, correlation of ASM/ASE demonstrates the potential utility of ASM as complements for ASE. These findings highlight the utility of long, CpG-methylation-sensitive SMRT reads in epigenetics study to construct comprehensive personal diploid methylomes and transcriptomes.

# MONITORING CHANGES IN GUT MICROBIOMES THROUGHOUT THE LIFE IN MICE

Lena Takayasu[1], Wataru Suda[1], Eiichiro Watanabe[1], Taichi Umeyama[1], Masahira Hattori[2,1]

[1]RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan, [2]Waseda University, Graduate School of Advanced Science and Engineering, Tokyo, Japan

There are age differences in the community structure of the human gut microbiome, but it is largely unknown about its consecutive alterations through an individual life from birth to death. Although the study for humans is not realistic, it is feasible to investigate alterations in the gut microbiome through the life in mice. We started to rear a pair of parent SPF mice more than two years ago, and have been collecting the fecal samples every two to seven days from the parent mice and their offspring that were born during this study. At present, fecal samples collected so far include those just after birth of the offspring and just before death of a few mice by some reason.
We sequenced the fecal DNA samples to obtain the 16S rRNA gene sequence data for evaluation of the microbial diversity and abundance in the gut microbiomes. The comparative analysis of the offspring samples revealed drastic changes in the gut microbiome structure in 0 to 20 days after birth, and the relatively small changes in 20 days to 3 months after birth. The analysis is in progress, and changes in the gut microbiome in mice by aging and by pregnancy and disease occurred during this study will also be presented.

# GRAMENE: UNIFYING COMPARATIVE GENOMICS AND PATHWAY RESOURCES FOR PLANT COMMUNITIES

Marcela K Tello-Ruiz[1], Sharon Wei[1], Andrew Olson[1], Justin Preece[2], Parul Gupta[2], Sushma Naithani[2], Joshua Stein[1], Yinping Jiao[1], Bo Wang[1], Sunita Kumari[1], Young K Lee[3], Demitri Muna[1], Daniel Bolser[3], Peter D'Eustacchio[4], Irene Papatheodorou[3], Paul Kersey[3], Pankaj Jasiwal[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]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, [4]New York University, School of Medicine, New York, NY, [5]USDA ARS, NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

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. It builds upon Ensembl and Reactome software, and is committed to open accesses and reproducible science based on the FAIR principles, providing both human and machine access to the data. Gramene provides integrated search capabilities and interactive views to visualize gene features, gene neighborhoods, phylogenetic trees, gene expression profiles, pathways, and cross-references.

Gramene's genomes portal hosts browsers for 53 complete plant 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 57 will include over 67,000 gene family trees comprised of more than 1.7 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 260 pathways curated in rice and inferred in 74 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.

Gramene is supported by NSF grant IOS-1127112, and USDA-ARS (1907-21000-030-00D).

# APPLICATION OF MACHINE LEARNING TO PRIMARY MELANOMA TRANSCRIPTOMES TO PREDICT PROGNOSIS IN A POPULATION-BASED COHORT

Rohit Thakur[1], Jérémie Nsengimana[1], Bram Vandekerckhove[2], Martin Lauss[3], Göran Jönsson[3], Tim Bishop[1], Julia Newton-Bishop[1], Jennifer H Barrett[1]

[1]Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, United Kingdom, [2]Eagle genomics, Wellcome Genome campus, Cambridge, United Kingdom, [3]Division of Oncology and Pathology, Lund University, Lund, Sweden

State-of-the-art machine learning methods have shown potential for identification of diagnostic and prognostic biomarkers in cancer. In this study, machine learning algorithms Random forest (RF) and Support Vector Machine (SVM) were applied to predict prognosis in one of the largest primary melanoma cohorts (n=687). Whole-genome transcriptome data were profiled for primaries of the Leeds Melanoma Cohort using the Illumina DASL-HT12-v4 array. Patients were stratified into two classes, class 0 (survived for at least 6 years) and class 1 (died of melanoma before 6 years). Patients followed up for less than 6 years were removed from the analyses. Transcriptomic data (n=525) was divided into training (70%) and test sets (30%). RF and SVM (with linear and non-linear functions) were applied to the training set to develop classification models. The training set was imbalanced (69% class 0 and 31% class 1); RF and SVM models were generated by undersampling the majority class. The model performance was assessed in the test set using sensitivity (sens), specificity (spec) and kappa (k). Clinical features such as stage, tumour site, age and sex were added to improve the prediction performance of the models.

The RF model generated after undersampling the majority class outperformed (sens= 0.54, spec= 0.86, k= 0.41) the RF model without undersampling (sens= 0.23, spec= 0.96, k= 0.23). The SVM model with undersampling (sens= 0.63, spec= 0.70, k= 0.31) outperformed the SVM model without undersampling (sens= 0.34, spec= 0.90, k= 0.27). In SVM, the linear function showed a better performance than several non-linear functions (SVM poly: k=0.10, SVM radial: k=0). With undersampling, RF had a better performance (k= 0.41) than SVM (k= 0.31), but the two methods had a good agreement overall (Cramer's V-0.73). Biological network analysis of the top 500 class predictors from RF identified central nodes as the genes involved in cell-cycling (CDK4, CKS1B and RPA3, as well as PSMD4, a subunit of the proteasome 26S involved in regulation of proliferation and apoptosis). The transcriptome based RF model had higher specificity in comparison to clinical feature based RF model (sens= 0.71, spec= 0.71, k=0.39). Combining transcriptomic and clinical features resulted in a RF model with improved sensitivity (in comparison to transcriptome-based data RF model) and specificity (in comparison to clinical feature based RF model) (sens= 0.63, spec=0.78, k=0.40).

Overall, the RF model performance was marginally better than the SVM model, but the two models showed broadly similar results. Integrating clinical features improved sensitivity of RF. Adjusting for clinical covariates in SVM and validation of findings in independent datasets is required. This data will be available before the conference.

# AN ENDEAVOR TO ASSEMBLE HUMAN CENTROMERES FROM LONG READS

Shubhakar R Tipireddy, Yuta Suzuki, Shinichi Morishita

The University of Tokyo, Department of Computational Biology and Medical Sciences, GSFS, Kashiwa, Japan

The human centromeres are comprised of ~171 bp tandem repeats referred to as alpha satellites. In the majority of centromeric regions, a multimer of alpha satellites monomers organizes a unit which itself iterates tandemly several hundreds of times resulting in a higher-order repeat (HOR) structure. Assembling this region from shotgun sequencing reads remains a challenge to date due to the high sequence identity (95~100%) among repeat units in the HOR structure.

Given the slight variation (0~5%) among the alpha satellites in each HOR unit, we propose a method to utilize these differences (single nucleotide variants a.k.a SNVs) as markers to prevent the centromere assembly from collapsing. Identifying these sparsely distributed SNVs from only PacBio long-reads is infeasible due to their high error rate. Hence we utilize both highly accurate short reads from the Illumina sequencing and erroneous long reads from the PacBio sequencing of a personal genome and attempt to assemble human centromeric regions.

Starting with an initial set of 1197 distinct monomers belonging to the human centromeric satellite family (Sevim et al., 2016), we generated 590 clusters of monomers each with a representative monomer sequence and a minimum intra-cluster sequence similarity of 90%. Each representative monomer sequence was aligned to the short-reads data to detect a set of possible variants present in the personal genome. Each monomer variant identified is characterized by a set of the SNVs present in it along with a unique identifier (ID) assigned to it. Occurrences of representative monomers in PacBio reads were detected through sequence alignment and a set of mismatches for each occurrence was obtained. For each representative monomer occurrence, the closest variant of it, if any, was assigned after careful comparison of the corresponding mismatch set with all the variant SNV sets. Replacing raw PacBio sequence with the assigned monomer variant sequence (i.e. indirectly the monomer sequence obtained from short-reads) effectively reduced the sequencing error in PacBio reads while carefully preserving the SNVs required to prevent the sequence assembly from collapsing. An overlap-layout graph was constructed after computing the alignments between encoded PacBio reads. Each edge in the graph represents an alignment between two reads supported by at least one marker (monomer variant). Upon graph traversal, we have successfully obtained read overlap-layouts, where some paths were extending up to ~500k base pairs. We are going to report latest analyses of obtained centromere assembly.

# JOINT MODELING OF GENETIC AND EPIGENETIC EFFECTS FOR COMPLEX PHENOTYPES.

Daniel Trejo Banos[1], Generation Scotland[2], Kathy L Evans[3,4], Andrew M McIntosh[5,4], Ian J Deary[4,6], Riccardo E Marioni[3,4], Matthew R Robinson[1]

[1]University of Lausanne, Department of computational biology, Lausanne, Switzerland, [2]University of Edinburgh, Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom, [3]University of Edinburgh, Centre for Genomic and Experimental Medicine, Edinburgh, United Kingdom, [4]University of Edinburgh, Centre for Cognitive Ageing and Cognitive Epidemiology, Edinburgh, United Kingdom, [5]University of Edinburgh, Division of Psychiatry, Edinburgh, United Kingdom, [6]University of Edinburgh, Department of psychology, Edinburgh, United Kingdom

Epigenetic modifications are hypothesized to play a role in many common diseases and quantitative traits. However, the magnitude of their effects and their joint contribution toward common disease risk remains to be elucidated. Methylome-wide association studies (MWAS) use methodology from genome-wide association studies (GWAS), even though the biochemical data is fundamentally different, and treat epigenetic effects as independent of genotype effects. Here, we present an approach that jointly estimates genetic and epigenetic effects on a given trait, with each having independent prior distributions over the variance explained by their effects. Our framework is flexible enough to model and account for some of the known biases in these types of data (e.g. cell counts), estimates effects (genetic and epigenetic) conditionally thus avoiding confounding, and by working in a Bayesian formalism we overcome the burden of multiple testing. We demonstrate our approach using a data set of 5000 individuals with measurements over SNP markers, methylation profiles, and a variety of traits (BMI, alcohol consumption and smoking habits). We estimate the independent contribution of methylation profile to the phenotypic variance, identify biomarkers that we replicate in an independent dataset, and determine the improvements in phenotypic prediction that can be obtained from a methylation array. Our framework is widely applicable, enabling improved understanding of the biological basis of complex disease and a better understanding of the value to epigenetic arrays to clinical diagnosis.

# HIGH-THROUGHPUT SINGLE-CELL DNA SEQUENCING USING DROPLET MICROFLUIDICS

Sebastian Treusch*, Maurizio Pellegrino*, Adam Sciambi*, Jennifer A Geis, Manimozhi Manivannan, Robert Durruthy-Durruthy, Kaustubh Gokhale, Jose Jacob, Tina X Chen, Pedro Mendez, Daniel Mendoza, William Oldham, Dennis J Eastburn, Keith W Jones

Mission Bio, Inc., South San Francisco, CA

* authors contributed equally

To enable the characterization of genetic heterogeneity in tumor cell populations, we developed a novel microfluidic approach that barcodes amplified genomic DNA from thousands of individual cancer cells confined to droplets. The barcodes are then used to reassemble the genetic profiles of cells from next generation sequencing data.

In a first application of this approach, we detected SNPs and indels across 19 disease relevant genes within more than 10.000 cells from longitudinally collected acute myeloid leukemia (AML) tumor populations. Our single-cell targeted sequencing method was able to sensitively identify tumor cells during complete remission and uncovered complex clonal evolution within AML tumors that was not observable with bulk sequencing.

We have now developed a targeted sequencing panel to assess more than 300 loci associated with a broad set of myeloid disorders, as well as the capability to build custom amplicon panels for specific applications of single cell DNA analysis, such as studies of solid tumor heterogeneity or CRISPR gene editing.

Here we will focus on technical aspects of our high-throughput single cell sequencing platform. We will demonstrate that our method results in low allele dropout, maintains a low number of cell doublets, and has high sensitivity to call rare subclone populations. Our single cell DNA sequencing approach enables the routine analysis of cellular heterogeneity and we anticipate that it will lead to improved stratification and therapy selection for AML and other cancers.

# FUNCTIONAL FINE-MAPPING OF ASTHMA AND IBD ASSOCIATION AT 11q13.5 SUGGESTS A TREG-SPECIFIC ENHANCER THAT DRIVES STAT5-RESPONSIVE EXPRESSION OF GARP

Rabab Nasrallah[1], Lara Bossini Castillo[2], Dafni A Glinos[2], Rahul Roychoudhuri[1], Gosia Trynka[2]

[1]Babraham Institute, Cambridge, United Kingdom, [2]Wellcome Sanger Institute, Cambridge, United Kingdom

The majority of variants associated to complex diseases map to protein non-coding regions of the genome. Previously we have shown that SNPs predisposing to common immune-mediated disease such as inflammatory bowel disease (IBD), rheumatoid arthritis and multiple sclerosis are enriched within active enhancers specific to regulatory T cells (Tregs). However, detailed functional consequences of these enhancer variants are poorly defined. Here, using a syntenic alignment approach and CRISPR/Cas9-based enhancer mutagenesis, we have identified a region of mouse chromosome 7 homologous to a human autoimmune/allergic risk locus 11q13.5 containing an enhancer element (hereinafter Lrrc32+70k) specifically commissioned in Foxp3+ Treg cells and required for Treg-specific expression of Lrrc32, encoding Glycoprotein A Repetitions Predominant (GARP). Highly conserved binding sites within Lrrc32+70k recruit the transcription factor STAT5 to mediate interleukin (IL)-2 driven expression of GARP on Treg cells. Treg-specific loss of GARP expression through Lrrc32+70k mutagenesis results in cytokine dysregulation that is rescued by provision of WT Treg cells. Finally, using gene expression from Tregs isolated from 100 healthy individuals we identify a Treg specific LRRC32 expression quantitative trait locus (eQTL). The eQTL signal and the GWAS associations for IBD and asthma are driven by the same variants. Finally, we prioritise the colocalised signal down to a single functional variant residing in a Treg enhancer marked by H3K27ac. These findings define a function for the 11q13.5 risk locus in Treg-mediated immunoregulation and directly implicate Treg cells in the pathophysiology of human autoimmune and allergic diseases.

# ASSESSING THE IMPACT OF GENE REGULATORY VARIATION ON MUTATIONAL PROCESSES ACTIVE IN CANCER

Lara Urban, PCAWG 3, PCAWG 8, Jan O Korbel, Oliver Stegle, Sebastian M Waszak

European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany

Cancer is a genetic disease and is caused by the accumulation of somatic mutations across the genome. Combinations of specific mutation types are termed mutational signatures and result from the activity of endogenous and exogenous mutational processes. Over 30 mutational signatures have been mapped across the most common cancer types, however, the molecular basis of most signatures remains unknown. Here, we tested within the Pan-cancer Analysis of Whole Genomes (PCAWG) for associations between mutational signatures and gene expression levels based on 1,178 patients that together cover 27 cancer types. Overall, we identified associations between 1,176 genes and 28 signatures, which were enriched for biological processes that were consistent with known etiologies of well-studied mutational processes, including APOBEC mutagenesis and tobacco smoking. We also explored the utility of gene expression-mutational signature associations to derive de novo functional annotations of signatures with previously unknown etiologies (*e.g.*, oxidative damage and Signature 38). Moreover, we studied the impact of *cis*-regulatory germline variation on gene expression and mutational signatures. We integrated *cis*-eQTL data from PCAWG and GTEx and identified 868 associations between germline variants, gene expression changes, and variation in mutational signatures. This analysis allowed us to disentangle the cause-consequence relationship of gene expression-mutational signature correlations. We were able to recapitulate known genetic associations, linking APOBEC mutagenesis to a common germline locus and showing that the effect of the variant is mediated via modulation of *APOBEC3A* and *APOBEC3B* expression. Our work further uncovered novel germline variants driving mutagenic processes, like a locus that modulates expression of the sterol transporter *ABCG8* and the liver-enriched Signature 12. We also identified that variation in *RAD51* expression as well as *cis*-eQTL for *RAD51* are associated with Signature 3, a mutational process typically observed in *BRCA*-deficient tumours. Hereby, *cis*-eQTL that down-regulate the expression of *RAD51* are also associated with a higher probability of Signature 3 mutations, suggesting that *RAD51*-deficiency might contribute to failure in double-strand break DNA repair by homologous recombination. Together, our study highlights the contribution of gene regulatory variation to modulating somatic mutational processes in cancer tissues.

# GENETICS EFFECTS ON ENHANCER ACTIVITY IN HUMAN PANCREATIC ISLETS

Arushi Varshney[1], Michael R Erdos[2], Narisu Narisu[2], John Hensley[3], Francis S Collins[2], Stephen C Parker[1,3]

[1]University of Michigan, Human Genetics, Ann Arbor, MI, [2]NIH, National Human Genome Research Institute, Bethesda, MD, [3]University of Michigan, Dept. Computational Medicine and Bioinformatics, Ann Arbor, MI

Genome-wide association studies (GWAS) have identified >100 single nucleotide polymorphisms (SNPs) that modulate risk for type 2 diabetes (T2D) and related traits. Using functional genomic profiling, we and others have shown that T2D GWAS SNPs are significantly enriched to overlap pancreatic islet stretch enhancer chromatin states, implicating these regulatory elements as biological mediators of disease risk. Recent studies that integrate genetic variation and islet gene expression have identified expression quantitative trait loci (eQTL) that nominate target effector transcripts for T2D GWAS SNPs. However, the subset of causal regulatory SNPs that mediate genetic control of gene expression are unknown. To help close this gap in knowledge, we have performed capped analysis of gene expression (CAGE) on RNA enriched for active enhancers (eRNA-CAGE) across 50 human islet samples. Comparison of these profiles to previously published islet chromatin maps revealed significant enrichment for active chromatin states, including enhancers ($P<0.001$). Interestingly, we observe unbalanced bidirectional transcription at chromatin-defined intronic enhancer regions, where the predominant direction of enhancer transcription is concordant with the transcribed direction of the endogenous gene, suggesting interactions between gene and enhancer activity. One such example is an ABCC8 intronic enhancer that we previously functionally validated using a mouse transgenic reporter assay. We further performed genotyping and imputation, which allowed us to run an eRNA-QTL scan. We identified 547 associations ($P<0.05$ for beta distribution adjusted P-value; QTLtools package), which are significantly enriched ($P=2.8x10^{-10}$) to occur in islet active enhancer chromatin states. Notably, we found islet eRNA-QTL that overlap T2D GWAS SNPs and islet eQTL. Together, these results form a genetic map of human islet enhancer activity that is a foundation for interpreting regulatory mechanisms at T2D GWAS loci.

# ITERATIVE REANALYSIS USING NOVEL GENOMIC RESEARCH TOOLS IMPROVES CLINICAL EXOME DIAGNOSTIC YIELD IN COMPLEX UNDIAGNOSED DISEASE CASES

Matt Velinder[1], John C Carey[2], Lorenzo D Botto[2], Ryan Layer[1], Brent Pedersen[1], Andrew Farrell[1], Ashley Andrews[2], Pinar Bayrak-Toydemir[3], Rong Mao[3], Aaron Quinlan[1], Gabor T Marth[1]

[1]USTAR Center for Genetic Discovery, Eccles Institute of Human Genetics, Salt Lake City, UT, [2]Division of Medical Genetics, Department of Pediatrics, Salt Lake City, UT, [3]Molecular Genetics and Genomics, ARUP Laboratories, Salt Lake City, UT

Precision medicine critically relies on the ability to rapidly identify clinically actionable genetic variants in affected patients. Technological and computational advances as well as cost reductions and data processing pipeline improvements have made rapid whole genome and whole exome sequencing increasingly feasible in clinical settings. However, most diagnoses made through clinical sequencing relate to well defined monogenic disorders for which causative genes have been established and known pathogenic variants have been identified. And despite the armamentarium of modern medicine and genomics many patients remain undiagnosed and unable to benefit from targeted molecular genetic therapies that could alter their clinical management and potentially improve their quality of life.

To address this disparity we have contributed our lab's bioinformatics and computational expertise to the Penelope Undiagnosed and Rare Disease Program. To date the Penelope Program has enrolled over 40 families with a wide range of complex undiagnosed clinical phenotypes. The project's diagnostic yield currently exceeds 50%. However, several cases remained undiagnosed despite exhaustive analysis by a team of trained bioinformaticians, medical geneticists and clinicians. We subsequently enrolled these cases into a research protocol where raw sequencing data was released to our lab and proceeded through an in-house alignment and variant calling pipeline (Sentieon Broad GATK best practices). We then applied genomic analysis tools developed at UCGD including gene.iobio (gene.iobio.io), RUFUS (github.com/jandrewrfarrell/RUFUS), GEMINI (github.com/brentp/gemini), Graphite (github.com/dillonl/graphite) and Peddy (github.com/brentp/peddy), among others. Using this reanalysis workflow we were able to diagnose an additional number of cases.

One of the cases that benefited from this reanalysis workflow was a 5-year-old boy with subependymal gray matter heterotopia, associated with pre- and post-natal growth delay, multiple congenital anomalies, bifid uvula and an unusual episode of altered behavior and consciousness with transient right hemiparesis. Upon applying this reanalysis workflow we readily identified a *de novo* frameshift variant in the *SON* gene. Based on the predicted functional impact, mode of inheritance and matching literature-reported phenotypes we concluded that this *SON* variant was causative of the disorder. This particular case and the others we have solved by this approach demonstrate its utility. More broadly, we propose releasing clinical exomes into an academic research setting where novel genomic algorithms and tools can be applied in an iterative reanalysis workflow could greatly improve diagnostic yield.

# T2D DISEASE STATUS ALTERS THE EFFECTS OF GENETIC VARIATION ON MOLECULAR PHENOTYPES: A DIRECT STUDY

A Viñuela[1], J Fernandez[2], A Kurbasic[3], MG Hong[4], S Sharma[5], C Brorsson[6], J Adamski[5], JM Schwenk[4], ER Pearson[7], S Brunak[6], PW Franks[3], MI McCarthy[2], ET Dermitzakis[1], IMI DIRECT consortium[7]

[1]U. of Geneva, Genetic Medicine & Development, Geneva, Switzerland, [2]U. of Oxford, WTCHG, Oxford, United Kingdom, [3]U. of Lund, Clinical Sciences, Malmö, Sweden, [4]Karolinska Institutet, SciLifeLab, Solba, Sweden, [5]Helmholtz Zentrum, Genome Analysis, Munich, Germany, [6]DTU, System Biology, Lyngby, Denmark, [7]U. of Dundee, Mol. & Clinical Medicine, Dundee, United Kingdom

While GWAS studies have produced many associations between genetic variants and complex diseases, we still lack an understanding of the cellular changes that occur during disease development. With this in mind, the DIRECT consortium built a cohort of 3,029 prediabetic and newly diagnosed T2D subjects, genotyped and phenotyped, with RNAseq, targeted metabolites (Biocrates) and proteins (immunoassays) measurements. Here, we present work on how genetic variants perturb molecular networks, and how these can also be perturbed by disease status.

Firstly, we looked for evidence of eQTLs in blood identifying 59,704 independent eQTLs associated with 94.2% genes (FDR 1%), with 78.7% genes having more than one eQTL. Compared to the eQTLs in 44 tissues produced by GTEx, we observed that 81% (testis) to 98% (brain) of the eQTLs were also active in blood, demonstrating the value of large studies in accessible tissues to uncover genetic effects in other tissues. Co-localization analysis between eQTLs and 153 GWAS variants for T2D identified 47 candidate genes mediating the signal in blood. Of those, the CaVEMaN fine-mapping method proposed 13 specific causal variants (Probability > 0.9). For example, the T2D associated GWAS variant at the AP3S2 locus (rs2028299) co-localized with the eSNPs for the same gene (RTC score 0.91) and the variant showed a high probability of being causal (P = 0.98). Adding information on 116 targeted metabolites and 263 proteins, we found that 67.2% of the genes (10,905) were associated with at least one metabolite, while 71.4% of the genes (11,577) were associated with at least one protein.

Finally, to investigate how the molecular network responds to T2D development, we looked for evidence of T2D specific genetic regulation. We identified 182 significant genotype-by-T2D interactions (FDR 1%), including rs28456 for FADS2 (a known T2D risk gene). FADS2 was also associated with 47 metabolites. We are currently further exploring the relationship between variants, gene expression, metabolites, and disease status to understand how they are mediated. Our results will increase the understanding on the complex ways genetic regulation affects multiple cellular phenotypes and leads to disease development.

# TRANS-eQTL META-ANALYSIS IN MORE THAN 31,000 BLOOD SAMPLES ASSOCIATES DOWNSTREAM GENES AND PATHWAYS WITH POLYGENIC RISK

Urmo Võsa*[1], Annique Claringbould[1], Harm-Jan Westra[1], Tõnu Esko[2], Lude Franke[1]

[1]University Medical Center Groningen, Department of Genetics, Groningen, Netherlands, [2]University of Tartu, Instititute of Genomics, Tartu, Estonia

The effects of trait-associated variants on the expression of nearby genes (expression quantitative trait loci; *cis*-eQTLs) have been well established, while the downstream consequences of such variants (*trans*-eQTLs) are mostly unclear. However, the downstream effects are especially relevant for prioritizing novel therapeutic targets in the future. Importantly, the identification of *trans*-eQTLs is non-trivial due to small effect sizes and a multiple testing burden.

To this end, we performed the largest meta-analysis to date, combining 31,684 blood samples from the eQTLGen Consortium. We used a standardised framework to test ~7 million variants for *cis*- and 10,563 genetic risk factors for *trans*-eQTL. Additionally, we calculated the polygenic risk scores for 1,263 traits and correlated those with gene expression (expression-polygenic risk scores; ePRS).

Using this approach, we identified *cis*-eQTL effects for 88% out of 19,250 blood-expressed genes. We observed *trans*-eQTL effects for 4,555 (43%) of tested genetic risk factors, impacting the expression of 7,772 (40%) genes. Out of significant genetic risk factors, 37% were linked with a *trans*-eQTL, suggesting the colocalization of signals.

We identified 335 "hub" SNPs, each influencing the expression of more than 50 distal genes. The majority of the *trans*-eQTLs (84%) also showed a *cis*-eQTL, enabling the investigation of local mechanisms leading to downstream effects. One example is rs17087335, a variant associated with coronary artery disease, affecting the gene expression of *REST* in *cis* and 98 genes in *trans*. *REST* is a transcriptional repressor, suppressing neuronal genes in non-neuronal tissues. The *trans*-eQTLs for rs17087335 were enriched for *REST* targets and for genes expressed in brain (FDR=1.2×10-45 and 1.5×10-17), which supports the proposed causal mechanism via *REST*. Conversely, we also identified clusters of genes that are affected by risk factors associated with the same disease. For example, the risk variants for inflammatory bowel disease affect gene clusters associated with B cell activation and interferon I signaling.

Finally, we identified 15,328 ePRSs, affecting 1,973 genes. For instance, the risk score for HDL cholesterol level was associated with the expression of genes involved in lipid metabolism (e.g. *ABCA1*, *ABCG1*) and familial hypercholesterolemia (*LDLR*). In brief, our study identifies a large number of downstream effects for trait-associated variants and helps to understand the potentially targetable molecular mechanisms involved in the disease.

# DISRUPTIONS TO RNA PROCESSING AND GENE REGULATION AS CONVERGENT MECHANISMS ASSOCIATED WITH MUTATION TO *CHD8*

A. Ayanna Wade, Kenneth Lim, Rinaldo Catta-Preta, Alex S Nord

University of California, Davis, Neurobiology, Physiology, & Behavior, Davis, CA

The packaging of DNA into chromatin determines the transcriptional potential of cells and is central to the differentiation process. Chromatin remodeling factors (CRFs) are key proteins that serve to control local chromatin state. Changes to dosage of CRFs can result in early embryonic lethality, severe phenotypic malformations, or cancer, indicating the importance of CRFs in development and cellular function. Recent sequencing of patient mutations has linked *de novo* loss-of-function mutations to CRFs with specific, causal roles in neurodevelopmental disorders, especially autism spectrum disorder (ASD). Characterizing cellular and molecular phenotypes arising from mutations to CRFs could reveal convergent mechanisms of pathology in patients. Chromodomain helicase DNA binding protein 8 (*CHD8*) is a CRF gene with one of the highest observed *de novo* loss-of-function mutations rates in patients with ASD. Mutations to *CHD8* are suggested to drive neurodevelopmental pathology through global disruptions to gene expression and chromatin state, which could give insight into biological underpinnings of ASD. However, mechanisms associated with Chd8 function have yet to be fully elucidated. Here, we analyzed published transcriptomic and epigenomic data across *Chd8 in vitro* and *in vivo* knockdown and knockout models and identify convergent mechanisms of gene regulation by Chd8. We find changes at the RNA level vary across studies and systems far more than genomic binding targets, suggesting that dysregulated gene expression is caused by both direct regulation and model-specific effects downstream of direct regulation. Our findings indicate conserved roles for Chd8 in RNA processing serving as a critical mechanism for gene regulation. Our findings are consistent with recent studies highlighting splicing in ASD and indicate RNA metabolism as an understudied, yet potentially widespread, causal factor in neurodevelopmental disorders.

# DISSECTING TRANSCRIPTIONAL REGULATION BY MACHINE LEARNING

Hai Wang[1,2], Maria K Mejía-Guerra[1], Karl A Kremling[1], Guillaume Ramstein[1], Ravi Valluru[1], Edward S Buckler[1], Jacob D Washburn[1]

[1]Cornell University, Department of Plant Breeding and Genetics, Ithaca, NY, [2]Chinese Academy of Agricultural Sciences, Biotechnology Research Institute, Beijing, China

Genetic diversity at the DNA level often leads to changes in transcriptional regulation, which in turn impacts transcript abundance, and ultimately, the phenotype. Recent advances have increased our ability to sequence genomes at low cost, but the potential of DNA-to-RNA prediction for identifying key sequence features remains largely unexplored. Here, we report the development of several machine learning (ML) models that take as input genomic regions flanking the coding sequences. First, we adapted natural language processing algorithms to model chromatin domains and transcription factor binding sites in the genome with auROC values above 0.90. We then developed convolutional neural network (CNN) models to discriminate between genes and pseudogenes with an auROC above 0.87. We further designed CNN models which predict absolute gene expression values with an $R^2$ of 0.51. Examination of models that predict pseudogenes and absolute expression values show that they place considerable weight on both the 5' and 3' UTR regions in their predictions. In crop breeding programs, breeders are often interested in relative expression among alleles in a population. To achieve allelic resolution, we predicted which allelic version of a gene is more strongly expressed by contrasting pairs of homologs from the two maize sub genomes, and between maize genes and their syntenic orthologs in sorghum. The model performs with auROC values above 0.90. These models can be applied to better understand the basis for gene regulation, by identification of cis-regulatory elements, and to increase genetic gains in plant breeding, by improvement of genomic prediction models.

# HYBRID *DE NOVO* ASSEMBLY OF *BRANCHIOSTOMA BELCHERI BEIHAI* AMPHIOXUS GENOME

Ming-Qiang Wang[1,2], Kevin Yi Yang[1,2], Junyuan Chen[3], Bingyu Mao[4], Stephen Kwok-Wing Tsui[1,2]

[1]The Chinese University of Hong Kong, School of Biomedical Sciences, Hong Kong SAR, China, [2]The Chinese University of Hong Kong, Hong Kong Bioinformatics Centre, Hong Kong SAR, China, [3]Chinese Academy of Sciences, Nanjing Institute of Paleontology and Geology, Nanjing, China, [4]Chinese Academy of Sciences, Kunming Institute of Zoology, Kunming, China

*Branchiostoma belcheri*, also known as Chinese amphioxus, is the closest living invertebrate relative of the vertebrates. It is widely used as a model system for studying evolutionary developmental biology and the origin of vertebrates. Beihai amphioxus is a subspecies of ***B. belcheri*** that inhabits in Guangxi Beihai, China. The previously reported amphioxus genome generated by short reads sequencing technologies are highly fragmentary, which hinders the downstream analysis and further applications. Here we report the sequencing and assembly of the 640Mb Beihai amphioxus genome using a hybrid approach that combined Pacific Bioscience Single Molecule Real-Time (SMRT) long reads and Illumina short reads sequencing technology. Specifically, a total of 80 Gb genomic data, including 20X PacBio long reads from Sequel sequencing platform, 66X paired-end reads and 56X mate-pair reads from Illumina HiSeq 2000 sequencing platforms, were generated to achieve a high-quality Beihai amphioxus genome. We performed hybrid *de novo* assembly, scaffolding, gap filling and polishing bioinformatics pipeline to obtain the assembly result which contains 92,511 contigs and 28,014 scaffolds with the contig N50 of 12 kb and scaffold N50 of 73 kb, respectively. The assembly genome contains 907 (92.7%, out of 978) BUSCO core genes collected from metazoa_odb9 database, with 796 (81.4%) of them being complete. The assembly result contains 15M repetitive sequences, which contains 67% simply repeats, 11% SINEs and 7% low-complexity sequence, accounting for 2.28% of the whole genome. Evidence-driven gene prediction method based on RNA-Seq data has identified 63,612 transcripts and 44,087 protein coding sequences in this Beihai amphioxus genome. We anticipate the Beihai amphioxus genome would improve our knowledge on the genetic diversity of this species, meanwhile providing a valuable genetic resource for the scientific community to further understand the vertebrate evolution.

# INCLUDING MEDICAL PROFESSIONALS IN THE HANDS-ON PRACTICE OF GENOMIC DATA ANALYSIS -- CLINICIAN-DRIVEN ANALYSIS OF GENOMIC PATIENT DATA

Alistair Ward[1,2], Matt Velinder[1], Tonya Di Sera[1], Gabor Marth[1,2]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]Frameshift Genomics, Boston, MA

Analyzing genomic data, specifically identifying genetic variants that may be causative for a patient's phenotype or disorder, is typically the job of a bioinformatician or diagnostic analyst. These individuals are experts in understanding and operating the complex tools required to analyze genomic data. It is critical that the analyst is fully aware of the full range of presented phenotypes, course of the disease, as well as comprehensive family history, where known. However, it is often the case that this information is not available to the analyst, severely limiting the efficiency and diagnosis rate for these analyses. Our approach to address this challenge is to more fully engage medical practitioners with such comprehensive understanding of the patient's disease, into the analysis process. The tools we are building allow them to analyze their own patients' genomic data, without help from a bioinformatician.

We have developed the IOBIO platform to support real-time visually driven analysis of complex genomic data. This project uses intelligent analysis of genomic data, coupled with an intuitive presentation of data that aims to open genomic analysis to a broader community, and address some of the problems highlighted above. The unique constraints and hurdles faced by individuals in the clinic need to be well understood in order to ensure that analysis tools are accessible to them. This includes understanding how these individuals perceive their role in the analysis process, and where they see their expertize being best utilized.

We have undertaken a study at the University of Utah to address these questions. We have engaged a community including cardiologists, neurologists, immunologists, at different points in the careers: from those still in medical school, to those operating their own clinics. This group has been independently analyzing the exomes of undiagnosed pediatric disorders using our IOBIO tools, leading to fundamental changes in the design of our analysis process. By comparing the diagnostic conclusions of the group for the same cases, along with the time taken to reach these conclusions, we have generated a goldmine of information; including the weighting given to different information in drawing a conclusion; the level of genetic education and knowledge held by many clinicians, which directly impacts our educational resources etc.

# SPECIES-SPECIFIC TRANSCRIPTIONAL CHANGES IN RESPONSE TO OXIDATIVE STRESS IN iPSC-DERIVED CARDIOMYOCYTES FROM HUMANS AND CHIMPANZEES

Michelle C Ward[1,2], Kristen M Patterson[1], Yoav Gilad[1,2]

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

The basis for complex polygenic diseases such as cardiovascular disease is unclear; however despite anatomical similarities, there are known differences in susceptibility between primates. For example, humans are prone to ischaemia-induced myocardial infarction, unlike our closest living relative, the chimpanzee. The ability to differentiate cardiomyocytes from induced pluripotent stem cells (iPSCs), in both species, now allows for direct inter-species comparisons. In order to understand human-specific regulatory adaptations in the heart, and to gain insight into the evolution of disease susceptibility and resistance, we have developed a model of hypoxia and oxidative stress in human and chimpanzee cardiomyocytes. We differentiated eight human and seven chimpanzee iPSC lines, together with six technical replicates, into cardiac troponin-positive iPSC-derived cardiomyocytes under normoxic conditions, subjected these cells to six hours of hypoxia, followed by six or 24 hours of re-oxygenation, and collected genome-wide gene expression data, as well as measurements of cellular stress at each time-point. The expression of thousands of genes is altered following oxygen stress in both species; however over a hundred genes show a species-specific response at one of the time-points. For example, *RASD1* is up-regulated specifically in human following hypoxia, and is a coronary artery disease GWAS hit of unknown function. Our results provide the first direct molecular measurements following cellular stress in primate cardiomyocytes, and suggest some degree of species-specificity in the transcriptional response in these two closely related species.

# GRAMENE MAIZE PAN-GENOME BROWSER

Sharon Wei[1], Joshua C Stein[1], Andrew Olson[1], Yinping Jiao[1], Bo Wang[1], Michael Campbell[1], Marcela K Tello-Ruiz[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA ARS NEA, Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Maize is the most genetically diverse crop in the world, with differences in gene content estimated between 5-20% among lines. Capturing the pan-Zea gene space and structural variation requires additional reference genomes, and the infrastructure to store, analyze and make accessible. To support this effort, Gramene has developed a dedicated genus-level browser resource: maize-pangenome.gramene.org, built upon the Ensembl infrastructure and guided by FAIR practices. Our first pass of this resource includes B73, W22 and PH207 complete reference genomes, along with 4 monocots, 3 dicots, 3 lower plants and 2 non-plant outgroup species. These served as input to generate phylogenetic resources based on protein and whole-genome DNA alignments. Insights into ancestrally conserved regions and structural rearrangements are defined by pairwise whole-genome alignments and displayed in a number of informative ways, including a multi-species view that allows graphical stacking of browsers and interspecies navigation. The gene trees can be used to programmatically identify gene expansions and losses between different maize accessions, which may explain evolutionary adaptations, inaccuracies in the gene models, or errors in the underlying reference genome assemblies. We anticipate maize accessions like the NAM populations being added to this resource. To test the utility of these resource and to assess quality of the gene structure predictions, Gramene outreach efforts include the first maize annotation jamboree co-organized with the MaizeCODE project. This work constitutes an initial prototype to support the infrastructure to identify misannotated gene structures and a process to correct these guided by the gene trees. In addition to providing resources to support quality assessment, as well as insights into many outstanding questions in the evolutionary history of the Zea genus, this resource will provide a basis for functional characterization of genes and the identification of targets for agronomic improvement of maize. This project is funded by NSF (IOS-1127112) and partially from USDA-ARS (1907-21000-030-00D).

# RETROTRANSPOSON EXPRESSION IN HUMAN CELLS

Paul Schaughency[1], Srinivasan Yegnasubramanian[1], Sarah J Wheelan[1,2]

[1]Johns Hopkins University School of Medicine, Oncology, Baltimore, MD, [2]Johns Hopkins Bloomberg School of Public Health, Biostatistics, Baltimore, MD

Studies in cell lines and tissues indicate that expression of repetitive elements, particularly Alu and L1 retrotransposons, varies across cell types; however, in nearly all cases this expression is low, suppressed by multiple cellular mechanisms, including methylation and post-transcriptional modification. To better understand transposable element expression, we developed an efficient capture-based high throughput sequencing strategy for enriching and sequencing these transcripts from total RNA. We are able to map transcripts to their originating locus and determine patterns of expression in cancer cell lines, including a cell line with disabled DNMT1 and DNMT3B, and we use several methods to correlate transposable element expression with various features genomewide.

# CHADFINDER: A COMPUTATIONAL METHOD TO EVALUATE MULTIPLEX PCR AMPLICON QUALITY AND PRIMER BEHAVIOR FOR ITERATIVE OPTIMIZATION

Heather C Wick[1], Marija Debeljak[2], James R Eshleman[2], Sarah J Wheelan[1,3]

[1]Johns Hopkins University School of Medicine, Institute of Genetic Medicine, Baltimore, MD, [2]Johns Hopkins University School of Medicine, Department of Pathology, Baltimore, MD, [3]Johns Hopkins University School of Medicine, Department of Oncology, Baltimore, MD

Multiplex PCR allows researchers and clinicians to simultaneously amplify multiple regions of the genome in a single experiment, saving time, resources, and cost associated with performing multiple single PCR experiments. However, extra care must be taken in order to plan a multiplex PCR experiment to avoid problems such as misbinding events, and to optimize the relative abundance of each resulting amplicon.

In addition to primer dimers, another type of misbinding event common in in multiplex PCR has been described wherein a primer intended for one target sequence instead binds within another target sequence. If amplified, the resulting faulty amplicon, or CHADs (Chimeric Heteromorphic Amplification Defects), can be mistaken for a variant or mutation. These false positive variants and mutations can negatively impact research and patient diagnostics.

Previously, we described CHADfinder, a computational method for identifying CHADs and other misbinding events from multiplex PCR experiments in order to remove them to improve variant calling analysis. CHADfinder also characterizes the behavior of the primers involved in the experiment, enabling the user to identify and replace problematic primers responsible for misbinding events or poor amplification in order to iteratively optimize primer design for multiplex PCR experiments.

Here we present analysis of a series of multiplex PCR experiments in which we used CHADfinder to iteratively identify and redesign problem primers for each subsequent rerun. For each multiplex experiment, we characterized amplicon quality and primer behavior in order to evaluate the level of improvement over the previous primer set.

# MOLECULAR TETHERS FOR VERY LARGE DNA MOLECULES

Eamon Winden[1,3], Samuel Krerowicz[1,2,3], David C Schwartz[2,3]

[1]University of Wisconsin-Madison, Laboratory of Genetics, Madison, WI, [2]University of Wisconsin-Madison, Department of Chemistry, Madison, WI, [3]University of Wisconsin-Madison, Laborator for Molecular and Computational Genomics, Madison, WI

Very large DNA molecules as objects present a unique set of advantages and challenges to be discovered and exploited for advancing the genomic sciences. These attributes hinge on the intrinsic physical characteristics of very large, stiff, random coils in solution: huge hydrodynamic dimensions giving rise to fragility under shear forces. Such considerations complicate common lab manipulations, which have been dealt with through inventions that include: Pulsed Field Gel Electrophoresis, Yeast Artificial Chromosomes (YACs) and Optical Mapping—systems that leverage many intrinsic physical and genomic advantages. Given this context, we aim to synthesize DNA molecules on the order of mammalian chromosomes.

# NANOPORE SEQUENCING REVEALS THE TRANSCRIPTIONAL COMPLEXITY OF NEUROPSYCHIATRIC DISEASE GENES IN HUMAN BRAIN.

Tomasz Wrzesinski[1], Michael Clark[2], Paul Harrison[2], Daniel Weinberger[3], Elizabeth Tunbridge[2], Wilfried Haerty[1]

[1]The Earlham Institute, Organisms and Ecosystems, Norwich, United Kingdom, [2]University of Oxford, Department of Psychiatry, Oxford, United Kingdom, [3]Lieber Institute For Brain Development, Baltimore, MD

Mutations leading to aberrant splicing are increasingly associated with human diseases and disorders. So far, we have relied on short read sequencing to assess splicing diversity and transcriptional convolution. However, the accurate annotation and quantification of full length transcript using short read sequencing still remains a challenging task. Fortunately, recent method developments have now made possible to sequence full length cDNA on the Oxford Nanopore platform enabling the in-depth annotation and analysis of alternatively spliced transcripts. As proof of principle, we first focused on splicing events within complex genes (>40 exons) such as the voltage-gated calcium channels (VGCCs - *CACNA1A*, *CACNA1C*, *CACNA1D*, *CACNA1S*). Many of these genes have been associated with human diseases, for instance, *CACNA1C* has been robustly linked with bipolar disorder through genome, transcriptome and methylome studies. We generated full length sequences for VGCC transcripts, in 3 individuals and 6 different regions of human brain to quantify transcript diversity and identify novel functional splicing events. In *CACNA1C* and *CACNA1D*, we were able to determine 38 and 33 novel exons, respectively, as well as novel frame-conserving splice junctions and micro-exons. Our approach allowed to further annotate 90 and 44 novel high-confidence isoforms in *CACNA1C* and *CACNA1D*, respectively, which abundance clearly distinguish between different brain tissues. In order to create a map of transcriptome-wide splicing changes across brain tissues, we extended our approach to cDNA capture followed by Nanopore sequencing and characterized a plethora of previously unknown coding and noncoding brain expressed transcripts. Our data demonstrate that long read sequencing including Oxford Nanopore allows to uncover unprecedented transcriptional complexity, leading to more accurate annotations which in turn facilitates the understanding of the functional implications of mutations altering splicing.

# A NOVEL lncRNA MAHAC IS ESSENTIAL FOR HYPOXIA-INDUCED HISTONE MODIFICATION AND TUMOR PROGRESSION

Kai-Wen Hsu[1], Yu-Cheng Tsai[1], Pei-Hua Peng[1], Der-Yen Lee[2], Chuan He[3], Kou-Juey Wu[1]

[1]China Medical Univ., Research Center for Tumor Medical Science, Taichung, Taiwan, [2]China Medical Univ., Integrated Medicine, Taichung, Taiwan, [3]University of Chicago, Chemistry, Chicago, IL

Intratumoral hypoxia promotes tumor progression through inducing epithelial-mesenchymal transition (EMT), cancer stemness, angiogenesis, and glycolysis1,2. Long noncoding RNAs (lncRNAs) are critical for many physiological processes, including tumor metastasis. Hypoxia induces histone modifications and regulates EMT marker gene expression 3,4. Here we show that a hypoxia-activated lncRNA (MAHAC, maintenance of histone acetylation) implicated in tumorigenesis from bioinformatics plays a crucial role in hypoxia-induced metastasis. Overexpression of MAHAC activates epithelial-mesenchymal transition (EMT) and tumor metastasis. More importantly, knockdown of MAHAC abolishes hypoxia-induced EMT. Due to the role of lncRNAs in serving as a scaffold for chromatin modifying complexes, we show that knockdown of MAHAC decreases the global acetylation of histone 4 lysine 5 (H4K5Ac) levels. RNA pull down and mass spectrometry analysis show that MAHAC anchors the human interleukin enhancer binding factor (ILF3) and its partner ILF2; together with CBP they mediate H4K5Ac. qChIP and qChIRP assays show that knockdown of MAHAC decreases the H4K5Ac levels in the promoters of EMT regulators, indicating the role of MAHAC and ILF3-ILF2-CBP complex in regulating the expression of these EMT regulators. This report is the first demonstration that MAHAC plays a crucial role in hypoxia-induced EMT and metastasis by scaffolding a chromatin modifying complex to mediate H4K5 acetylation. The epigenetic regulation of MAHAC will be discussed.

References:
1. Yang MH, et al. Nature Cell Biology 10, 295 (2008).
2. Yang MH, et al. Nature Cell Biology 12, 982 (2010).
3. Wu MZ, et al. Mol. Cell 43, 811 (2011).
4. Wu CY, et al. Trends in Genetics 28, 454 (2012).

# DEGENERATIVE EXPANSION OF A YOUNG "SOCIAL CHROMOSOME" SUPERGENE

Eckart Stolle, Rodrigo Pracana, Yannick Wurm

Queen Mary University of London, Organismal Biology, London, United Kingdom

Suppressed recombination is thought to reduce the efficacy of selection due to the Hill-Robertson effect. The long-term consequences of this effect have been widely studied in sex chromosome systems: loss of chromosomal segments and accumulation of repetitive elements lead to a dramatic reduction in size, gene number and gene density. In contrast, few empirical studies have examined the effects of suppressed recombination over shorter timescales, or in other “supergene” chromosomal regions that are inherited as a single block and determine complex phenotypes.

The fire ant Solenopsis invicta carries a supergene region that controls a key social dimorphism: whether a colony accepts one or multiple queens. This supergene can be used to study the effect of suppressed recombination on chromosome evolution because the homozygous state of one of its variants (Sb) has low fitness, meaning that this variant rarely recombines. Here, we test whether Sb is affected by the accumulation of structural mutations.

Using long-molecule optical mapping of haploid individuals carrying each of the two variants of this supergene (SB and Sb), we show that the variant with restricted recombination (Sb) has undergone a ~15% increase in length during the last ~400 000 years, and that this has arisen via an accumulation of large additional sequences in the supergene. We found this pattern in two additional fire ant species expected to carry a supergene region similar to that of S. invicta.

Our results support the hypothesis that regions with recently restricted recombination should undergo degenerative expansion. Degenerative expansion has been previously shown to occur in the young sex chromosome systems of several plant species, and is thought to occur when selection is less effective for insertions, which are generally mildly deleterious, than for deletions, which are generally more strongly deleterious. Our study represents the first empirical evidence for degenerative expansion in an animal supergene.

# BRAIN GENE EXPRESSION RESPONSE TO PESTICIDE EXPOSURE INDICATES EFFECTS ON COGNITION.

Isabel K Fletcher[1], Thomas J Colgan[1], Andres Arce[2], Richard Gill[2], Yannick Wurm[1]

[1]Queen Mary University of London, Organismal Biology, London, United Kingdom, [2]Imperial College, London, United Kingdom

Insect pollinators including social bees are key to ecosystem stability as well as agricultural yields. Recent declines in social bees worldwide are thus concerning. A major factor implicated in these declines is the application of pesticides to crops. These are intended to control pest species, but can also negatively affect non-target wild pollinators. Behavioral and field studies have clearly demonstrated that exposure to neonicotinoid pesticides negatively impacts learning and memory abilities, foraging behavior and colony survival of social bees.

We know relatively little however about the molecular mechanisms by which pesticide exposure affects bee cognition. To address this, we exposed Bombus terrestris bumblebee colonies to commonly used pesticides. We find widespread effects of pesticide exposure on the gene expression in the brain, identifying candidate pathways involved in detoxification. We describe the signatures of selection on these genes in wild populations. Our work demonstrates a novel, straightforward manner of quantifying the effects of pesticide exposure.

# DISSECTING THE CAUSAL MECHANISM OF X-LINKED DYSTONIA-PARKINSONISM BY INTEGRATING GENOME AND TRANSCRIPTOME ASSEMBLY

Rachita Yadav*[1,2,3], Tatsiana Aneichyk*[1,2,3], William T Hendriks*[2,4], David Shin*[2,4], Dadi Gao*[1,2,3], Christine A Vaine[2,4], Ryan L Collins[1,3], Aloysius Domingo[1,2,3,4], Benjamin Currall[1,3], Nutan Sharma[2,4], Xandra O Breakefield[2,4], Laurie J Ozelius[2,4], D. Cristopher Bragg[2,4], Michael E Talkowski[1,2,3,4]

[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, [2]Department of Neurology, Massachusetts General Hospital, Boston, MA, [3]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, [4]The Collaborative Center for X-linked Dystonia-Parkinsonism, Massachusetts General Hospital, Boston, MA

X-linked Dystonia Parkinsonism (XDP) is an adult-onset neurodegenerative disorder that is indigenous to the Philippines and exhibits features of both dystonia and parkinsonism in a characteristic temporal progression. Over the last two decades, conventional genetic approaches have linked XDP to a 410 Kb founder haplotype that included the same seven markers shared by all probands, five single nucleotide variations known as Disease Specific Changes (DSCs), 2627 bp sine-VNTR-Alu (SVA) retrotransposon and 48bp deletion; yet, the causal variant and pathogenic mechanism have remained unknown.

We integrated *de novo* genome and transcriptome assembly methods using short-read, multiple long-read, linked-read and targeted sequencing technologies among XDP probands, carriers, and controls (n=792). We characterized transcriptomes from fibroblasts among 46 subjects (probands, carriers, and unaffected family members) as well as a subset of 24 clones from iPSC-derived neural stem cells (NSCs) and induced neurons.

These analyses identified a set of 54 alleles defining the most abundant haplotype as well as five independent recombination events that narrowed the putative causal genomic segment to 219 Kb including only the *TAF1* gene. Intriguingly, the de novo transcriptome assembly in NSCs revealed a striking expression signature involving aberrant splicing and intron retention (IR) of the *TAF1* gene, and partial retention of intronic sequence proximal to the SVA insertion within intron 32 of *TAF1* that has never been observed in controls. Canonical TAF1 transcripts were significantly reduced in XDP probands in iPSC-derived NSCs, and this reduction was driven by decreased exon usage 3' to exon 32. Both the aberrant splicing and reduced *TAF1* expression signatures were rescued following CRISPR/Cas9 excision of the SVA in patient-derived NSCs. Transcriptome-wide analyses also revealed expression alterations of neurodegenerative disorders genes as well as genes associated with synaptic transmission and neurodevelopment.

These data implicate aberrant splicing and intron retention as a consequence of a noncoding SVA insertion as a possible pathogenic mechanism in XDP, and propose a potential roadmap for integrated, reference-free genome and transcriptome assemblies in genomic studies of population isolates.

# GENETIC RECOMBINATION IS NOT LIMITED TO HOTSPOTS IN DOG MEIOSIS

Qi Yu[1], Fatima Smagulova[2,3], Kevin M Brick[1], Sarah Thibault[2], Daniel R Camerini-Otero#[1], Galina V Petukhova#[2]

[1]National Institutes of Health, National Institute of Diabetes, Digestive and Kidney Diseases, bethesda, MD, [2]Uniformed Services University of the Health Sciences, Department of Biochemistry and Molecular Biology, bethesda, MD, [3]IRSET, Rennes, France

Meiotic recombination is a fundamental process that generates genetic variation between homologous chromosomes. In most vertebrates, the DNA double strand breaks (DSBs) that initiate recombination occur at hotspots, 2-4 Kb regions determined by the DNA binding specificity of the PRDM9 protein. Mice that lack PRDM9 are infertile, but despite its apparent importance, the Prdm9 gene has been lost in several lineages, including canids. To investigate the mechanism of recombination initiation in mammals without functional PRDM9, we built high-resolution and genome-wide maps of meiotic DSBs in male dogs.

In contrast to other mammals, a large proportion of meiotic DSBs in dogs occur outside of narrowly defined hotspots. In addition to 17,162 classic DSB hotspots we detected 110 Recombination initiation Domains (RiDs) that display elevated DSB frequency. Importantly, crossover rate is also elevated in these domains. Both DSB hotspots and RiDs are conserved between three dogs of different breeds, consistent with the loss of PRDM9-mediated DSB targeting. DSB frequency at hotspots and RiDs is disproportionately high at subtelomeric regions, and like hotspots, RiDs are also CpG rich, suggesting that CpGs play a major role in recombination initiation in dogs. Indeed, we can simulate DSB formation and recapitulate the genome-wide patterns of most hotspots and RiDs by using CpGs as the only determinant of DSB targeting. Finally, the observed patterns of nucleotide substitutions is consistent with GC biased gene conversion, and support the hypothesis that recombination in male dogs drive the evolution of GC content in canid species.

# PARLIAMENT2: BENCHMARKING A STRUCTURAL VARIANT CONSENSUS CALLER COMPARED TO INDIVIDUAL METHODS

Andrew Carroll[1], Samantha Zarate[1], William J Salerno[2], Fritz Sedlazeck[2], Olga Krasheninina[2], Richard Gibbs[2]

[1]DNAnexus, Science, Mountain View, CA, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

Structural variants are large (>50bp) genomic rearrangements relative to a reference genome. Detecting these events is challenging because their sizes are similar to or larger than those of Illumina reads, forcing programs to use indirect methods (e.g. coverage, insert size distribution, read orientation, and split-mapping) to infer their presence. As a result, they are more poorly characterized than smaller variants; though a number of tools in the field focus on one or a few of these methods, the respective accuracies of individual methods remain low.

Here we present Parliament2, an updated version of the Parliament consensus SV-calling infrastructure. Parliament2 runs a number of individual short-read callers (Breakdancer, Breakseq, CNVnator, Delly, Manta, and Lumpy) on a whole-genome sample, overlaps the resulting calls to create a consensus using SURVIVOR, and genotypes each candidate to confirm its presence in the sample with high accuracy.

We benchmark each of the individual methods as well as Parliament2 on the v0.5 Genome in a Bottle structural variant truth set, identifying the accuracy and strengths of each individual method and demonstrating that the call-overlap-genotype approach of Parliament2 yields higher accuracy -- especially precision -- than the individual methods do alone.

Through optimized concurrent execution of the tools on a single machine, Parliament2 can finish running all tools in about the same time as the slowest individual method, requiring approximately 60 CPU-hours for a full genome.

# INTEGRATION AND ASSEMBLY OF EXTANT SEQUENCING TECHNOLOGIES TO ENABLE COMPLETE STRUCTURAL VARIATION DISCOVERY AND PHASING

Xuefang Zhao[4,5], Mark J Chaisson[1,2], Ashley D Sanders[3], Human Genome Structural Variation Consortium[6]

[1]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [2]University of Southern California, Molecular and Computational Biology, Los Angeles, CA, [3]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [4]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [5]Massachusetts General Hospital, Center for Genomic Medicine, Boston, MA, [6]The Jackson Laboratory for Genomic Medicine, Farmington, CT

Despite significant advances in technologies and algorithms to routinely process whole-genome sequences(WGS), full-scale discovery of structural variation (SV) remains a major challenge due to limitations in short-read WGS and associated SV detecting methods. In an effort to determine the scale of this problem, and to benchmark the capabilities of divergent data types to capture SV in the human genome, we have applied 7 genomics technologies and 23 independent algorithms (most run by the developer) to three parent-child trios from diverse ancestries to discover haplotype-resolved SVs. We designed integration methods that adjudicate SVs from short and long reads respectively, which were further integrated into a full-scale SV set that described more than 800,000 indels (1-49 bp) and 30,000 SVs (≥ 50bp) within each child, representing an ~7-fold increase over previous 1000 Genomes Project reports derived from standard short read libraries. We also discovered 156 large inversions and complex SVs per genome using this complementation of technologies, which were often omitted or misinterpreted by individual technologies or algorithms. We further show that near complete physical haplotype resolution with 10X Genomics and strand-seq data allows long-read sequences to be accurately partitioned and assembled by haplotype, significantly increasing our yield of SVs. Our results suggest that haploid resolution of diploid genomes is a realistic goal using combinations of orthogonal technologies, depending on variant size and class, though no single technology or algorithm was capable of achieving these benchmarks alone. We anticipate that such datasets will provide unique insights into the evolutionary and functional consequences of SV in human populations, and serve as a developing gold standard for discovering and understanding human genetic variation.

# IDENTIFICATION OF A BENEFICIAL, COMPLEX, GENETIC REARRANGEMENT IN *RCAN-1*, AN ORTHOLOG OF THE DOWN SYNDROME CRITICAL REGION PROTEIN 1 (DSCR1/RCAN1), IN A LABORATORY STRAIN OF *C. ELEGANS*

Yuehui Zhao, Patrick McGrath

Georgia Institute of Technology, School of Biological Sciences, Atlanta, GA

There is general interest in identifying beneficial mutations that can increase fitness of animals in new environments. In pursuit of these genetic variants, we performed QTL mapping on 89 recombinant inbred lines (RILs) between two laboratory strains of *C. elegans*, LSJ2 and N2, using relative fitness as a phenotype. Interestingly, we found one of the RILs displayed transgressive segregation, i.e. its relative fitness was higher than either of the parental strains used to construct this strain. Additionally, this RIL showed changes in foraging behavior not present in either parental strain. To determine if these changes in fitness and behavior were caused by a *de novo* mutation or higher order epistasis between specific combinations of N2 and LSJ2 alleles, we constructed new RILs from this outlier RIL and its two parental strains and performed bulk segregant analysis on the foraging behavior. This analysis identified a single locus on chromosome III, suggesting a *de novo* mutation was responsible. Using resequencing experiments, we identified a genomic rearrangement in *rcan-1*, an ortholog of the human DSCR1/RCAN1 gene. DSCR1/RCAN1 encodes a repressor of calcineurin A, and has been proposed to play a role in trisomy 21 caused Down Syndrome in humans. We found that this genetic change is complex, and consists of at least 6 genomic inversion/duplication events resulting in 2-8x increased copy number of the *rcan-1* locus. These genetic changes are predicted to 1) increase the copy number of the gene, 2) modify the promoter region of the gene, and 3) create a second truncated RCAN-1 protein. Using a near isogenic line surrounding the *rcan-1* rearrangement, we confirmed its role in the behavioral changes and fitness gains of this RIL. Together, our study demonstrates that adaptation can occur rapidly within the lab, and can result from genes implicated in human disease.

# MEGA-ANALYSIS OF ODDS RATIO (MEGAOR): A CONVERGENT METHOD FOR DEEP UNDERSTANDING THE GENETIC EVIDENCE IN SCHIZOPHRENIA

Peilin Jia[1], Zhongming Zhao[1,2,3]

[1]University of Texas Health Science Center at Houston, Center for Precision Health, School of Biomedical Informatics, Houston, TX, [2]University of Texas Health Science Center at Houston, Human Genetics Center, School of Public Health, Houston, TX, [3]Vanderbilt University, Department of Biomedical Informatics, Nashville, TN

Numerous high throughput omics studies have been conducted in schizophrenia, providing an accumulated catalogue of susceptible variants and genes. The results from these studies, however, are highly heterogeneous and complex. The variants and genes nominated by genetic studies often have limited overlap with those from transcriptomic and epigenetic studies, and vice versa. There is thus a pressing need for integrative analysis to unify the different types of data and provide a convergent view of candidate genes in schizophrenia. In this study, we collected a comprehensive, multidimensional dataset, including 7475 brain-expressed genes. The data hosted genome-wide association evidence in genetics (e.g., genome-wide association studies, linkage studies, copy number variations, *de novo* mutations), epigenetics (methylation), transcriptomes, and literature mining. We developed a method named Mega-analysis of Odds Ratio (MegaOR) to prioritize candidate genes. Application of MegaOR in the multidimensional data resulted in consensus sets of candidate genes associated with schizophrenia (SZgenes), which were enriched with condensed evidence from multi-dimensions. We proved that our SZgenes had highly tissue-specific expression in brain and nerve and had intensive interactions significantly higher than randomly expected. Furthermore, we found that our SZgenes were involved in the human brain development by showing strong spatiotemporal expression patterns, which were replicated in independent brain expression data sets. Finally, we found the SZgenes were enriched in critical functional gene sets involved in neuronal activities, ligand gated ion signaling, and Fragile X mental retardation protein targets. In summary, these results assessed rich association evidence to demonstrate a converged set of schizophrenia candidate genes, whose characteristics provided insights into the pathophysiology underlying schizophrenia.

# USING HUMAN GENETICS TO GIVE THE RIGHT PATIENT THE RIGHT DRUG: INFLAMMATORY BOWEL DISEASE AS AN ILLUSTRATIVE EXAMPLE

Jeffrey Barrett

Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom

Nearly 85% of candidate drugs that enter clinical trials fail, often because poor efficacy is discovered only after progression to expensive late phase trials. Fewer new medicines are coming to market, and those that do are more expensive because their success must pay for themselves as well as the development costs of the failures. Even for well-known, approved medicines we don't understand why they're effective in some patients but not others.

Genetics can be used to identify new drug targets, and to stratify patients for existing treatments. Because of the unambiguous causality of genetic associations, they can be used to test whether modulating a particular target will achieve therapeutic benefit in a particular disease. Indeed, previous estimates have suggested that even relatively basic application of GWAS can double the success rates for candidate drugs.

Inflammatory bowel disease (IBD) has been one of the most successfully genetically characterized complex diseases, and offers great opportunity for using these discoveries to find new therapeutic targets and stratify patients. I will describe how we are building on the 250 GWAS-implicated loci in inflammatory bowel disease to connect prioritized genes to cell-specific functions, and how these approaches can be the basis of a framework for the identification and prioritisation of new targets. I will show how open innovation partnerships involving pharma industry scientists working closely with academics can best bring these cutting edge datasets to bear on the problem. For example, the Open Targets platform, which is open to users worldwide, in industry or academia, integrates a dozen databases for prioritising targets in a single analysis framework enabled by new statistical techniques and disease ontologies.

I will also present the largest study in the world of the genetics of response to anti-TNF treatment in IBD. Anti-TNF medications are among the most widely prescribed drugs in the world, and have transformative benefits to patients with a range of diseases, including IBD, rheumatoid arthritis, and psoriasis. Despite this efficacy, nearly half of IBD patients on anti-TNF to not achieve long term remission. We followed up a cohort of >1600 patients for over a year to try to understand genetic and other factors that predict loss of response. We found an association to HLA-DQA1*05 that is strong enough to be potentially clinically useful in adjusting treatment approach.

Overall IBD offers an excellent illustration of both the achievements and promise of using genomics to improve the development and prescription of medicines.

# TRANS-ACTING EFFECTS ON GENE EXPRESSION DRIVE OMNIGENIC INHERITANCE OF COMPLEX TRAITS

Jonathan K Pritchard[1], Xuanyao Liu[2], Yang I Li[2]

[1]Stanford University, Biology, Genetics, & HHMI, Stanford, CA,
[2]University of Chicago, Human Genetics & Genetic Medicine, Chicago, IL

In a 2017 paper we showed that, for typical complex traits, most of the genome is tightly linked to variation that affects trait inheritance and, furthermore, that most of the trait heritability is mediated through genes that do not have close functional relationships to the trait in question. We argued these observations do not fit neatly within standard conceptual models of human genetics, and that the field needs to develop new frameworks for understanding GWAS data. As part of our "omnigenic model", we proposed that the genetic basis of complex traits is due to the combined action of core genes, which have direct effects on the trait, and peripheral genes, which are trans-acting regulators. Perhaps unexpectedly, most heritability is driven by peripheral genes.

Here we present a more formal theoretical model that relates core and peripheral genes in complex traits to cis- and trans-variation affecting gene expression (i.e., eQTLs). Under this model, the key role for peripheral genes is a predictable consequence of known features of trans-acting variation in the heritability of gene expression. In particular, we find that if core genes for a trait tend to be co-regulated--as seems highly likely--then these co-regulated networks act as amplifiers for peripheral variation such that nearly all of the genetic variance is transferred to peripheral genes. Qualitatively similar results are obtained in other models that do not hypothesize a special class of core genes. We further model the extent to which core genes are likely to produce the largest signals in GWAS studies, and end by discussing approaches for finding core genes.

# REPROGRAMMED TRANSLATION OF THE INSERTION SEQUENCE ISTRA IN THE PATHOGENIC BACTERIUM STREPTOCOCCUS PYOGENES

Yun-Juan Bao, Victoria A Ploplis, Francis J Castellino

University of Notre Dame, W.M. Keck Center for Transgene Research, Notre Dame, IN

The short mobile elements or insertion sequences are commonly found to infect prokaryotes. The insertion sequences infecting prokaryotes usually encode transposase genes that autonomously catalyze the binding, cleavage, and strand transfer during the transposition reactions. The autonomous transposition activity has been shown to confer functional advantages and genetic diversity to the host genomes by promoting acquisition of accessory genes or shuffling the host genome structures.

In the efforts to systematically identify the insertion sequences and their evolutionary dynamics in bacterial species, we discovered a unique family of insertion sequences IStra in the pathogenic bacterium Streptococcus pyogenes. IStra frequently occurs in this genome and causes extensive structural rearrangements. These rearrangements lead to a highly asymmetrical genome architecture rarely observed in prokaryotes.

Sequence characterizations showed that the sequences of IStra contain variable C-termini by translating distinct flanking sequences downstream the insertion sites. The sequence variability at the C-termini was caused by reprogramming translation via a 1-bp deletion at the 3'-terminus of the gene. The reprogrammed translation results in polymorphism of the C-terminal protein sequences. Tests of selection pressure based on the relative rates of non-synonymous and synonymous substitution rate predicted that the C-terminal region of IStra is probably under relaxed functional constraints, thus allowing more amino acid changes in this region and flexible insertion sites without affecting the transposition activity. Furthermore, the lineage-wide evolutionary dynamic analysis indicated that the flexible C-termini of IStra might be a host-adaptation consequence conferring advantages for the host bacterium based on two evidences: (1) The sequences of IStra are preferentially inserted in conserved genomic regions neighboring housekeeping genes, such as tRNAs, rRNAs, or ribosomal proteins without interrupting the translation of these housekeeping genes. (2) Estimation of the ages of IStr indicated that IStra may have been inserted in the genus Streptococcus very recently roughly 1000 years ago, close to the time of the first record of the occurrence of scarlet fever, a typical disease phenotype caused by Streptococcus pyogenes.

Further experimental validation using recombinant fluorescent proteins supported our hypothesis. In vitro expression of the fusion protein of IStra and the fluorescent protein in E. coli showed that the copies of IStra were translated into different varieties of C-termini albeit with varied intensity. Therefore, we demonstrated a paradigm of reprogrammed C-terminal translation of the mobile element IStra in prokaryotes.

# POPULATION STRUCTURE OF THE HUMAN GUT MICROBIOME ACROSS ETHNICALLY DIVERSE SUB-SAHARAN AFRICANS

Matthew Hansen*[1], Meagan A Rubel*[1], Aubrey G Bailey[2,3], Alessia Ranciaro[1], Simon R Thompson[1,4], Michael C Campbell[1,5], William Beggs[1], Jaanki R Dave[1,6], Elizabeth Eyermann[1], George Mokone[7], Sununguko W Mpoloka[8], Thomas Nyambo[9], Christian Abnet[10], Stephen J Chanock[10], Frederic D Bushman[2], Sarah A Tishkoff[1,11]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]University of Pennsylvania, Microbiology, Philadelphia, PA, [3]University of North Carolina, Vironomics Core, Chapel Hill, NC, [4]Queen Mary University of London, Genomics, London, United Kingdom, [5]Howard University, Biology, Washington D.C., DC, [6]The Commonwealth Medical College, Biology, Scranton, PA, [7]University of Botswana School of Medicine, Biomedical Sciences, Gaborone, Botswana, [8]University of Botswana, Biological Sciences, Gaborone, Botswana, [9]Muhimbili University of Health and Allied Sciences, BiochemistryDar es Salaam, Tanzania, [10]National Cancer Institute, Division of Cancer Epidemiology and Genetics, Bethesda, MD, [11]University of Pennsylvania, Biology, Philadelphia, PA

Background: Analysis of the gut microbiota of rural African populations provides insight into functions of the healthy human gut that may be altered in industrialized western populations. Here we analyze fecal microbiota of seven ethnically diverse rural populations in Tanzania and Botswana (N=114) with hunter-gatherer, pastoralist, or agropastoralist subsistence practices, and compare them to a Western population. A subset were also genotyped, allowing association of ethnicity and gut microbiome composition.

Results: Most Africans had a high abundance of bacteria from the genus Prevotella, while the U. S. population was high in Bacteroides. Within Africa, there were significant differences in microbiota composition between hunter-gatherer and agropastoralist populations, and between sexes in the Hadza and the Maasai. The Hadza gut bacteria were outliers in comparison with the other populations, with the most unique microbiomes and the highest abundances of Prevotella and Treponema. The gut bacteria of the San hunter-gatherers in Botswana and Burunge agropastoralists in Tanzania had an unusually broad range of bacterial community compositions. Modeling of bacterial gene content suggested that U. S. samples were enriched compared to African samples in pathways for metabolism of environmental toxins and industrial pollutants. Host BMI was negatively correlated with bacterial taxonomic diversity. Structure of the microbiota was correlated with host genotype and geographic separation between ethnic groups.

*Authors contributed equally

# TRANSLATING METAGENOMICS

Ami S Bhatt

Stanford University, Palo Alto, CA

There are more than 1,000 species of bacteria, viruses and fungi that live in the human gut. Far from being passive passengers, these organisms strongly interact with host metabolism, the immune system, and more. For all of this interaction, the dynamics between human hosts and bacteria (microbiome) has only been explored in earnest for the last fifteen to twenty years. Compelling early experiments have shown that intestinal microbiome composition is associated with obesity, cardiovascular diseases, and the effectiveness of certain cancer chemotherapies. Therefore, understanding the impact of microbiomes speciation on noncommunicable diseases such as cancer, hematological and cardiometabolic disorders is fundamental to our health care. But how does one begin to model the dynamics of >1,000, mostly un-sequenced species and strains of bacteria, viruses and fungi? In this presentation, I will discuss two approaches that our translational laboratory has developed and applied - these novel molecular and computational tools allow us to study strain level dynamics of the microbiome, to understand how microbial genomes change over time, and predict the functional output of microbiomes. In particular, I will introduce (a) a Read Cloud assembly approach that allows improved generation of genomes from metagenomic sequencing, and (b) MetaRiboSeq - a method that allows for the indirect measurement of protein abundances by quantification of translated transcripts by adapting ribosomal profiling to metagenomic mixtures.

# COMMON GENETIC VARIATION CONTRIBUTES TO RISK OF RARE DEVELOPMENTAL DISORDERS

Mari E Niemi[1], Hilary C Martin[1], Scott Gordon[2], Kerrie McAloney[2], Sui Yu[3], Nicholas G Martin[2], Jozef Gecz[3], Matthew E Hurles[1], Jeffrey C Barrett[1]

[1]Wellcome Trust Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]QIMR Berghofer Medical Research Institute, Brisbane, Australia, [3]University of Alelaide, Adelaide, Australia

Most known genetic causes of severe childhood developmental disorders (DDs) are rare, deleterious, protein-coding changes that cause Mendelian disorders. Here, we show that common genetic variation also affects risk of severe DDs.

The Deciphering Developmental Disorders (DDD) Study has recruited 14,000 families with a child suffering from a previously undiagnosed DD. These patients are generally severely affected with abnormalities in multiple organ systems. We previously exome sequenced patients and their parents, and showed that that 50% of patients carry a diagnostic *de novo* coding mutation. To investigate the role of common variation in DDs, we conducted a GWAS of 6,987 DDD patients (GBR ancestry) with neurodevelopmental abnormalities, and 9,270 ancestry-matched controls from UK Household Longitudinal Study. The proportion of variation in 'DD risk' attributable to common variants was 13.8% (SE=3.7%), which contradicts the general assumption that such severe DDs are purely Mendelian. We confirmed this finding by showing over-transmission of DD-risk alleles in 728 independent DDD trios (P=0.007).

We found that polygenic DD risk was significantly negatively genetically correlated with educational attainment (EA) ($r_g$=-0.45, P=$6.3x10^{-8}$) and intelligence ($r_g$=-0.44, P=$2.2x10^{-5}$), and positively with schizophrenia ($r_g$=0.29, P=$5.9x10^{-5}$). This suggests that alleles associated with neurodevelopment in the general population contribute to rare DD risk. We replicated depletion of EA- and intelligence-increasing alleles (P=$3.9x10^{-4}$ and $7.6x10^{-4}$ respectively) in an independent cohort of 1,270 intellectual disability cases from Australia, and 1,688 ancestry-matched controls.

Furthermore, we found no difference in polygenic burden between patients with diagnostic coding variants and those without, suggesting that common variant risk is not confined to patients without a monogenic diagnosis. We found that patients with mild/moderate developmental delay had lower polygenic EA scores than patients with severe delay (P=$2.0x10^{-4}$), suggesting that common variants are a larger contributor to milder DDs.

Overall, our results demonstrate that common genetic variation plays a role in disorders traditionally viewed as purely monogenic. This may have important implications for understanding variable clinical presentation and searching for secondary genetic modifiers.

# VARIATION IN MICROBIOME COMPOSITION IMPACTS HUMAN GENE EXPRESSION BY CHANGING CHROMATIN ACCESSIBILITY

Allison Richards[1], Amanda Muehlbauer[2], Francesco Messina[1], Adnan Alazizi[1], Michael Burns[2], Trevor Gould[2], Camilla Cascardo[1], Roger Pique-Regi[1], Ran Blekhman[2], Francesca Luca[1]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Minnesota, Department of Genetics, Cell Biology and Development, Minneapolis, MN

Variation in gut microbiome is associated with human disease, yet the underlying molecular mechanisms are not well understood. Recent studies have shown that the microbiome may cause changes in gene regulation in interfacing host epithelial cells in the gut. Here, we treated colonic epithelial cells with live microbiota from five healthy individuals and quantified changes in transcriptional regulation and chromatin accessibility in host cells.

We identified over 5,000 host genes that change expression, including 588 distinct associations between specific taxa and host genes (FDR<10%), corresponding to 121 host genes with changes in expression associated with the abundance of 46 taxa. Remarkably, we reproduced 24 of these associations *in vivo*, in colonic tissue biopsies from 15 healthy individuals. Five of the six taxa with the strongest influence on gene expression (>30 associated genes) alter the response of genes associated with complex traits (OR>2.7, p-value<0.005). To validate and further demonstrate the effect of specific microbes on host gene expression, we focused on a single microbe, *Collinsella aerofaciens*, which we found to be correlated with the expression of multiple relevant genes. We created a manipulated microbial community with titrated doses of *Collinsella aerofaciens*, and identified 1,570 genes differentially expressed (FDR<10%), including 19 out of the originally identified 29 genes (OR=4.1, p-value=0.0002). To investigate the molecular mechanism for gene regulatory changes in response to the microbiome, we used ATAC-seq and identified 234 regions that were differentially accessible between the treated samples and untreated controls. We found an enrichment for differentially accessible regions within 10kb of differentially expressed genes (OR=2.22, p-value=0.013). Differential footprinting analysis identified footprints induced by the microbiome treatment for several factors (e.g., MEF2A and SUM1 motifs or equivalent homologs).

Together, our results show that specific microbes play an important role in regulating expression of individual host genes involved in human complex traits. Our work also supports the hypothesis that one of the mechanisms by which the microbiome regulates host gene expression is through changes in chromatin accessibility and transcription factor binding in host cells. Finally, the ability to fine tune the expression of host genes by manipulating the microbiome suggests future therapeutic routes for human wellness.

**BRAINSEQ PHASE II**: SCHIZOPHRENIA-ASSOCIATED EXPRESSION DIFFERENCES BETWEEN THE HIPPOCAMPUS AND THE DORSOLATERAL PREFRONTAL CORTEX

Leonardo Collado-Torres, Emily E Burke, Joo Heon Shin, Stephen A Semick, BrainSeq Consortium, Ran Tao, Amy Deep-Soboslay, Thomas M Hyde, Joel E Kleinman, Daniel R Weinberger+, Andrew E Jaffe+

Lieber Institute for Brain Development, Translational Sciences, Baltimore, MD

+ Co-corresponding authors.

**Background**: We previously identified widespread genetic, developmental, and schizophrenia-associated changes in polyadenylated RNAs in the dorsolateral prefrontal cortex (DLPFC), but the landscape of hippocampal (HIPPO) expression using RNA sequencing is less well-explored.

**Methods**: We performed RNA-seq using RiboZero on 900 tissue samples across 551 individuals (286 with schizophrenia) in DLPFC (N=453) and HIPPO (N=447). We quantified expression of multiple feature summarizations of the Gencode v25 reference transcriptome, including genes, exons and splice junctions. Within and across brain regions, we modeled age-related changes in controls using linear splines, integrated genetic data to perform expression quantitative trait loci (eQTL) analyses, and performed differential expression analyses controlling for observed and latent confounders.

**Results**: We identified widespread transcriptional regulation in the DLPFC and the hippocampus over development, with 10,807 genes differentially expressed across age and brain regions (Bonferroni < 0.01) that are nominally replicated in BrainSpan (n = 79 tissue samples, 40 DLPFC and 39 HIPPO). Of these genes, 5,931 (~55%) contain differentially expressed exons and splice junctions that replicated in BrainSpan. We found 197 genes differentially expressed (at FDR < 5%) in hippocampus between schizophrenia patients and non-psychiatric controls, with 133 (~68%) genes additionally implicated via exonic differential expression. Of these 197 genes, only 45 (~22%) nominally replicated in DLPFC (at $p < 0.05$), suggesting regional heterogeneity of the molecular correlates of schizophrenia diagnosis. We characterized extensive genetic regulation of gene expression with significantly different effects across the two brain regions: we identified 115,787 region-dependent eQTLs (at FDR < 1%), corresponding to 1,484 genes, at the gene, exon, or splice junction level (99 across all three). These region-dependent eQTLs included clinically relevant risk variants - five schizophrenia risk loci showed significant differential regional regulation.

**Discussion**: We show extensive regional specificity of developmental and genetic regulation, and schizophrenia-associated expression differences between the hippocampus and DLPFC. These findings suggest that some components of the transcriptional correlates of developmental and genetic risk for schizophrenia may be brain region-specific.

# TECHNICAL ABSTRACTS
# FOR WORKSHOPS

INTERESTED IN HEARING ABOUT NANOPORE SEQUENCING AT BIOLOGY OF GENOMES?

Hear about the latest technology updates from Oxford Nanopore as well as hearing from Winston Timp, Johns Hopkins University, about his latest project using nanopore sequencing.

James Brayer, Oxford Nanopore Technologies
Winston Timp, Johns Hopkins University:

Applications of modification detection in nanopore sequencing
Nanopore sequencing is a single molecule characterization method, allowing direct DNA sequencing of stretches ranging from kilobases to even megabase long reads. Unlike traditional sequencing-by-synthesis methods, it can distinguish covalently modified nucleotides directly through their modulation of the electrolytic current. And the long reads allow for straightforward detection of structural variations, large insertions, deletions or transpositions that are often difficult to detect with short-read sequencing.
We demonstrate the power of exogenous labeling to perform an integrative, single molecule characterization of the epigenome. Specifically, we have used M.CviPI, a GpC methyltransferase, to label accessible DNA in cancer and normal cell lines. This allows for us to simultaneously correlate nucleosome positioning and the native CpG methylation along long stretches of the genome (~10kb) on a single molecule. Using this methodology, we also assess the variability of the epigenetic patterns between cancer and normal cell lines.
We have also made significant progress towards more general base modification training, specifically attempting to train different bacterial methyltransferases provided from our collaborators at NEB. We demonstrate initial results from this work including N6-methyladenine, 5-methylcytosine in non-CG motifs, and 4-methylcytosine. An immediate application of these new trained models is polishing of nanopore sequencing genome assemblies. Typical assembly polishing leveraging the electrical data to correct errors has difficulty in areas corresponding to methylation motifs. We demonstrate incorporating the methylation trained models allows higher accuracy assemblies.
Finally, we have begun to apply Cas9-based sequencing enrichment to nanopore sequencing as part of our library preparation methodology. For a first pass, we examined the promoter region of human telomerase (hTERT) in thyroid cancer derived cell lines; this region frequently either has mutations, aberrant DNA methylation, or both. This region is particularly hard to probe with conventional bisulfite amplicon sequencing due to its high level of CG dinucleotides. We demonstrate focused coverage of long native nanopore sequencing reads in this region, measuring single molecule methylation patterns and SNVs alike.

**NOTES**

**NOTES**

**NOTES**

**NOTES**

**NOTES**

## Participant List

Dr. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Mr. Shaked Afik
UC Berkeley
safik@berkeley.edu

Ms. Ipsita Agarwal
Columbia University
ia2337@columbia.edu

Dr. Keiko Akagi
MD Anderson Cancer Center
kakagi@mdanderson.org

Ms. Noor Al Hajri
Sidra Medicine
nalhajri@gmail.com

Ms. Shouq Alanezi
Catholic University of America
galbooshi@hotmail.com

Mr. Patrick Albers
University of Oxford
patrick.albers@bdi.ox.ac.uk

Dr. Jessica Alfoldi
Broad Institute of MIT and Harvard
jalfoldi@broadinstitute.org

Ms. Nayra Al-Thani
Weill Cornell Medicine Qatar
nma2008@qatar-med.cornell.edu

Ms. Maris Alver
Institute of Genomics, University of Tartu
maris.alver@ut.ee

Mr. Lin An
Pennsylvania State University
lua137@psu.edu

Dr. Simeon Andrews
Weill Cornell Medicine - Qatar
andrewsssa@gmail.com

Dr. Barbara Arbeithuber
Pennsylvania State University
bxa15@psu.edu

Dr. Alexander Arguello
National Institute of Mental Health
alexander.arguello@nih.gov

Dr. Maria Artesi
University of Liege
maria.artesi@gmail.com

Ms. Kathryn Asalone
American University
ka5144a@student.american.edu

Ms. Samira Asgari
Harvard Medical School
sasgari@broadinstitute.com

Mr. Tal Ashuach
UC Berkeley
tal_ashuach@berkeley.edu

Mr. Ziga Avsec
Technical University Munich
zigaavsec@gmail.com

Dr. Julien Ayroles
Princeton University
jayroles@princeton.edu

Dr. Courtney Babbitt
University of Massachusetts, Amherst
cbabbitt@bio.umass.edu

Dr. Taejeong Bae
Mayo Clinic
bae.taejeong@mayo.edu

Ms. Erin Baggs
Earlham institute
erin.baggs@earlham.ac.uk

Dr. Orli Bahcall
Nature
o.bahcall@us.nature.com

Dr. Brunilda Balliu
Stanford University School of Medicine
bballiu@stanford.edu

Dr. Sara Ballouz
CSHL
sballouz@cshl.edu

Dr. Jessie YJ Bao
University of Notre Dame
ybao2@nd.edu

Dr. Damir Baranasic
Imperial College London
damir.baranasic@imperial.ac.uk

Dr. Jeffrey Barrett
Wellcome Trust Sanger Institute
jb26@sanger.ac.uk

Mr. Rajbir Batra
University of Cambridge
rajbir.batra@cruk.cam.ac.uk

Dr. Alexis Battle
Johns Hopkins University
ajbattle@jhu.edu

Dr. Serafim Batzoglou
Illumina, Inc.
sbatzoglou@illumina.com

Dr. Sarah Bay
Genetics Society of America
sbay@genetics-gsa.org

Ms. Avery Davis Bell
Harvard Medical School
averydavis@fas.harvard.edu

Dr. Pascal Belleau
Cold Spring Harbor Laboratory
belleau@cshl.edu

Dr. Shifra Ben-Dor
Weizmann Institute of Science
shifra.ben-dor@weizmann.ac.il

Dr. Tone Berge
OsloMet - Oslo Metropolitan University
tone.berge@hioa.no

Dr. Semir Beyaz
Cold Spring Harbor Laboratory
semirbeyaz@gmail.com

Mr. Vivek Bhardwaj
MPI of Immunobiology and Epigenetics
bhardwaj@ie-freiburg.mpg.de

Dr. Ami Bhatt
Stanford University
asbhatt@stanford.edu

Dr. Preetida Bhetariya
University of Utah
preetida.bhetariya@utah.edu

Dr. Wendy Bickmore
MRC Human Genetics Unit
Wendy.Bickmore@igmm.ed.ac.uk

Dr. Tim Bigdeli
SUNY Downstate Medical Center
kim.brown@downstate.edu

Mr. Kostas Billis
EMBL EBI
kbillis@ebi.ac.uk

Dr. Minou Bina
Purdue University
bina@purdue.edu

Dr. Jef Boeke
New York University Langone Medical Center
Jef.Boeke@nyumc.org

Dr. Lorenzo Bomba
Wellcome Trust Sanger Institute
lb17@sanger.ac.uk

Ms. Beatrice Borsari
Centre for Genomic Regulation (CRG)
beatrice.borsari@crg.eu

Dr. Lara Bossini-Castillo
Wellcome Trust Sanger Institute
lbc@sanger.ac.uk

Ms. Megan Bowler
University of Utah
megan.bowler@genetics.utah.edu

Dr. Harrison Brand
Massachusetts General Hospital/Harvard Medical Sch
hbrand1@mgh.harvard.edu

Dr. James Brayer
Oxford Nanopore Technologies Ltd
james.brayer@nanoporetech.com

Ms. Katie Brenner
Cold Spring Harbor Laboratory
kbrenner@cshl.edu

Ms. Gwenna Breton
Human Evolution
gwenna.breton@ebc.uu.se

Dr. Christopher Brown
University of Pennsylvania
chrbro@pennmedicine.upenn.edu

Dr. Andrew Brown
University of Geneva
andrew.brown@unige.ch

Mr. John Bryan
National Eye Institute, NIH
John.Bryan@nih.gov

Dr. Suhong Bu
Fujian Agriculture and Forestry University
busuhong@163.com

Ms. Emily Burke
Lieber Institute for Brain Development
emily.burke@libd.org

Mr. Diego Calderon
Stanford University
dcal@stanford.edu

Dr. Minal Caliskan
University of Pennsylvania
minal.caliskan@gmail.com

Ms. Bobbie Cansdale
University of Sydney
bobbie.cansdale@sydney.edu.au

Dr. Lucia Carbone
Oregon Health & Science University
carbone@ohsu.edu

Mr. David Carlson
Stony Brook University
david.carlson@stonybrook.edu

Dr. Piero Carninci
RIKEN Center for Life Science Technologies
carninci@riken.jp

Dr. Andrew Carroll
DNAnexus
acarroll@dnanexus.com

Dr. Anne-Ruxandra Carvunis
University of Pittsburgh
anc201@pitt.edu

Dr. Mahul Chakraborty
University of California Irvine
mahulchak@gmail.com

Ms. Pritha Chanana
Mayo Clinic
chanana.pritha@mayo.edu

Dr. Lesley Chapman
National Institute of Standards and Technology
lmc2@nist.gov

Dr. Alexander Charney
Mount Sinai
alexander.charney@mssm.edu

Dr. Sidi Chen
Yale University
sidi.chen@yale.edu

Mr. Surya Chhetri
The University of Alabama in Huntsville
sbc0011@uah.edu

Prof. Jeffrey Chuang
The Jackson Laboratory
jeff.chuang@jax.org

Dr. Deanna Church
10x Genomics Inc.
deanna.church@10xgenomics.com

Dr. Leonardo Collado Torres
Lieber Institute for Brain Development
leo.collado@libd.org

Dr. Mahlon Collins
University of Minnesota
mahlon@umn.edu

Dr. Vincenza Colonna
National Research Council
vincenza.colonna@igb.cnr.it

Dr. Chris Cotsapas
Yale School of Medicine
cotsapas@gmail.com

Dr. Mark Cowley
Garvan Institute of Medical Research
m.cowley@garvan.org.au

Dr. Sarah Craig
Penn State University
sarah@bx.psu.edu

Dr. Joanna Cross
National Institutes of Health
Joanna.Cross@nih.gov

Dr. Megan Crow
Cold Spring Harbor Laboratory
mcrow@cshl.edu

Ms. Hannah Currant
EMBL-EBI
currant@ebi.ac.uk

Ms. Ciara Curtin
GenomeWeb
ccurtin@genomeweb.com

Mr. Matthew Dapas
Northwestern University
matthew.dapas@northwestern.edu

Dr. Gargi Dayama
University of Minnesota
gdayama@umn.edu

Dr. Aaron Day-Williams
Merck
aaron.day-williams@merck.com

Dr. Jacob Degner
Abbvie
jacob.degner@abbvie.com

Dr. Job Dekker
University of Massachusetts Medical School
job.dekker@umassmed.edu

Dr. Ricardo del Rosario
Broad Institute
rcdelros@broadinstitute.org

Dr. Olivier Delaneau
University of Geneva
olivier.delaneau@unige.ch

Dr. Jennifer DeLeon
Genome Research
deleon@cshl.edu

Ms. Danielle Denisko
University of Toronto
danielle.denisko@mail.utoronto.ca

Dr. Emmanouil Dermitzakis
University of Geneva
emmanouil.dermitzakis@unige.ch

Dr. Scott Devine
University of Maryland School of Medicine
sdevine@som.umaryland.edu

Dr. Julia di Iulio
The Scripps Research Institute
jdiiulio@scripps.edu

Dr. Federica Di Palma
Ealrham Institute
federica.di-palma@earlham.ac.uk

Ms. Tonya Di Sera
University of Utah
tonyads@genetics.utah.edu

Mr. Joey Mark Diaz
University of Leeds
umjmsd@leeds.ac.uk

Dr. Junjun Ding
Sun Yat-sen Universtiy
dingjunj@mail.sysu.edu.cn

Dr. Sarah Djebali
Inra Genphyse
sarah.djebali-quelen@inra.fr

Dr. Alexander Dobin
CSHL
dobin@cshl.edu

Dr. Olga Dolgova
Population Genomics Group
olga.dolgova@cnag.crg.eu

Mr. Ricardo D'Oliveira Albanus
University of Michigan
albanus@umich.edu

Dr. Egor Dolzhenko
Illumina, Inc
edolzhenko@illumina.com

Dr. Paul Donat
Stony Brook University
paul.donat@stonybrook.edu

Prof. Peter Donnelly
University of Oxford
donnelly@well.ox.ac.uk

Mr. Max Dougherty
University of Washington
mldough@uw.edu

Mr. Timothy Dreszer
Stanford University
tdreszer@stanford.edu

Dr. Olga Dudchenko
Baylor College of Medicine
olga.dudchenko@bcm.edu

Mr. Noah Dukler
Cold Spring Harbor Labs
ndukler@cshl.edu

Dr. Keith Durkin
University of Liege
keithdurkin@gmail.com

Dr. Michael Eberle
Illumina, Inc
meberle@illumina.com

Dr. Aditya Ekawade
University of Utah School of Medicine
aditya.ekawade@gmail.com

Ms. Susan Fairley
EMBL-EBI
fairley@ebi.ac.uk

Dr. Shaohua Fan
University of Pennsylvania
shaohuaf@pennmedicine.upenn.edu

Dr. Emma Farley
University of California San Diego
efarley@ucsd.edu

Dr. Andrew Farrell
University of Utah
jandrewrfarrell@gmail.com

Dr. Elise Feingold
NIH/National Human Genome Research Institute
feingole@mail.nih.gov

Ms. Lynn Fellman
Fellman Studios
lynn@fellmanstudio.com

Dr. Adam Felsenfeld
National Institutes of Health
adam_felsenfeld@nih.gov

Ms. Anna Fijarczyk
Universite Laval
aniafijarczyk@gmail.com

Mr. Gregory Findlay
University of Washington
gf2@uw.edu

Mr. Anthony Findley
Wayne State University
afindley@med.wayne.edu

Mr. Ryan Fischer
Parent Project Muscular Dystrophy
Ryan@parentprojectmd.org

Dr. Tomas Fitzgerald
The European Bioinformatics Institute (EMBL-EBI)
tomas@ebi.ac.uk

Dr. Paul Flicek
EMBL-EBI
flicek@ebi.ac.uk

Dr. Liliana Florea
Johns Hopkins School of Medicine
florea@jhu.edu

Ms. Ana Flores-Bojorquez
University of California San Diego
abojorqu@ucsd.edu

Ms. Talitha Forcier
Cold Spring Harbor Laboratory
talitha@cshl.edu

Dr. Mallory Freeberg
Johns Hopkins University
mfreeberg@ebi.ac.uk

Dr. Laure Fresard
Stanford University School of Medicine
lfresard@stanford.edu

Dr. Julien Gagneur
Technical University Munich
gagneur@in.tum.de

Mr. Jonathan Gaige
Stony Brook University
jonathan.gaige@stonybrook.edu

Dr. Pedro Galante
Hospital Sirio Libanes
pgalante@mochsl.org.br

Dr. David Galas
Pacific Northwest Research Institute
djgalas@gmail.com

Mr. Craig Gambogi
Univeristy of Pennsylvania
gambogi@pennmedicine.upenn.edu

Dr. Dadi Gao
Massachusetts General Hospital
dgao2@mgh.harvard.edu

Dr. Kate Gao
Springer Nature
kate.gao@us.nature.com

Dr. Eugene Gardner
Wellcome Sanger Institute
eg15@sanger.ac.uk

Mr. Diego Garrido
Centre for Genomic Regulation (CRG)
diego.garrido@crg.eu

Mr. Erik Garrison
Wellcome Trust Sanger Institue
erik.garrison@gmail.com

Ms. Molly Gasperini
University of Washington
gasperim@uw.edu

Mr. Evan Geller
Yale University
evan.geller@yale.edu

Mr. Giulio Genovese
Broad Institute
giulio@broadinstitute.org

Dr. Michel Georges
University of Liège
michel.georges@ulg.ac.be

Ms. Stephanie Georges
University of Utah
stephanie.georges@genetics.utah.edu

Dr. Mark Gerstein
Yale University
pi@gersteinlab.org

Mr. Marius Gheorghe
University of Oslo
marius.gheorghe@ncmm.uio.no

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. Daniel Gilchrist
NIH/National Human Genome Research Institute
daniel.gilchrist@nih.gov

Mr. Manraj Gill
Cold Spring Harbor Laboratory
mgill@cshl.edu

Ms. Emily Glassberg
Stanford University
eglassbe@stanford.edu

Ms. Dafni Glinos
Wellcome Trust Sanger Institute
dg22@sanger.ac.uk

Mr. Rohan Gnanaolivu
Mayo Clinic
gnanaolivu.rohandavid@mayo.edu

Dr. Andreas Gnirke
Broad Institute
gnirke@broadinstitute.org

Mr. Alexander Godfrey
MIT/Whitehead Institute
agodfrey@wi.mit.edu

Dr. Alon Goren
UCSD
agoren@ucsd.edu

Dr. Liangke (Connie) Gou
UCLA
lgou@ucla.edu

Dr. Julie Granka
AncestryDNA
jgranka@ancestry.com

Mr. Jonathan Griffiths
CRUK Cambridge Institute
jag216@cam.ac.uk

Dr. Jeremy Grushcow
Sequence Bio
jeremy@sequencebio.com

Dr. Rodrigo Gularte Merida
Memorial Sloan Kattering Cancer Center
gularter@mskcc.org

Dr. Brad Gulko
Cornell University/CSHL
bgulko@cs.cornell.edu

Dr. Meenal Gupta
University of Utah
meenal002@gmail.com

Dr. Melissa Gymrek
University of California San Diego
mgymrek@ucsd.edu

Dr. Wilfried Haerty
Earlham Institute
wilfried.haerty@earlham.ac.uk

Mr. Vincent Hahaut
University of Liege
vincent.hahaut@uliege.be

Dr. Iman Hajirasouliha
Weill Cornell Medicine of Cornell University
imh2003@med.cornell.edu

Dr. Christopher Hammell
Cold Spring Harbor Laboratory
chammell@cshl.edu

Mr. Bob Handsaker
Broad Institute
handsake@broadinstitute.org

Dr. Roberta Hannibal
Second Genome
roberta@secondgenome.com

Dr. Matthew Hansen
University of Pennsylvania
mhansen@pennmedicine.upenn.edu

Dr. Ross Hardison
The Pennsylvania State University
rch8@psu.edu

Mr. Simon Hardwick
Garvan Institute of Medical Research
s.hardwick@garvan.org.au

Dr. Ronald Harris
Baylor College of Medicine
rharris1@bcm.edu

Dr. Christopher Hart
Ionis Pharmaceuticals
chart@ionisph.com

Dr. Shinichi Hashimoto
Kanazawa University
hashimoto@med.kanazawa-u.ac.jp

Prof. Masahira Hattori
Waseda University
m-hattori@aoni.waseda.jp

Mr. James Havrilla
University of Utah
semjaavria@gmail.com

Ms. Laura Hayward
Columbia university
lauhayward@gmail.com

Dr. Tim Hefferon
NIH
theffero@mail.nih.gov

Dr. Hussein Hejase

Mr. Pyry Helkkula
Institute for Molecular Medicine Finland
pyry.helkkula@helsinki.fi

Dr. Javier Herrero
UCL Cancer Institute
javier.herrero@ucl.ac.uk

Mr. Andrew Hill
University of Washington
ajh24@uw.edu

Dr. Gabriel Hoffman
Icahn School of Medicine at Mount Sinai
gabriel.hoffman@mssm.edu

Dr. Eurie Hong
Ancestry
ehong@ancestry.com

Dr. Yi-Fei Huang
Cold Spring Harbor Laboratory
yihuang@cshl.edu

Dr. Xiaomeng Huang
University of Utah
xm01.huang@gmail.com

Dr. Zhuoyi Huang
Berry Genomics
joeyfuerimmer@gmail.com

Mr. Christopher Hubel
King's College London
christopher.huebel@kcl.ac.uk

Dr. Matthew Hurles
Wellcome Trust Genome Campus
meh@sanger.ac.uk

Dr. Carolyn Hutter
National Human Genome Research Institute
carolyn.hutter@nih.gov

Ms. Elizabeth Hutton
Cold Spring Harbor Laboratory
ehutton@cshl.edu

Dr. Trey Ideker
University of California, San Diego
trey.ideker@gmail.com

Dr. Hae Kyung Im
The University of Chicago
haky@uchicago.edu

Dr. Marcin Imielinski
Weill Cornell Medicine
mai9037@med.cornell.edu

Ms. Rosario Isasi
University of Miami
Risasi@miami.edu

Dr. Sadahiro Iwabuchi
Kanazawa University
s_iwabuchi@staff.kanazawa-u.ac.jp

Dr. Lakshmanan Iyer
LabCorp Inc
iyerl@labcorp.com

Mr. Ray Jackson
DuPont / Industrial Biosciences
raymond.e.jackson@dupont.com

Dr. Mattias Jakobsson
Uppsala University
mattias.jakobsson@ebc.uu.se

Dr. Jeffrey Jensen
Arizona State University
Jeffrey.D.Jensen@asu.edu

Dr. Chao Jiang
Stanford University
jiangch@stanford.edu

Dr. Xin Jin
BGI-Shenzhen
jinxin@genomics.cn

Mr. Thomas Juettemann
EMBL EBI
juettemann@ebi.ac.uk

Dr. Goo Jun
University of Texas Health Science Center Houston
goo.jun@uth.tmc.edu

Dr. Irwin Jungreis
MIT
iljungr@csail.mit.edu

Dr. Monica Justice
Hospital for Sick Children
monica.justice@sickkids.ca

Dr. Tugce Karaderi
Eastern Mediterranean University
tugce@well.ox.ac.uk

Dr. Konrad Karczewski
Broad Institute
kon10004@gmail.com

Ms. Snehal Karpe
National Centre for Biological Sciences(NCBS-TIFR)
karpesnehal@gmail.com

Ms. Yukie Kashima
The University of Tokyo
0799458414@edu.k.u-tokyo.ac.jp

Dr. Hideya Kawaji
RIKEN
kawaji@gsc.riken.jp

Ms. Katarzyna Kedzierska
University of Virginia
kzk5f@virginia.edu

Dr. Manolis Kellis
MIT / Harvard
manoli@mit.edu

Mr. Derek Kelly
University of Pennsylvania
derkelly@pennmedicine.upenn.edu

Ms. Elsa Kentepozidou
EMBL EBI
elsa@ebi.ac.uk

Dr. Peter Kerpedjiev
Harvard Medical School
pkerp@hms.harvard.edu

Ms. Akanksha Khare
Agilent Technologies
akanksha.khare@agilent.com

Dr. Daehwan Kim
University of Texas Southwestern Medical Center
Daehwan.Kim@UTSouthwestern.edu

Ms. Young Kim
Stony Brook University
young.c.kim@stonybrook.edu

Dr. Jaemin Kim
National Institutes of Health
jaemin.kim@nih.gov

Dr. Sarah Kim-Hellmuth
New York Genome Center
skim@nygenome.org

Mr. James King
MRC London Institute of Medical Sciences
jk2014@ic.ac.uk

Dr. Amnon Koren
Cornell University
koren@cornell.edu

Dr. Ilya Korsunsky
Harvard Medical School
ilya.korsunsky@gmail.com

Dr. Ksenia Krasileva
Earlham institute
ksenia.krasileva@earlham.ac.uk

Dr. Vivek Kumar
CSHL
vkumar@cshl.edu

Dr. Anshul Kundaje
Stanford University
akundaje@stanford.edu

Dr. Kousik Kundu
WellCome Sanger Institute
kk8@sanger.ac.uk

Mr. Shyam Lakshmanan
National Eye Institute, NIH
shyam.lakshmanan@nih.gov

Dr. Jason Lambert
Univeresity of California, Davis
jtlambert@ucdavis.edu

Dr. Eric Lander
The Broad Institute of MIT & Harvard
lander@broadinstitute.org

Dr. Billy Lau
Stanford University
billylau@stanford.edu

Dr. Ryan Layer
University of Utah
ryan.layer@gmail.com

Dr. Amanda Lea
Princeton University
amandalea7180@gmail.com

Mr. Dillon Lee
University of Utah
dlee123@gmail.com

Dr. Christina Leslie
Memorial Sloan Kettering Cancer Center
cleslie@cbio.mskcc.org

Dr. Stephen Levene
The University of Texas at Dallas
stephen.levene@utdallas.edu

Dr. Dawei Li
University of Vermont
dawei.li@uvm.edu

Dr. Robert Li
USDA-ARS
robert.li@ars.usda.gov

Ms. Jingqiu Liao
Cornell University
jl3374@cornell.edu

Dr. Jayon Lihm
Cold Spring Harbor Laboratory
jlihm@cshl.edu

Ms. Emi Ling
Harvard Medical School
eling@fas.harvard.edu

Dr. Leonard Lipovich
Wayne State University
llipovich@med.wayne.edu

Dr. Jianjun Liu
Genome Institute of Singapore
liuj3@gis.a-star.edu.sg

Dr. Yunlong Liu
Indiana University Purdue University - Indianapolis
yunliu@iu.edu

Dr. Nicole Lockhart
NIH/NHGRI
lockhani@mail.nih.gov

Dr. Quan Long
University of Calgary
quan.long@ucalgary.ca

Dr. Francesca Luca
Wayne State University
fluca@wayne.edu

Dr. Jianzhu Ma
University of California, San Diego
majianzhu@gmail.com

Dr. Daniel MacArthur
Broad Institute
danmac@broadinstitute.org

Dr. Niklas Mahler
Umea University
niklas.mahler@umu.se

Ms. Lauren Mak
University of Calgary
lauren.mak@ucalgary.ca

Dr. Joel Malek
Weill Cornell Medicine - Qatar
mar2078@qatar-med.cornell.edu

Ms. Naveen Malik
Black Hills State University
naveen.malik@yellowjackets.bhsu.edu

Ms. Shani Marom
Ben Gurion university
shanimarom1@gmail.com

Prof. Gabor Marth
University of Utah
gmarth@genetics.utah.edu

Dr. Hilary Martin
Wellcome Sanger Institute
hcm@sanger.ac.uk

Dr. Baishali Maskeri
National Institutes of Health-NIH
maskerib@mail.nih.gov

Mr. Bansho Masutani
University of Tokyo
banmasutani@gmail.com

Dr. Debra Mathews
Johns Hopkins University
dmathews@jhmi.edu

Dr. Eric Mbunwe
University of Pennsylvania
embunwe@gmail.com

Mr. Christopher McAllester
University of Wisconsin-Madison
cmcallester@gmail.com

Dr. Davis McCarthy
EMBL-EBI
davis@ebi.ac.uk

Dr. Richard McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Dr. David McGaughey
National Eye Institute, NIH
david.mcgaughey@nih.gov

Mr. James McKenna
Uppsala University
jamesmck2013@gmail.com

Dr. David McKinnon
Stony Brook University
david.mckinnon@stonybrook.edu

Dr. Francis McMahon
National Institutes of Health
mcmahonf@mail.nih.gov

Dr. John McPherson
UC Davis
jdmcpherson@ucdavis.edu

Dr. Stephen Meyn
University of Wisconsin-Madison, SMPH
stephen.meyn@wisc.edu

Dr. Chase Miller
University of Utah, Center for Genetic Discovery
chmille4@gmail.com

Dr. Ryan Mills
University of Michigan
remills@umich.edu

Dr. Josine Min
University of Bristol
Josine.Min@Bristol.ac.uk

Dr. Dan Mishmar
Ben-Gurion University of the Negev
dmishmar@bgu.ac.il

Dr. Arman Mohammad
The Broad Institute
mohammad@broadinstitute.org

Dr. Pejman Mohammadi
The Scripps Research Institute
pejman@scripps.edu

Mr. Jaaved Mohammed
Stanford University
jaavedm@stanford.edu

Prof. Michele Morgante
Universita di Udine
michele.morgante@uniud.it

Dr. Shinichi Morishita
University of Tokyo
moris@edu.k.u-tokyo.ac.jp

Prof. Leonid Moroz
University of Florida
moroz@whitney.ufl.edu

Dr. Matthew Moscou
The Sainsbury Laboratory
matthew.moscou@tsl.ac.uk

Dr. Jonathan Mudge
EMBL EBI
jmudge@ebi.ac.uk

Dr. Kamaldeen Muili
Bowling Green State University
kamuili@bgsu.edu

Dr. Kasper Munch
Aarhus University
kaspermunch@birc.au.dk

Mr. Manuel Munoz
Centre for Genomic Regulation (CRG)
manuel.munoz@crg.eu

Ms. Priyanka Nakka
Brown University
priyanka_nakka@brown.edu

Mr. Sahin Naqvi
Whitehead Institute/MIT
sahin.naqvi@gmail.com

Mr. Fabio Navarro
Yale University
fabio.navarro@yale.edu

Dr. Tal Nawy
Nature Research
t.nawy@us.nature.com

Dr. Holly Neibergs
Washington State University
neibergs@wsu.edu

Mr. Dominic Nelson
McGill University
dominic.nelson@mail.mcgill.ca

Dr. Huck-Hui Ng
Genome Institute of Singapore
nghh@gis.a-star.edu.sg

Dr. Thomas Nicholas
University of Utah
thomas.nicholas@utah.edu

Ms. Mari Niemi
Wellcome Trust Sanger Institute
mn2@sanger.ac.uk

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. Jim Notwell
Circuit Therapeutics
jnotwell@circuittx.com

Mr. Ninad Oak
Baylor College of Medicine
ninad.oak@bcm.edu

Ms. Arisa Oda
University of Tokyo
odar@g.ecc.u-tokyo.ac.jp

Dr. Charles O'Donnell
Marauder
odonnell@maraudertx.com

Dr. Anne O'Donnell-Luria
Broad Institute, Boston Children's Hospital
odonnell@broadinstitute.org

Ms. Meritxell Oliva
University Of Chicago
meritxellop@uchicago.edu

Ms. M. Kathrina Onate
Stanford University
kconate@stanford.edu

Ms. Yoko Ono
Kyowa Hakko Kirin Co., Ltd
yoko.ono@kyowa-kirin.co.jp

Dr. Roel Ophoff
UCLA
ophoff@ucla.edu

Dr. Jakub Orzechowski Westholm
Science for Life Laboratory
jakub.westholm@scilifelab.se

Mr. Omead Ostadan
Illumina, Inc.
oostadan@illumina.com

Dr. Elaine Ostrander
National Institutes of Health
eostrand@mail.nih.gov

Dr. David Page
Whitehead Institute; MIT/HHMI
dcpage@wi.mit.edu

Dr. Luisa Pallares
Princeton Univeristy
pallares@princeton.edu

Dr. Aarno Palotie
Massachusetts General Hospital
aarno@broadinstitute.org

Dr. Brent Pedersen
University of Utah
bpederse@gmail.com

Dr. Dana Pe'er
Sloan Kettering Institute, MSKCC
peerster@gmail.com

Dr. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Dr. Dmitri Pervouchine
Skolkovo Institute for Science and Technology
d.pervouchine@skoltech.ru

Dr. Susanne Pfeifer
Arizona State University
Susanne.Pfeifer@asu.edu

Dr. Lon Phan
NIH
lonphan@mail.nih.gov

Dr. Adam Phillippy
National Human Genome Research Institute
adam.phillippy@nih.gov

Dr. Luca Pinello
Massachusetts General Hospital/ Harvard
lpinello@mgh.harvard.edu

Mr. Yishay Pinto
Bar-Ilan University
yipinto@gmail.com

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Dr. Sharon Plon
Baylor College of Medicine
splon@bcm.edu

Prof. Christopher Ponting
University of Edinburgh
chris.ponting@igmm.ed.ac.uk

Ms. Maryam Pouryahya
Stony Brook University
maryam.pouryahya@gmail.com

Dr. Jonathan Pritchard
Stanford University
pritch@stanford.edu

Dr. Teresa Przytycka
NIH/NLM/NCBI
przytyck@mail.nih.gov

Dr. Yi Qiao
University of Utah
yi.qiao@genetics.utah.edu

Prof. Francis Quetier
GENOPOLE-evry & University EVRY
francis.quetier@genopole.fr

Dr. Aaron Ragsdale
McGill University
aaron.ragsdale@mail.mcgill.ca

Dr. Guillaume Ramstein
Cornell University
gr226@cornell.edu

Dr. Joshua Randall
Wellcome Trust Sanger Institute
jr17@sanger.ac.uk

Dr. Gunnar Ratsch
ETH Zurich
raetsch@ethz.ch

Dr. Sarah Ratzel
American Journal of Human Genetics
sratzel@ajhg.net

Dr. Parisa Razaz
MGH/Harvard/Broad
prazaz@mgh.harvard.edu

Mr. Guillermo Reales
Universidade Federal do Rio Grande do Sul
grealesm@gmail.com

Ms. Bethany Reman
Black Hills State University
bethany.reman@yellowjackets.bhsu.edu

Dr. Mark Reppell
AbbVie Inc
mark.reppell@abbvie.com

Dr. Arang Rhie
NIH
arang.rhie@nih.gov

Dr. Samuli Ripatti
University of Helsinki
samuli.ripatti@helsinki.fi

Dr. Jose Rodriguez-Martinez
University of Puerto Rico - Rio Piedras
jose.rodriguez233@upr.edu

Dr. Christian Roedelsperger
Max Planck Institute Tuebingen
christian.roedelsperger@tuebingen.mpg.de

Dr. Jeffrey Rogers
Baylor College of Medicine
jr13@bcm.edu

Ms. Masa Roller
EMBL-EBI
roller@ebi.ac.uk

Dr. Mostafa Ronaghi
Illumina, Inc.
mronaghi@illumina.com

Ms. Kate Rosenbloom
UC Santa Cruz Genomics Institute
kate@soe.ucsc.edu

Dr. Jeffrey Rosenfeld
Rutger Cancer Institute of NJ
jeffrey.rosenfeld@rutgers.edu

Dr. Tanmoy Roychowdhury
Mayo Clinic
roychowdhury.tanmoy@mayo.edu

Dr. Joel Rozowsky
Yale University
joel.rozowsky@yale.edu

Dr. Douglas Ruderfer
Vanderbilt University Medical Center
douglas.ruderfer@vanderbilt.edu

Dr. Lucky Runtuwene
The University of Tokyo
luckyruntuwene@edu.k.u-tokyo.ac.jp

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Dr. Tim Sands
Genome Biology
tim.sands@biomedcentral.com

Dr. Sriram Sankararaman
UCLA
sriram.sankararaman@gmail.com

Mr. Sergio Santos
EMBL-EBI
ssantos@ebi.ac.uk

Mr. Thomas Sasani
University of Utah
tom.sasani@utah.edu

Dr. Fah Sathirapongsasuti
23andMe
fsathira@23andMe.com

Dr. Michael Schatz
CSHL and JHU
mschatz@cshl.edu

Prof. Mikkel Heide Schierup
Aarhus University
mheide@birc.au.dk

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@mail.nih.gov

Dr. Gary Schroth
Illumina, Inc.
gschroth@illumina.com

Dr. Gregory Schwartz
University of Pennsylvania
gsch@pennmedicine.upenn.edu

Dr. David Schwartz
University of Wisconsin - Madison
dcschwartz@wisc.edu

Dr. Fritz Sedlazeck
Baylor College of Medicine
Fritz.Sedlazeck@bcm.edu

Mr. Stephen Semick
Lieber Institute for Brain Development
stephen.semick@libd.org

Mr. Sarun Sereewattanawoot
The University of Tokyo
sereewattanawoot_sarun_15@stu-cbms.k.u-tokyo.ac.jp

Dr. Eilon Sharon
Stanford University
eilon@stanford.edu

Ms. Anna Shcherbina
Stanford University
annashch@stanford.edu

Mr. Max Shen
Massachusetts Institute of Technology
maxwshen@mit.edu

Dr. Ning Shen
Fulcrum Therapeutics
nshen@fulcrumtx.com

Dr. Gavin Sherlock
Stanford University
gsherloc@stanford.edu

Dr. Xinghua Shi
University of North Carolina at Charlotte
x.shi@uncc.edu

Dr. Atsushi Shimizu
Iwate Medical University
ashimizu@iwate-med.ac.jp

Mr. Patrick Short
Wellcome Trust Sanger Institute
ps14@sanger.ac.uk

Dr. Massa Shoura
Stanford University
massa86@stanford.edu

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Michelle Simon
MRC Harwell Institute
M.Simon@har.mrc.ac.uk

Ms. Nicolette Sipperly
Stony Brook University
nicolette.sipperly@stonybrook.edu

Dr. Alexei Slesarev
BioReliance MilliporeSigma
alexei.slesarev@sial.com

Ms. Pola Smirin-Yosef
Ariel University
psmirin@gmail.com

Dr. Michael Smith
NIH
smithmw@mail.nih.gov

Dr. Matthew Smonskey
Thermo Fisher Scientific
matthew.smonskey@thermofisher.com

Dr. Michael Snyder
Stanford University
mpsnyder@stanford.edu

Dr. Wei Song
NIH
songw2@mail.nih.gov

Prof. Nicole Soranzo
Wellcome Trust Sanger Institute
ns6@sanger.ac.uk

Dr. John Stamatoyannopoulos
Altius Institute for Biomedical Sciences
jstam@altius.org

Dr. Dwight Stambolian
University of Pennsylvania
stamboli@pennmedicine.upenn.edu

Dr. Oliver Stegle
European Molecular Biology Laboratory
oliver.stegle@ebi.ac.uk

Dr. Arnold Stein
Purdue University
steina@purdue.edu

Dr. Barbara Stranger
University of Chicago
bstranger@medicine.bsd.uchicago.edu

Dr. J Seth Strattan
Stanford University
jseth@stanford.edu

Mr. Benjamin Strober
Johns Hopkins University
bstrober3@gmail.com

Ms. Meena Subramaniam
UCSF
meena.subramaniam@ucsf.edu

Dr. Wataru Suda
RIKEN
wataru.suda@riken.jp

Ms. Ida Surakka
University of Michigan
isurakka@umich.edu

Dr. Hillary Sussman
Genome Research, Executive Editor
hsussman@cshl.edu

Dr. Yoichi Sutoh
Iwate Medical University
ysutoh@medicalgenome.info

Mr. Yuta Suzuki
The University of Tokyo
esperhacone@gmail.com

Dr. Lena Takayasu
RIKEN
lena.takayasu@riken.jp

Dr. Michael Talkowski
Massachusetts General Hospital and Harvard Medical
mtalkowski@mgh.harvard.edu

Dr. Jiaying Tan
CELL PRESS
jtan@cell.com

Ms. Catherine Tang
Stony Brook University
catherine.tang.1@stonybrook.edu

Dr. Marcela Tello-Ruiz
Cold Spring Harbor Laboratory
mmonaco@cshl.edu

Mr. Rohit Thakur
University of Leeds
umrth@leeds.ac.uk

Dr. James Thomas
NIH
thomasjw4@mail.nih.gov

Dr. Geng Tian
Geneis (Beijing) Co,. Ltd
chenlf@geneis.cn

Dr. Winston Timp
Johns Hopkins University
wtimp@jhu.edu

Mr. Shubhakar Reddy Tipireddy
University of Tokyo
r.shubhakar@edu.k.u-tokyo.ac.jp

Dr. Sarah Tishkoff
University of Pennsylvania
tishkoff@mail.med.upenn.edu

Dr. Daniel Trejo Banos
University of Lausanne
daniel.trejobanos@unil.ch

Dr. Richard Trembath
King's College London
dean-folsm@kcl.ac.uk

Dr. Sebastian Treusch
Mission Bio Inc
treusch@missionbio.com

Dr. Jennifer Troyer
National Human Genome Research Institute
troyerj@mail.nih.gov

Dr. David Truong
New York University Langone Medical Center
David.Truong@nyumc.org

Dr. Gosia Trynka
Wellcome Trust Sanger Institute
gosia@sanger.ac.uk

Dr. Christopher Tuggle
Iowa State University
cktuggle@iastate.edu

Dr. Taru Tukiainen
Institute for Molecular Medicine Finland FIMM
taru@atgu.mgh.harvard.edu

Dr. Taichi Umeyama
RIKEN
taichi.umeyama@riken.jp

Ms. Lara Urban
EMBL-EBI
lara.h.urban@ebi.ac.uk

Dr. Fyodor Urnov
Altius Institute for Biomedical Sciences
urnov@altius.org

Dr. Flora Vaccarino
Yale University
flora.vaccarino@yale.edu

Dr. Anton Valouev
Grail Bio
valouev@gmail.com

Dr. Cristopher Van Hout
Regeneron
cristopher.vanhout@regeneron.com

Ms. Arushi Varshney
University of Michigan
arushiv@umich.edu

Dr. Krishna Veeramah
Stony Brook University
krishna.veeramah@stonybrook.edu

Dr. Matthew Velinder
University of Utah
matt.velinder@utah.edu

Dr. Ana Vinuela
University of Geneva
ana.vinuela@unige.ch

Dr. Irina Voineagu
University of New South Wales
i.voineagu@unsw.edu.au

Dr. Urmo Vosa
University Medical Center Groningen
urmo.vosa@gmail.com

Dr. Deven Vyas
Stony Brook Univeristy
deven.vyas@stonybrook.edu

Ms. Alexandria Wade
University of California, Davis
aawade@ucdavis.edu

Dr. Hai Wang
Cornell University
hw449@cornell.edu

Mr. Ming-Qiang Wang
The Chinese University of Hong Kong
mqwang@link.cuhk.edu.hk

Dr. Michelle Ward
University of Chicago
mcward@uchicago.edu

Dr. Alistair Ward
University of Utah
alistairnward@gmail.com

Dr. Jacob Washburn
Cornell University
jdw297@cornell.edu

Dr. Sharon Wei
CSHL
weix@cshl.edu

Dr. Jia Wen
UNC Charlotte
jwen6@uncc.edu

Dr. Sarah Wheelan
The Johns Hopkins University School of Medicine
swheelan@jhmi.edu

Dr. Nicola Whiffin
Imperial College London
n.whiffin@imperial.ac.uk

Ms. Heather Wick
Johns Hopkins University School of Medicine
hwick1@jhmi.edu

Dr. Cristen Willer
University of Michigan
cristen@umich.edu

Dr. Richard Wilson
Nationwide Children's Hospital
richard.wilson@nationwidechildrens.org

Mr. Eamon Winden
University of Wisconsin, Madison
ewinden@wisc.edu

Dr. Kim Worley
Baylor College of Medicine
kworley@bcm.edu

Dr. Tomasz Wrzesinski
Earlham Institute
tomasz.wrzesinski@earlham.ac.uk

Dr. Kou-Juey Wu
China Medical University
wukj@mail.cmu.edu.tw

Dr. Yannick Wurm
Queen Mary U Lonodn
y.wurm@qmul.ac.uk

Dr. Simon Xi
Pfizer
hualin.xi@pfizer.com

Mr. Guanjue Xiang
Pennsylvania State University
gzx103@psu.edu

Ms. Rachita Yadav
Massachusetts General Hospital
ryadav1@mgh.harvard.edu

Prof. Huanming Yang
BGI-China
yanghm@genomics.cn

Dr. Tao Yang
The Pennsylvania State University
xadmyangt@gmail.com

Dr. Jimmie Ye
UCSF
jimmie.ye@ucsf.edu

Dr. B. LInju Yen
National Health Research Institutes
blyen@nhri.org.tw

Mr. Jake Yeung
Ecole Polytechnique Federale de Lausanne
jake.yeung@epfl.ch

Dr. Lynn Young
National Institutes of Health
lynny@mail.nih.gov

Dr. Qi Yu
NIH
dryuqi@gmail.com

Dr. Fuli Yu
Berry Genomics
yufuli@berrygenomics.com

Dr. Laura Zahn
AAAS/Science
lzahn@aaas.org

Ms. Samantha Zarate
DNAnexus
szarate@dnanexus.com

Dr. Feng Zhang
The Broad Institute of MIT and Harvard
zhang_f@mit.edu

Dr. Xuefang Zhao
Massachusetts General Hospital
XZHAO12@mgh.harvard.edu

Dr. Zhongming Zhao
University of Texas Health Science Center Houston
zhongming.zhao@uth.tmc.edu

Dr. Yixin Zhao
Cold Spring Harbor Laboratory
yizhao@cshl.edu

Mr. Yuehui Zhao
Georgia Institute of Technology
yzhao349@gatech.edu

Dr. Shida Zhu
BGI-Shenzhen
zhushida@genomics.cn

Dr. Martine Zilversmit
American Museum of Natural History, Central Park W
mzilversmit@amnh.org

Oxford NANOPORE Technologies

# Applications of modification detection in nanopore sequencing

Interested in hearing about nanopore sequencing at Biology of Genomes?

**Attend our lunch workshop on Wednesday 9th May.**

Hear from Winston Timp, Assistant Professor of Biomedical Engineering at Johns Hopkins University, about his latest research using nanopore sequencing plus updates from the Oxford Nanopore Technologies team.

**Register now**

**https://register.nanoporetech.com/biologyofgenomes**

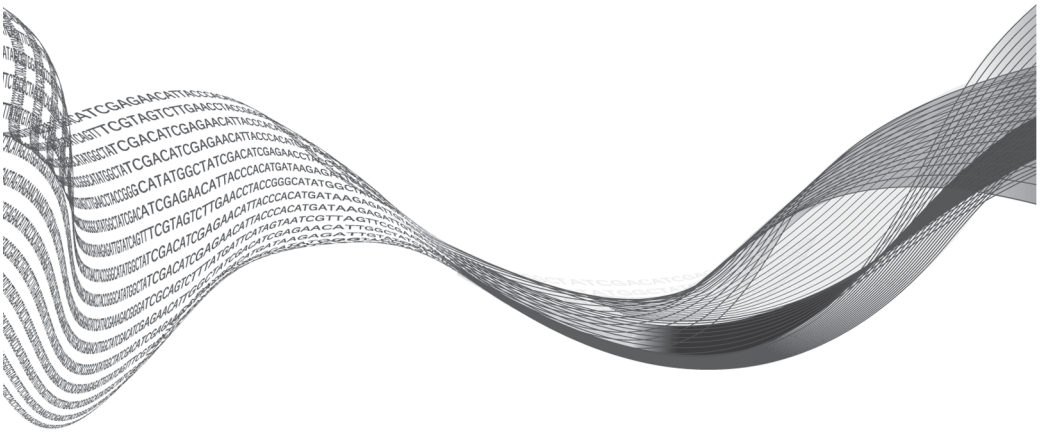
Oxford Nanopore Technologies and the Wheel icon are registered trademarks of Oxford Nanopore Technologies in various countries. All other brands and names contained are the property of their respective owners. © 2018 Oxford Nanopore Technologies. All rights reserved.

illumina®

# We never stop seeking

We are driven to know more—to relentlessly search for the answers that will advance the understanding of genomics to improve human health. And we realize we can't do it alone. We're counting on the next generation of scientific minds to help us keep up the momentum. As the leading developer of genomic solutions and services, we help accelerate genetic research and its use in the fields of cancer, hereditary disease, reproductive health, infectious disease, and agriculture. Together, we'll realize the promise of personalized medicine.

**www.lllumina.com**



©2018 Illumina, Inc. All rights reserved.
All trademarks are the property of Illumina, Inc. or their respective owners. For specific trademark information, see www.illumina.com/company/legal.html.

## 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**<br>**Huntington Hospital**<br>270 Park Avenue, Huntington | **631-351-2300**<br>**(1037)** |
| **Dentists**<br>Dr. William Berg<br>Dr. Robert Zeman | <br>**631-271-2310**<br>**631-271-8090** |
| **Doctor**<br>mediCenter<br>365 W. Jericho Tpke., Huntington | **631-423-5400**<br>(**1034**) |
| **Drugs - 24 hours, 7 days**<br>Eckerd<br>391 W. Main Street, Huntington | **631-549-9400**<br>(**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*** - Obtain PIN from Meetings & Courses Office to enter Library after hours. See Library staff for photocopier code.

**Computers, E-mail, Internet access**

Grace Auditorium
Upper level: E-mail only
Lower level: Word processing and printing.
STMP server address: mail.optonline.net
*To access your E-mail, you must know the name of your home server.*

**Dining, Bar**

Blackford Hall

Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late

***Helpful tip*** - If there is a line at the upper dining area, try the lower dining room

**Messages, Mail, Faxes**

Message Board, Grace, lower level

**Swimming, Tennis, Jogging, Hiking**

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m.
Two tennis courts open daily.

**Pool Table, Ping Pong, Television**

Blackford Hall Rec Room - downstairs

**Russell Fitness Center**

Dolan Hall, west wing, lower level
***Helpful tip*** – Obtain PIN from Meetings & Courses Office

**Concierge**

**On duty daily at Meetings & Courses Office.**
***After hours – From tan house phones, dial x8870 for assistance***

**Pay Phones, House Phones**

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

| **1-800 access #'s:** | |
|---|---|
| **AT&T** | **9-1-800-321-0288** |
| **MCI** | **9-1-800-674-7000** |

**Local Interest**

| | |
|---|---|
| Fish Hatchery | 631-692-6768 |
| Sagamore Hill | 516-922-4447 |
| Whaling Museum | 631-367-3418 |
| Heckscher Museum | 631-351-3250 |
| CSHL DNA Learning Center | x 5170 |

**New York City**

***Helpful tip -***

Take Syosset Taxi to Syosset Train Station
($8.00 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**

| | | |
|---|---|---|
| Classic Limousine | 631-567-5100 | (**1033**) |
| Syosset Limousine | 516-364-9681 | (**1031**) |
| To head west of CSHL - Syosset train station | | |
| Syosset Taxi | 516-921-2141 | (**1030**) |
| To head east of CSHL - Huntington Village | | |
| Orange & White Taxi | 631-271-3600 | (**1032**) |
| Executive Limo | 631-696-8000 | (**1047**) |

**Trains**

| | |
|---|---|
| Long Island Rail Road | 822-LIRR |
| *Schedules available from the Meetings & Courses Office.* | |
| Amtrak | 800-872-7245 |
| MetroNorth | 800-638-7646 |
| New Jersey Transit | 201-762-5100 |

**Ferries**

| | |
|---|---|
| Bridgeport / Port Jefferson | 631-473-0286 **(1036)** |
| Orient Point/ New London | 631-323-2525 **(1038)** |

**Car Rentals**

| | |
|---|---|
| Avis | 631-271-9300 |
| Enterprise | 631-424-8300 |
| Hertz | 631-427-6106 |

**Airlines**

| | |
|---|---|
| American | 800-433-7300 |
| America West | 800-237-9292 |
| British Airways | 800-247-9297 |
| Continental | 800-525-0280 |
| Delta | 800-221-1212 |
| Japan Airlines | 800-525-3663 |
| Jet Blue | 800-538-2583 |
| KLM | 800-374-7747 |
| Lufthansa | 800-645-3880 |
| Northwest | 800-225-2525 |
| United | 800-241-6522 |
| US Airways | 800-428-4322 |