Abstracts of papers presented
at the 2019 meeting on

# THE BIOLOGY OF GENOMES

May 7–May 11, 2019

Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2019 meeting on

# THE BIOLOGY OF GENOMES

May 7–May 11, 2019

Arranged by

Matthew Hurles, *Wellcome Trust Sanger Institute*
Elaine Ostrander, *National Human Genome Research Institute*
Dana Pe'er, *Sloan Kettering Institute*
Jay Shendure, *University of Washington*

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

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

**Corporate Benefactors**

Estée Lauder Companies
Regeneron
Thermo Fisher Scientific

**Corporate Sponsors**

Agilent Technologies
Bayer
Bristol-Myers Squibb Company
Calico Labs
Celgene
Genentech, Inc.
Merck
New England BioLabs
Pfizer

**Corporate Partners**

Alexandria Real Estate
Enzo Life Sciences
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:* Illustration by Anna Minkina, University of Washington, Seattle.

**THE BIOLOGY OF GENOMES**
Tuesday, May 7 – Saturday, May 11, 2019

| | | |
|---|---|---|
| Tuesday | 7:30 pm | **1** Functional Genomics |
| Tuesday | *following eve. session* | *Happy Hour*<br>*Sponsored by PacBio* |
| Wednesday | 9:00 am | **2** Computational Genomics |
| Wednesday | 2:00 pm | **3** Poster Session I |
| Wednesday | 4:30 pm | *Wine and Cheese Party* |
| Wednesday | 7:30 pm | **4** Cancer and Medical Genomics |
| Thursday | 9:00 am | **5** Population Genomics |
| Thursday | 2:00 pm | **6** Poster Session II |
| Thursday | 4:30 pm | **7** ELSI Panel and Discussion |
| Thursday | 7:30 pm | **8** Evolutionary Non-Human Genomics |
| Friday | 9:00 am | **9** Genome Engineering and Editing |
| Friday | 12:00 pm | *Meet the Speakers Lunch* |
| Friday | 2:00 pm | **10** Poster Session III |
| Friday | 4:30 pm | GUEST SPEAKERS |
| Friday | 6:00 pm | *Cocktail Hour*<br>*Sponsored by Thermo Fisher* |
| | 7:00 pm | Banquet Dinner |
| Saturday | 9:00 am | **11** Complex Traits and Microbiome |

*Workshops (immediately following morning sessions)*
PacBio, Wednesday *(p. T-1)*
Oxford Nanopore, Thursday *(p. T-2)*

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

Cold Spring Harbor Laboratory is committed to maintaining a safe and respectful environment for all meeting attendees, and does not permit or tolerate discrimination or harassment in any form. By participating in this meeting, you agree to abide by the Code of Conduct, which is available both online and at the back of this book.

---

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 7—7:30 PM

**SESSION 1** FUNCTIONAL GENOMICS

**Chairpersons:** **Arjun Raj,** University of Pennsylvania, Philadelphia
**Helena Kilpinen,** University College London, United Kingdom

Arjun Raj.
Presenter affiliation: University of Pennsylvania, Philadelphia, Pennsylvania. 1

**Decoding regulatory DNA sequence and regulatory networks in keratinocyte differentiation**
Daniel S. Kim, Viviana Risca, Arwa Kathiria, Adam Rubin, James Chappell, Howard Chang, Michael Snyder, Will Greenleaf, Paul Khavari, Anshul Kundaje.
Presenter affiliation: Stanford University, Stanford, California. 2

**Joint assay of single cell RNA-seq and transcription factor binding**
Arnav Moudgil, Michael N. Wilkinson, Xuhua Chen, June He, Alex J. Cammack, Michael J. Vasek, Tomas Lagunas, Zongtai Qi, Samantha A. Morris, Joseph D. Dougherty, Robi D. Mitra.
Presenter affiliation: Washington University in St. Louis School of Medicine, St. Louis, Missouri. 3

**Long-range single-molecule mapping reveals the multikilobase-scale organization of accessible chromatin**
Georgi K. Marinov, Zohar Shiphony, Matthew P. Swaffer, Anshul Kundaje, William J. Greenleaf.
Presenter affiliation: Stanford University, Stanford, California. 4

**Cell type-specific genetic regulation of expression in the granule cell layer of the human dentate gyrus**
Andrew E. Jaffe, Daniel J. Hoeppner, Takeshi Saito, Blanpain Lou, Joy Ukaigwe, Emily E. Burke, Ran Tao, Katsunori Tajinda, Amy Deep-Soboslay, Joo Heon Shin, Joel E. Kleinman, Daniel R. Weinberger, Mitsuyuki Matsumoto, Thomas M. Hyde.
Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland; Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland. 5

**Robustness of gene expression to stochastic, environmental and genetic variation is encoded in the genome**
Olga Sigalova, Mattia Forneris, Katharina Bender, Judith Zaugg, Eileen E. Furlong.
Presenter affiliation: European Molecular Biology Laboratory, Heidelberg, Germany. 6

**Massively parallel characterization of regulatory dynamics during neural induction**
Anat Kreimer, Fumitaka Inoue, Tal Ashuach, Nadav Ahituv, Nir Yosef.
Presenter affiliation: University of California, Berkeley, Berkeley, California; University of California San Francisco, San Francisco, California. 7

**Functional annotation of long non-coding RNAs based on their spatial localization**
Saumya Agrawal, Jordan Ramilowski, Chi Wai Yip, Chung Chau Hon, Kayoko Yasuzawa, Masayoshi Ito, Takeya Kasukawa, Naoto Kondo, Ken Yagi, Harukazu Suzuki, Josée Dostie, Jay W. Shin, Michiel de Hoon, Piero Carninci.
Presenter affiliation: RIKEN, Yokohama, Japan. 8

**Happy Hour**
*Sponsored by* **PacBio**

WEDNESDAY, May 8—9:00 AM

**SESSION 2** COMPUTATIONAL GENOMICS

**Chairpersons:** **Eleazar Eskin,** University of California, Los Angeles
**Barbara Engelhardt,** Princeton University, New Jersey

**Leveraging alleleic hetegeneity in association studies**
Eleazar Eskin, Farhad Hormozdiari, Jong Wha Joo.
Presenter affiliation: University of California, Los Angeles, Los Angeles, California. 9

**Human genetics of complex retinal anatomy revealed by large scale imaging**
Hannah Currant, Tomas Fitzgerald, Pirro G. Hysi, Charles A. Reisman, Qi Yang, Christopher J. Hammond, Peng T. Khaw, Paul J. Foster, Praveen Patel, Anthony P. Khawaja, Ewan Birney.
Presenter affiliation: European Molecular Biology Laboratory (EMBL), Cambridge, United Kingdom. 10

**A rapid phylogeny-based method for accurate community profiling of large-scale metagenomic datasets**
Lenore Pipes, Rasmus Nielsen.
Presenter affiliation: University of California-Berkeley, Berkeley, California. 11

**Whole-body maps of postzygotic mutations in humans**
Nicole B. Rockweiler, Avinash Ramu, Ellen Gelfand, Casey Drubin, Ni Huang, Malachi Griffith, Kristin Ardlie, The GTEx Consortium, Donald F. Conrad.
Presenter affiliation: Washington University School of Medicine, St. Louis, Missouri. 12

**Discovering and distinguishing as many cell types as possible**
Barbara Engelhardt, Bianca Dumitrascu, Karen Feng, Soledad Villar.
Presenter affiliation: Princeton University, Princeton, New Jersey.

**100 genomes in 100 days—The structural variant landscape in tomato genomes**
Michael C. Schatz, Sebastian Soyk, Xingang Wang, Michael Alonge, Zachary Lemmon, Srividya Ramakrishnan, Sara Goodwin, Fritz Sedlazeck, W. Richard McCombie, Esther van der Knaap, Joyce Van Eck, Zachary Lippman.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; Johns Hopkins University, Baltimore, Maryland. 13

**Dissecting the genetic architecture of complex traits through local polygenicity using summary statistics from genome-wide association studies**
Ruth D. Johnson, Kathryn Burch, Kangcheng Hou, Bogdan Pasaniuc, Sriram Sankararaman.
Presenter affiliation: University of California, Los Angeles, Los Angeles, California. 14

**scBayes—A computational method to study tumor subclone-specific gene expression and chromatin accessibility using single cell RNA sequencing and single cell ATAC sequencing in combination of bulk DNA sequencing**
Yi Qiao, Xiaomeng Huang, Gabor Marth.
Presenter affiliation: USTAR, Salt Lake City, Utah; University of Utah, Salt Lake City, Utah. 15

WEDNESDAY, May 8—2:00 PM

**SESSION 3** POSTER SESSION I

**Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes**
Haley J. Abel, David E. Larson, Colby Chiang, Indraniel Das, Krishna L. Kanchi, Ryan M. Layer, Benjamin M. Neale, William J. Salerno, Catherine Reeves, Steven Buyske, Tara C. Matise, Donna M. Muzny, Michael C. Zody, Eric S. Lander, Susan K. Dutcher, Nathan O. Stitziel, Ira M. Hall.
Presenter affiliation: Washington University, St. Louis, Missouri. 16

**Chromatin organization modulates the origin of heritable structural variations in human genome**
Alexej Abyzov, Tanmoy Roychowdhury.
Presenter affiliation: Mayo Clinic, Rochester, Minnesota. 17

**Genomic analysis of germline and somatic variation in high-grade serous ovarian cancer**
Aaron W. Adamson, Yuan Chun Ding, Mihaela C. Cristea, Sharon Wilczynski, Mark T. Wakabayashi, Ernest S. Han, Thanh H. Dellinger, Pin S. Lin, Amy A. Hakim, Linda Steele, Charles D. Warden, Shu Tao, Susan L. Neuhausen.
Presenter affiliation: Beckman Research Institute of City of Hope, Duarte, California. 18

**Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks**
Vikram Agarwal, Jay Shendure.
Presenter affiliation: University of Washington, Seattle, Washington. 19

**Coding and noncoding DNA variants underlie *trans*-acting hotspots of gene expression variation**
Sheila Lutz, Margaret Kliebhan, Christian Brion, Frank W. Albert.
Presenter affiliation: University of Minnesota, Minneapolis, Minnesota. 20

**Identifying very rare genetic variants in DNA microarray data using a population-specific reference panel for imputation**
Maris Alver, Timo T. Sikka, Tiit Nikopensius, Mart Kals, Reedik Mägi, Tõnu Esko.
Presenter affiliation: University of Tartu, Tartu, Estonia. 21

**Genome-wide analyses of Ni-Vanuatu populations reveal diverse ancestries and temporal differences in admixture patterns**
Lara R. Arauna, Jeremy Choin, Javier Mendoza-Revilla, Christine Harmant, Laure Lemee, Olivier Cassar, Antoine Gessain, Etienne Patin, Lluis Quintana-Murci.
Presenter affiliation: Institut Pasteur, Paris, France. 22

**Fine-scale population structure and history in British-Pakistanis and implications for disease risk**
Elena Arciero, Theofanis Tsismentzoglou, Dan Mason, Neil Small, Eamonn Sheridan, Richard Trembath, Matthew E. Hurles, David A. van Heel, John Wright, Mark Iles, Hilary C. Martin.
Presenter affiliation: The Wellcome Sanger Institute, Hinxton, United Kingdom. 23

**Single cell multi--omics profiling reveals a hierarchical epigenetic landscape during mammalian germ layer specification**
Ricard Argelaguet, Stephen J. Clark, Hisham Mohammed, Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Yunlong Xiang, Courtney Hanna, Guido Sanguinetti, Felix Krueger, Wei Xie, Peter Rugg--Gunn, Gavin Kelsey, Wendy Dean, Jennifer Nichols, Oliver Stegle, John C. Marioni, Wolf Reik.
Presenter affiliation: European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom. 24

**Best practices for discovering mosaic variants in whole-genome sequencing data**
Taejeong Bae, Liana Fasching, Flora Vaccarino, Alexej Abyzov, Brain Somatic Mosaicism Network.
Presenter affiliation: Mayo Clinic, Rochester, Minnesota. 25

**Strategies for differential omics of responders and non-responders in patient-derived tumor organoids**
Matthew H. Bailey, Brent S. Pedersen, Katrin P. Guillen, Sandra Scherer, Yi Qiao, Xiaomeng Huang, Andrew Butterfield, Alana Welm, Bryan E. Welm, Gabor T. Marth.
Presenter affiliation: University of Utah, Salt Lake City, Utah 26

**Is it time to change the reference genome?**
Sara Ballouz, Alexander Dobin, Jesse Gillis.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 27

**scRNA-seq identifies changes in cell types, cell states, and expression levels in patients with idiopathic pulmonary fibrosis.**
Nicholas E. Banovich, Linh Bui, Austin Gutierrez, Chris A. Habermann, Mei-i Chung, Jonathan A. Kropski.
Presenter affiliation: Translational Genomics Research Institute (TGen), Phoenix, Arizona. 28

**Longitudinal molecular trajectories of diffuse glioma in adults**
Floris P. Barthel, Kevin C. Johnson, Frederick S. Varn, Anzhela D. Moskalik, Georgette Tanner, The GLASS Consortium, Jason T. Huse, John F. de Groot, Lucy F. Stead, Roel G. Verhaak.
Presenter affiliation: The Jackson Laboratory, Farmington, Connecticut; VU Medical Center, Amsterdam, the Netherlands. 29

**Expression of exogenous mRNA in human cancer cell lines leads to RISC-mediated cell death**
William E. Putzbach, Ashley Haluck-Kangas, Elizabeth T. Bartom, Marcus E. Peter.
Presenter affiliation: Northwestern University, Chicago, Illinois. 30

**Inferring African population differences from gene expression**
Sanjit S. Batra, Jonathan Fischer, Matthew Hansen, Derek Kelly, Yun S. Song, Sarah Tishkoff.
Presenter affiliation: University of California, Berkeley, California. 31

**Leveraging public epigenomic datasets to examine the role of regulatory variation in the three-dimensional organization of the genome**
Brittany A. Baur, Shilu Zhang, Sushmita Roy.
Presenter affiliation: University of Wisconsin-Madison, Madison, Wisconsin. 32

**Critical assessment of computational methods to predict the impact of enhancer variation**
Nicholas Eng, Dustin Shigaki, Ayoti Patra, James Chen, Jin-Woo Oh, Wang Xi, Michael A. Beer.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 33

**Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors—diffTF**
Ivan Berest, Christian Arnold, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Kristian Helin, Judith Zaugg.
Presenter affiliation: EMBL, Heidelberg, Germany. 34

**Bioinformatics pipeline for detection and prioritization of very low-frequency somatic cancer mutations in tumor-derived circulating cell-free DNA sequencing data**
Preetida J. Bhetariya, Gabor T. Marth.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 35

**Mercury—High quality visualization and reproducible exploratory analysis of scRNA-seq data**
Daniel Blibaum, Alexander Dobin.
Presenter affiliation: Cold Spring Harbor Lab, Cold Spring Harbor, New York. 36

**Characterizing RNA stability genome wide through combined analysis of PRO-seq and RNA-seq data**
Amit Blumberg, Yifei Huang, Yixin Zhao, Charles G. Danko, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 37

**The fitness landscape of clonal haematopoiesis**
Caroline J. Watson, Alana Papula, Yeuk G. Poon, Wing H. Wong, Andrew L. Young, Todd E. Druely, Daniel S. Fisher, Jamie R. Blundell.
Presenter affiliation: University of Cambridge, Cambridge, United Kingdom. 38

**An integrated index of public gene expression databases and its application for meta-analysis of hypoxic transcriptomes**
Hidemasa Bono.
Presenter affiliation: Research Organization of Information and Systems, Mishima, Japan. 39

**Co-expression patterns define variation-intolerant chromatin regulators associated with neuropsychiatric phenotypes.**
Leandros Boukas, James Havrilla, Peter Hickey, Aaron Quinlan, Hans Bjornsson, Kasper Hansen.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 40

**Assessment of human male gonadal mosaicism reveals mutational origins and stratifies risk of transmission**
Martin W. Breuss, Jonathan Sebat, Joseph Gleeson.
Presenter affiliation: University of California, San Diego, La Jolla, California; Rady Children's Institute for Genomic Medicine, San Diego, California. 41

**Mapping *trans*-acting genetic variation in gene expression by simultaneous quantification of mRNA and protein levels in single cells**
Christian Brion, Sheila Lutz, Frank W. Albert.
Presenter affiliation: University of Minnesota, Minneapolis, Minnesota. 42

**Comparison of gene- and exon-based cis-eQTLs in human subcutaneous adipose tissue**
S M. Brotman, C K. Raulerson, A Ko, P Pajukanta, M Laakso, K L. Mohlke.
Presenter affiliation: UNC, Chapel Hill, North Carolina. 43

**Determining the impact of missed inversions in the human genome by droplet digital PCR (ddPCR)**
Marta Puig, Jon Lerga-Jaso, Carla Giner-Delgado, Sarai Pacheco, David Izquierdo, Alejandra Delprat, Jack F. Regan, George Karlin-Neumann, Mario Cáceres.
Presenter affiliation: Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain; Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. 44

**Common mitochondrial DNA variants implicated in variations in mitochondrial oxidative activity**
Na Cai, Aurora Gomez-Duran, Ekaterina Yonova-Doing, Marc Jan Bonder, Bogdan Mirauta, John Danesh, Joanna Howson, Oliver Stegle, Patrick Chinnery, Nicole Soranzo.
Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom; European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom. 45

**Tracking the interdisciplinarity and misappropriation of genomics research through social media**
Jedidiah Carlson, Kelley Harris.
Presenter affiliation: University of Washington, Seattle, Washington. 46

**Identifying *de novo* noncoding mutations associated with autism spectrum disorders**
Christopher Castro, Shengcheng Dong, Stephanie Bielas, Alan P. Boyle.
Presenter affiliation: The University of Michigan, Ann Arbor, Michigan. 47

**The genomic architecture of introgression among sibling species of bacteria**
Maria Izabel A. Cavassim, Sara Moeskjær, Camous Moslemi, Bryden Fields, Asger Bachmann, Bjarni Vilhjalmsson, Mikkel H. Schierup, J Peter W. Young, Stig U. Andersen.
Presenter affiliation: Aarhus University, Aarhus, Denmark. 48

**A map of mitotic recombinations in Bloom Syndrome family trios revealed by direct haplotype tagging**
Marek Kucka, Andreea Dréau, Christopher M. Cunniff, Yingguang Frank Chan.
Presenter affiliation: Friedrich Miescher Laboratory of the Max Planck Society, Tuebingen, Germany. 49

**Penetrance of Hirschsprung disease depends on cell fate decisions in the enteric nervous system (ENS)**
Sumantra Chatterjee, Elizabeth Vincent, Dallas Auer, Loyal Goff, Aravinda Chakravarti.
Presenter affiliation: NYU School of Medicine, New York, New York. 50

**EpiML—A web server for epistasis analysis using machine learning methods**
Junjie Chen, Jia Wen, Xinghua Shi.
Presenter affiliation: University of North Carolina at Charlotte, Charlotte, North Carolina. 51

**Identification of structural variation associated with cardiometabolic traits in the Finnish population**
Lei Chen, Liron Ganel, Haley J. Abel, David E. Larson, Allison Regier, Indraniel Das, Susan K. Service, FinMetSeq Consortium, Aarno Palotie, Samuli Ripatti, Markku Laakso, Nelson B. Freimer, Michael Boehnke, Adam E. Locke, Nathan O. Stitziel, Ira M. Hall.
Presenter affiliation: Washington University School of Medicine, St Louis, Missouri. 52

**Re-examining the evidence of polygenic adaptation at height-associated loci in mainland Europe and Sardinia**
Minhui Chen, Masato Akiyama, Carlo Sidore, Kazuyoshi Ishigaki, Yukinori Okada, Yoichiro Kamatani, David Schlessinger, Charleston Chiang.
Presenter affiliation: University of Southern California, Los Angeles, California. 53

**Robust and window-insensitive mixture model approaches for localizing balancing selection**
Xiaoheng Cheng, Michael DeGiorgio.
Presenter affiliation: Penn State University, University Park, Pennsylvania. 54

**The impact of global and local Polynesian genetic ancestry on complex traits in Native Hawaiians**
Hanxiao Sun, Meng Lin, Lynne Wilkens, Loic Le Marchand, Christopher A. Haiman, Charleston W. Chiang.
Presenter affiliation: University of Southern California, Los Angeles, California. 55

**The adaptive history of African rainforest hunter-gatherers using whole-exome sequencing**
Jeremy Choin, Marie Lopez, Martin Silvert, Martin Sikora, Helene Quach, Christine Harmant, Patrick Mouguiama-Daouda, Alain Froment, George H. Perry, Luis B. Barreiro, Paul Verdu, Etienne Patin, Lluis Quintana-Murci.
Presenter affiliation: Institut Pasteur, Paris, France. 56

**A reference database and detection method for actively translating ORFs from Ribo-seq data**
Saket Choudhary, Wenzheng Li, Andrew D. Smith.
Presenter affiliation: University of Southern California, Los Angeles, California. 57

**Characterizing genetic and transcript variation using long single molecule sequencing in maize**
Kapeel M. Chougule, Sharon Wei, Yinping Jiao, Bo Wang, Josh Stein, Jianing Liu, Brett Hannigan, Chai Fungtammasan, Victor Llaca, Kevin Fengler, Candice Hirsch, Matthew Hufford, Kelly Dawe, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 58

**Propagating massive local relatedness matrices locus-by-locus—Probabilistic haplotype screening in large datasets via doubly decoupled HMMs**
Ryan Christ, Louis Aslett, David Steinsaltz, Chris Holmes, Chris Spencer, Ira Hall.
Presenter affiliation: Washington University School of Medicine in St. Louis, St. Louis, Missouri; Alan Turing Institute, London, United Kingdom. 59

**Outlying levels of DNA methylation are associated with rare genetic variation and altered gene-expression**
V. Kartik Chundru, James G. Prendergast, Riccardo E. Marioni, Allan J. Beveridge, Grant W. Montgomery, Ian J. Deary, David A. Hume, Naomi R. Wray, Peter M. Visscher, Allan F. McRae.
Presenter affiliation: University of Queensland, Brisbane, Australia. 60

**Predicting the structure and butyrate production of synthetic human gut microbiome communities**
Ryan L. Clark, David M. Stevenson, Daniel Amador-Noguez, Ophelia S. Venturelli.
Presenter affiliation: University of Wisconsin - Madison, Madison, Wisconsin. 61

**mtDNA eQTLs and the $m^1$A 16S rRNA modification explain mtDNA tissue-specific gene expression pattern in humans**
Tal Cohen, Chen Mordechai, Alal Eran, Dan Mishmar.
Presenter affiliation: Ben-Gurion University of the Negev, Be'er Sheva, Israel. 62

***recount-brain*—A curated repository of human brain RNA-seq datasets metadata**
Ashkaun Razmara, Shannon E. Ellis, Dustin J. Sokolowski, Sean Davis, Michael D. Wilson, Jeffrey T. Leek, Andrew E. Jaffe, Leonardo Collado-Torres.
Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland. 63

**Mutational and selective principles of structural variation across human populations**
Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Amit V. Khera, Laurent C. Francioli, Laura D. Gauthier, Harold Wang, Matthew Solomonson, Matthew R. Stone, The gnomAD Consortium, Benjamin M. Neale, Sekar Kathiresan, Mark J. Daly, Eric Banks, Daniel G. MacArthur, Michael E. Talkowski.
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts. 64

**The regulatory landscape underlying epithelial regeneration in *Drosophila***
Elena Vizcaya-Molina, Cecilia C. Klein, Carlos Camilleri-Robles, Raziel Amador-Rios, Florenci Serras, Roderic Guigo, Montserrat Corominas.
Presenter affiliation: Universitat de Barcelona, Barcelona, Spain. 65

**GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child and adult BMI**
Alexessander Couto Alves, Vincent Jaddoe, Struan Grant, Sylvain Sebert, Mark I. McCarthy, Marjo-Riitta Järvelin.
Presenter affiliation: Imperial College London, London, United Kingdom; Surrey University, Guildford, United Kingdom. 66

***In-silico* edgetic profiling of population-specific variation in human genome**
Hongzhu Cui, Anastasia Leshchyk, Suhas Srinivasan, Dmitry Korkin.
Presenter affiliation: Worcester Polytechnic Institute, Worcester, Massachusetts. 67

**Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression**
Anna Cuomo, Daniel Seaton, Davis J. McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Ludovic Vallier, John C. Marioni, Oliver Stegle, Mariya Chhatriwala.
Presenter affiliation: EMBL-EBI, Cambridge, United Kingdom. 68

**Genome-wide mapping of DNA methylation variants affecting gene expression levels in gliomas with respect to their grade and IDH gene mutation status**
Michal J. Dabrowski, Agata Dziedzic, Rafal Guzik, Michal Draminski, Bartosz Wojtas, Karolina Stepniak, Bartlomiej Gielniewski, Jacek Koronacki, Bozena Kaminska.
Presenter affiliation: Institute of Computer Science, Warsaw, Poland. 69

**Functional annotation of lncRNAs in the human brain**
Anupriya Dalmia, Natalia Savytska, Salvador Rodriguez-Nieto, Tenzin Nyima, Noemia-Rita Alves-Fernandes, Stefan Bonn, Melissa G. Castillo Lizardo, Peter Heutink.
Presenter affiliation: German Centre for Neurodegenerative Diseases, Tuebingen, Germany; RIKEN, Yokohama, Japan. 70

**Identifying and functionally characterizing the proximate effects of fetal-specific cardiac regulatory variants on molecular phenotypes**
Matteo D'Antonio, Margaret K. Donovan, William W. Young Greenwald, David Jakubosky, Paola Benaglio, Erin N. Smith, Agnieszka D'Antonio-Chronowska, Kelly A. Frazer.
Presenter affiliation: University of California, San Diego, La Jolla, California. 71

**How specific is tumor-specific splicing?**
Julianne K. David, Sean K. Maden, Abhinav Nellore, Reid F. Thompson.
Presenter affiliation: Oregon Health & Science University, Portland, Oregon. 72

**Analyzing genetic diversity in marmoset colonies**
Ricardo C. del Rosario, Qiangge Zhang, Curtis J. Melo, Xian Gao, Monika A. Burns, Fenna Krienen, Guoping Feng, Steven A. McCarroll.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Harvard Medical School, Boston, Massachusetts. 73

**A pan-cancer transcriptome analysis reveals pervasive regulation through tumor-associated alternative promoters**
Deniz Demircioglu, Martin Kindermans, Engin Cukuroglu, Tannistha Nandi, Claudia Calabrese, Nuno A. Fonseca, Andre Kahles, Kjong Lehmann, Oliver Stegle, Alvis Brazma, Angela N. Brooks, Gunnar Rätsch, Patrick Tan, Jonathan Göke.
Presenter affiliation: Genome Institute of Singapore, Singapore. 74

**The interplay between 3D genome organization and genetic variation**
Olivier Delaneau, Mariana Zazhytska, Christelle Borel, Guillaume Rey, Giuliana Giannuzzi, Stylianos Antonarakis, Alexandre Reymond, Emmanouil Dermitzakis.
Presenter affiliation: University of Geneva, Geneva, Switzerland. 75

**STARsolo—Single-cell RNA-seq analysis beyond gene expression**
Daniel Blibaum, Alexander Dobin.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 76

**The chromatin accessibility information landscape of human transcription factor-DNA interactions**
Ricardo D'Oliveira Albanus, Yasuhiro Kyono, John Hensley, Jacob Kitzman, Stephen C.J. Parker.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 77

**Characterizing the genomic determinants of cis-regulatory element evolution with phylogenetic models**
Noah E. Dukler, Yi-Fei Huang, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 78

**Using citizen science to identify cattle for a changing climate**
Harly J. Durbin, Troy N. Rowan, Helen Yampara-Iquise, Robert D. Schnabel, Jared E. Decker.
Presenter affiliation: University of Missouri, Columbia, Missouri. 79

**Generating a list of prioritized genes associated with a patient's disease and phenotype using the genepanel.iobio web tool**
Aditya Ekawade, Tonya Di Sera, Chase Miller, Alistair Ward, Matt Velinder, Gabor Marth.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 80

**A global survey of transcription in 18,000 samples identifies putative functions of highly repetitive, cancer-associated RNAs originating from endogenous retroviral elements**
Christine Eng, Shumei Chia, Engin Cukuroglu, Lam-Ha Ly, Iwona Szczerbinska, Winston Chan, Ramanuj DasGupta, Jonathan Goeke.
Presenter affiliation: Genome Institute of Singapore, Singapore. 81

**Developing a comprehensive view of the *PRDM1* negative and positive regulatory landscape**
Melissa Englund, Torrin McDonald, Sierra Nishizaki, Jessica Switzenberg, Alan P. Boyle.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 82

**Extending the resources created by the 1000 Genomes Project in the International Genome Sample Resource (IGSR)**
Susan Fairley, Ernesto Lowy, Gabriel Aldam, Paul Flicek.
Presenter affiliation: European Molecular Biology Laboratory, Hinxton, United Kingdom. 83

**An atlas of human benign variation inferred from primate population sequencing**
Kyle Kai-How Farh, Hong Gao, Laksshman Sundaram, Serafim Batzoglou, Lukas F. Kuderna, Marta Gut, Ivo Gut, Felipe Ennes, Takashi Hayakawa, Amanda D. Melin, Katerina Guschanski, Jean P. Boubli, Christian Roos, Jeffrey Rogers, Tomas Marques-Bonet.
Presenter affiliation: Illumina, San Diego, California. 84

**The unbiased length spectrum of human de novo mutations in 4,330 children**
Andrew Farrell, William Richards, Anna Docherty, Hilary Coon, Gabor T. Marth.
Presenter affiliation: University of Utah, Salt Lake Cityh, Utah. 85

**Genetic and environmental determinants of cardiomyocyte gene expression**
A S. Findley, A L. Richards, K Rhodes, M C. Ward, C A. Kalita, A Alazizi, X Wen, D E. Lanfear, R Pique-Regi, Y Gilad, F Luca.
Presenter affiliation: Wayne State University, Detroit, Michigan. 86

**Deciphering the human genetic panorama of Basques**
André Flores-Bello, Frédéric Bauduer, Jasone Salaberria, Bernard Oyharçabak, Lluis Quintana-Murci, David Comas.
Presenter affiliation: Institute of Evolutionary Biology (CSIC – UPF), barcelona, Spain. 87

**Human enhancers with complex evolutionary architectures are more active than enhancers with simple architectures**
Sarah L. Fong, John A. Capra.
Presenter affiliation: Vanderbilt University, Nashville, Tennessee. 88

**Refining reference protein-coding gene annotation for human and mouse in GENCODE**
Adam Frankish, Paul Flicek, The GENCODE Consortium.
Presenter affiliation: European Molecular Biology Laboratory, Cambridge, United Kingdom. 89

**Transcriptomics data goldmine on the ENCODE portal**
Idan Gabdank, Jason Hilton, Yunhai Luo, Jennifer Zamanian, Khine Z. Lin, Jennifer P. Jou, Zachary A. Myers, Emma O'Neill, Paul Sud, Forrest Tanaka, Jin W. Lee, Casey R. Litton, Keenan Graham, Ulugbek K. Baymuradov, Philip Adenekan, Benjamin C. Hitz, J Seth Strattan, J Michael Cherry.
Presenter affiliation: Stanford University School of Medicine, Stanford, California. 90

**SANDY—A straightforward and complete next-generation sequencing read simulator**
Thiago Miller, Gabriela Guardia, Helena Conceicao, Felipe Santos, Rodrigo Barreiro, Jose L. Buzzo, Fernanda O. Rego, Pedro A. Galante.
Presenter affiliation: Hospital Sirio-Libanes, Sao Paulo, Brazil. 91

**Population differentiation of the *ADCY9* pharmacogene reveals a potential epistatic interaction with *CETP***
Isabel Gamache, Magdalena Burchert, Rocio Sanchez, Éric Rhéaume, Jean-Christophe Grenier, Jean-Claude Tardif, Marie-Pierre Dubé, Julie Hussin.
Presenter affiliation: Montreal Heart Institute, Montréal, Canada; Université de Montréal, Montréal, Canada. 92

**Effect of large copy number variants on general cognition in 18,329 healthy individuals**
Eugene J. Gardner, Mari Niemi, George Kirov, Adam S. Butterworth, Hilary C. Martin, Matthew E. Hurles.
Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom. 93

**Molecular profiling of primary cutaneous melanomas reveals pathways predisposing to brain metastases**
Manik Garg, Roy Rabbie, Nuno A. Fonseca, Dominique L. Couturier, Alvis Brazma, David J. Adams.
Presenter affiliation: European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Cambridge, United Kingdom. 94

**Cross-species analysis across 450 million years of evolution reveals conservation and divergence of the microglia program**
Laufey Geirsdottir, Eyal David, Hadas Keren-Shaul, Daniel Erny, Marco Prinz, Ido Amit.
Presenter affiliation: Weizmann Institute of Science, Rehovot, Israel. 95

**On the importance of long-range phasing for non-invasive prenatal testing at low fetal fractions**
Giulio Genovese, Po-Ru Loh, Robert E. Handsaker, Seva Kashin, Steven A. McCarroll.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Harvard Medical School, Boston, Massachusetts. 96

***Oncogene.iobio*—A web app for real-time, integrative examination and functional prioritization of tumor mutations**
Stephanie J. Georges, Yi Qiao, Xiaomeng Huang, Gabor T. Marth.
Presenter affiliation: University of Utah, USTAR Center for Genetic Discovery, Salt Lake City, Utah. 97

**Comprehensive functional genomic resource and integrative model for the human brain**
Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Daniel Geschwind, James Knowles, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 98

**Biological insights from the integrated genomic analysis of 143 human embryonic stem cell lines**
Sulagna Ghosh, Florian Merkle, Giulio Genovese, Robert Handsaker, Seva Kashin, Konrad Karczewski, Steven McCarroll, Kevin Eggan.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Harvard University, Cambridge, Massachusetts. 99

**Open targets genetics—Integrating genetics and functional genomics to identify and prioritize drug targets**
Maya Ghoussaini, on behalf of Open Targets.
Presenter affiliation: Open Targets, Hinxton, United Kingdom; Wellcome Sanger Institute, Hinxton, United Kingdom. 100

**Signatures of transcriptomic individuality revealed by the nine-banded armadillo**
Sara Ballouz, Megan Crow, Maria Pena, Jesse Gillis.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 101

***Cis*-regulation provides a link between the human Y chromosome and sex differences in non-reproductive tissues**
Alexander K. Godfrey, Sahin Naqvi, Lukas Chmatal, Joel M. Chick, Richard N. Mitchell, Steven P. Gygi, David C. Page.
Presenter affiliation: Whitehead Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts. 102

**Single-cell DNA sequencing reveals extensive clonal heterogeneity and biased T-cell infiltration in dedifferentiated liposarcoma**
Rodrigo Gularte Mérida, Timour Baslan, Richard W Cass, Yawei Shen, Arnaud Da Cruz Paula, Christina Antonescu, Nicholas Socci, Tomoyo Okada, Samuel Singer.
Presenter affiliation: Memorial Sloan-Kettering Cancer Center, New York, New York. 103

**New IDEAS for GWAS—Epigenetic "fingerprints" to systematically identify disease causal cell types for thousands of traits in the GWAS Catalog and UK Biobank**
Cong Guo, Karsten B. Sieber, Aidan MacNamara, Matthew R. Nelson, Yu Zhang.
Presenter affiliation: GlaxoSmithKline, Collegeville, Pennsylvania. 104

**Targeted genotyping of Oxford Nanopore reads empowers clinical applications through high single molecule reference-anchored accuracy**
Jeffrey A. Rosenfeld, Marcus Stoiber, Melissa delaBastide, Tim Massingham, Sara Goodwin, Shridar Ganesan.
Presenter affiliation: Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey. 105

WEDNESDAY, May 8—4:30 PM

**Wine and Cheese Party**

WEDNESDAY, May 8—7:30 PM

**SESSION 4** CANCER AND MEDICAL GENOMICS

**Chairpersons:** **Sekar Kathiresan,** Massachusetts General Hospital, Boston
**Catherine Wu,** Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts

**Polygenic scoring and risk for common, complex disease**
Sekar Kathiresan, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts.

**From genetic risk variants to convergent protein networks—An integrative approach to elucidate the causal molecular mechanisms of schizophrenia**
Kasper Lage, Eugene Nacu, April Kim, Edyta Malolepsza, Natalie Petrossian, William Crotty, Tarjinder Singh, Ruize Liu, Greta Pintacuda, Matthew Johnson, Kamal Sharma, SCHEMA Consortium, Richard Huganir, Beth Stevens, Mark Daly, Hailiang Huang, Jake Jaffe, Kevin Eggan.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts. 106

**Putative non-coding cis-regulatory drivers in chronic lymphocytic leukaemia and skin cancer**
Halit Ongen, Olivier Delaneau, Nikolaos M. Lykoskoufis, Aline Real, Christelle Borel, Michael W. Stevens, Cedric Howald, Emmanouil T. Dermitzakis.
Presenter affiliation: University of Geneva, Geneva, Switzerland. 107

**Identifying disorder-specific methylation signatures in patients with severe developmental disorders**
Juliet E. Handsaker, Eugene J. Gardner, Patrick J. Short, Guiseppe Gallone, Leopold Parts, Jeffrey C. Barrett, Matthew E. Hurles.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 108

**Building better personal cancer vaccines**
Catherine J. Wu.
Presenter affiliation: Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts. 109

**A novel method for joint analysis and colocalization applied to 10,000 GWAS datasets identifies molecular mechanisms underpinning GWAS hits, reveals surprising shared biology between diseases, and increases the power of polygenic risk scores across populations**
Peter Donnelly, Gerton Lunter, Eva Krapohl, Andreas Heger, Reka Nagy, Peter Sorensen, Mike Sivley, Mike Weale, Jeffrey Barrett, Gilean McVean, Vincent Plagnol, Chris Spencer.
Presenter affiliation: Genomics plc, Oxford, United Kingdom. 110

**Surveying structural variation landscapes in single cells by strand-specific sequencing reveals complex genomic rearrangement processes**
Ashley D. Sanders, Sascha Meiers, Maryam Ghareghani, David Porubsky, Tobias Marschall, Jan O. Korbel.
Presenter affiliation: European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. 111

**TREX1 and APOBEC3B induce chromothripsis and kataegis**
Aikaterini Chatzipli, John Maciejowski, Peter J. Campbell, Titia de Lange.
Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York. 112

THURSDAY, May 9—9:00 AM

**SESSION 5** POPULATION GENOMICS

**Chairpersons:** **Hilary Martin,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
**Steven McCarroll,** Broad Institute of MIT and Harvard, Cambridge, Massachusetts

**Quantifying the contribution of X-linked coding variants to developmental disorders**
Hilary C. Martin, Ana Lisa Taylor Tavares, Kaitlin E. Samocha, Joanna Kaplanis, Nadia Akawi, Alejandro Sifrim, David R. FitzPatrick, Helen V. Firth, Matthew E. Hurles.
Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom. 113

**Mapping the functions of human disease risk alleles during neuronal development using multiplexed differentiation of human induced pluripotent stem cells**
Julie Jerber, Daniel Seaton, Daniel Pearce, Oliver Stegle, Florian Merkle, Daniel Gaffney.
Presenter affiliation: Open Targets, Hinxton, United Kingdom; Wellcome Sanger Institute, Hinxton, United Kingdom. 114

**Generation of the first regulatory map of human primary microglia from 141 neurosurgical patients.**
Fiona Calvert, Adam Young, Natsuhiko Kumasaka, Robin Franklin, Daniel Gaffney.
Presenter affiliation: Wellcome Sanger Institute, Cambridgeshire, United Kingdom. 115

**Tracing the peopling history of the Southwest Pacific through whole genomes**
Javier Mendoza-Revilla, Jeremy Choin, Lara R. Arauna, Christine Harmant, Maximilian Larena, Olivier Cassar, Jean F. Deleuze, Antoine Gessain, Mattias Jakobsson, Mark Stoneking, Etienne Patin, Lluis Quintana-Murci.
Presenter affiliation: Institut Pasteur, Paris, France. 116

**Monogenic and polygenic inherited forms of clonal selection**
Po-Ru Loh, Giulio Genovese, Steven A. McCarroll.
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Harvard Medical School, Boston, Massachusetts. 117

**Neanderthal fragments lengths and mutational signatures show different historical generation times in non-African population over the past 50,000 years**
Moises Coll, Laurits Skov, Mikkel H. Schierup.
Presenter affiliation: Aarhus University, Aarhus, Denmark. 118

**Tissue variability of regulatory variant effects and the role of transcription factors**
Elise D. Flynn, Stephane Castel, Xiaoting Li, Hanrui Zhang, GTEx Consortium, Raag Agrawal, Harmen J. Bussemaker, Pejman Mohammadi, Tuuli Lappalainen.
Presenter affiliation: Columbia University, New York, New York; New York Genome Center, New York, New York. 119

**Detecting signatures of selection and local adaptation in United States *Bos taurus* beef cattle**
Troy N. Rowan, Robert D. Schnabel, Jared E. Decker.
Presenter affiliation: University of Missouri, Columbia, Missouri. 120

**SESSION 6** POSTER SESSION II

**Population-scale analysis of copy number variation from whole genome sequencing**
Robert E. Handsaker, Seva Kashin, Christina Usher, Christopher Whelan, Michele T. Pato, Carlos N. Pato, Michael Boehnke, Steven A. McCarroll.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Harvard Medical School, Boston, Massachusetts. 121

**HARLEE (Hadoop ARchitecture LakE of Exomes) facilitates the discovery of candidate Mendelian disease genes through a genocentric approach**
Adam W. Hansen, Mullai Murugan, He Li, Michael Khayat, Liwen Wang, Jill Rosenfeld, Shalini Jhangiani, Zeynep Coban Akdemir, Fritz Sedlazeck, Donna Muzny, Aniko Sabo, Jennifer E. Posey, Michael F. Wangler, Christine Eng, Vernon R. Sutton, James R. Lupski, Eric Boerwinkle, Richard A. Gibbs.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 122

**A cryptocurrency funded, blockchain notarized and reviewed, publicly available and usable reference genome of *Cannabis sativa L***
Timothy T. Harkins, Brittany Sexton, Yvonne Helbert, Liam T. Kane, Heather Ebling, Lei Zhang, Biao Liu, Zachary Eaton, Luo Sun, Eileen Dimalanta, Sarah Kingan, Primo Baybayan, Maximilian Press, William Barbazuk, Keerthana Krishnan, Kevin J. McKernan.
Presenter affiliation: Medicinal Genomics, Beverly, Massachusetts. 123

**Associating germline variants with somatic chromosome breaks in cancer**
Meng Xiao He, Sahar Shahamatdar, Amaro Taylor-Weiner, Travis Zack, Sohini Ramachandran, Eliezer Van Allen.
Presenter affiliation: Harvard University, Boston, Massachusetts; Dana-Farber Cancer Institute, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge,. 124

**Understanding regulation of mRNA N6-methyladenosine (m6A) and its contributions to human diseases by QTL mapping**
Zijie Zhang, Kaixuan Luo, Maguanyun Qiu, Matthew Stephens, Xin He, Chuan He.
Presenter affiliation: University of Chicago, Chicago, Illinois. 125

**Mechanisms of tissue specific genetic regulation revealed by latent factors across eQTLs**
Yuan He, Kaushik Srinivasan, GTEx Consortium, Christopher D. Brown, Alexis Battle.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 126

**Application of variation graph for the detection of additional somatic mutations**
Cecilia Wieder, Garima Khandelwal, Javier Herrero.
Presenter affiliation: UCL Cancer Institute, University College London, London, United Kingdom. 127

**vlinc273 is required for normal DNA replication timing of chromosome 6 in the GM12878 cell line**
Michael B. Heskett, Leslie Smith, Takayo Sasaki, David M. Gilbert, Matthew J. Thayer, Paul T. Spellman.
Presenter affiliation: Oregon Health and Science University, Portland, Oregon. 128

**The foundational data initiative for Parkinson disease (FOUNDIN-PD)**
Peter Heutink, David W. Graig, Kendall Van Keuren-Jensen, Steve Finkbeiner, Mark Cookson, Mike Nalls, Raphael J. Gibbs, Andrew Singleton.
Presenter affiliation: German Center for Neurodegenerative Diseases, Tuebingen, Germany. 129

**Functional annotation of non-coding genomic regions in non-human primates**
Naoki Hirose, Masatsugu Ema, Erika Sasaki, Jun Kawai, Yasuhiro Murakawa, Hideya Kawaji.
Presenter affiliation: RIKEN, Yokohama, Japan. 130

**Genome-wide germline correlates of the epigenetic landscape of prostate cancer**
Kathleen E. Houlahan, Yu-Jia Shiah, Musaddeque Ahmed, Michael Fraser, Housheng H. He, Robert G. Bristow, Paul C. Boutros.
Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada; University of Toronto, Toronto, Canada; Vector Institute, Toronto, Canada. 131

**Multi-omics data analysis in cardiovascular research**
Julie Hussin.
Presenter affiliation: Université de Montréal, Montreal, Canada. 132

**Genomics2020—An overview of the ongoing NHGRI strategic planning process**
Carolyn M. Hutter, on behalf of the NHGRI Strategic Planning Group.
Presenter affiliation: National Human Genome Research Institute, Bethesda, Maryland. 133

**A human protein-protein interaction network for PD GWAS risk loci to catalyze genomic interpretation**
Anastasia Illarionova, Anamika Giri, Joachim Taeger, Peter Heutink.
Presenter affiliation: Deutsches Zentrum fur Neurodegenerative Erkrankungen (DZNE), Tuebingen, Germany. 134

**Comparison of the gene expression profile in the different cultured condition by single cell transcriptome analysis**
Sadahiro Iwabuchi, Saki Kurokawa, Yu Kamide, Shinichi Hashimoto.
Presenter affiliation: Kanazawa University, Kanazawa, Japan. 135

**Understanding genetic variation using targeted and single-cell long-read sequencing**
Shruti Iyer, Sara Goodwin, W. Richard McCombie.
Presenter affiliation: Cold Spring Harbor Laboratory, Woodbury, New York; Stony Brook University, Stony Brook, New York. 136

**CNet—A multi-omics approach to detecting clinically associated, combinatory genomic signatures**
Peilin Jia, Guangsheng Pei, Zhongming Zhao.
Presenter affiliation: The University of Texas Health Science Center at Houston, Houston, Texas. 137

**The ENCODE transcriptome pipelines—Reproducible, portable tools for RNA-seq analysis**
Otto Jolanki, Ulugbek K. Baymuradov, Idan Gabdank, Keenan Graham, Ben C. Hitz, Jin W. Lee, Sorena Rahmanian, Vanessa Sochat, J Seth Strattan, Dana Wyman, Ali Mortazavi, J Michael Cherry, ENCODE Data Coordination Center.
Presenter affiliation: Stanford University, Stanford, California. 138

**Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci**
Thomas M. Keane, Mouse Genomes Project Consortium.
Presenter affiliation: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom. 139

**Identification of distinct molecular subtypes of Crohn's disease through gene expression and chromatin accessibility**
Benjamin P. Keith, Takahiko Toyonaga, Michelle Hoffner O'Connor, Praveen Sethupathy, Shehzad Z. Sheikh, Terrence S. Furey.
Presenter affiliation: The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina. 140

**Single-cell dissection of Alzheimer's pathology—Genetics, epigenomics, transcription, sex, age, and disease phenotypes**
Manolis Kellis, Hansruedi Mathys, Jose Davila-Velderrain, Liang He, Shahin Mohammadi, Carles Boix, Maria Kousi, David A. Bennett, Li-Huei Tsai.
Presenter affiliation: MIT, Cambridge, Massachusetts; Broad Institute, Cambridge, Massachusetts. 141

**Gene-environment interactions in the UK Biobank**
Matthew Kerin, Kevin Sharp, David Steinsaltz, Jonathan Marchini.
Presenter affiliation: University of Oxford, Oxford, United Kingdom. 142

**Telomere-to-telomere assembly of a complete human X chromosome**
Karen H. Miga, Sergey Koren, Arang Rhie, Shelise Brooks, Gerard G. Bouffard, Valerie Maduro, Nancy Hansen, Edmund Howe, Jennifer L. Gerton, Jo Wood, Milinn G. Kremitzki, Matthew Loose, Urvashi Surti, Kerstin Howe, Tina G. Lindsay, Alice Young, James Mullikin, Adam M. Phillippy.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 143

**μSPLiT—High-throughput single-cell RNA sequencing platform for microbial communities**
Anna Kuchina, Leandra M. Brettner, William R. DePaolo, Georg Seelig.
Presenter affiliation: University of Washington, Seattle, Washington. 144

**The Primate Genome Sequencing Initiative—Assessing primate genomic diversity at a global scale**
Lukas F.K. Kuderna, Marta Gut, Ivo Gut, Felipe Ennes, Takashi Hayakawa, Amanda D. Melin, Katerina Guschanski, Jean P. Boubli, Christian Roos, Jeffrey Rogers, Kyle Kai-How Farh, Tomas Marques-Bonet.
Presenter affiliation: Institut de Biologia Evolutiva, Barcelona, Spain. 145

**Predictive modeling in biological systems using multi-omics datasets in open-science, community platform of KBase**
Vivek Kumar, Sunita Kumari, Zhenyuan Lu, Doreen Ware, Bob Cottingham, Chris Henry, Adam Arkin.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 146

**Efficient and high-throughput genotyping of elusive human polymorphic inversions mediated by inverted repeats**
Marina Laplana, Sergi Villatoro, Roser Zaurín, José Luis Royo, Marta Puig, Mario Cáceres.
Presenter affiliation: Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain. 147

**Detection of tissue-specific RNA in the plasma of cancer patients—A Circulating Cell-free Genome Atlas (CCGA) Substudy**
Matthew H. Larson, Hyunsung J. Kim, Wenying Pan, Sarah Stuart, Yiqi Zhou, Archana Shenoy, Monica Pimentel, Per Knudsgaard, Vasiliki Demas, Earl Hubbell, Alexander M. Aravanis, Arash Jamshidi.
Presenter affiliation: GRAIL, Inc., Menlo Park, California. 148

**Mining thousands of genomes for reliable structural variant population allele frequency estimates**
Ryan M. Layer, Brent S. Pedersen, Aaron R. Quinlan, Fritz J. Sedlazeck, Eric Boerwinkle.
Presenter affiliation: University of Colorado, Boulder, Colorado. 149

**Contiguous assembly of Bionano DLS chemistry reveals structural variants at challenging regions**
Joyce Lee, Andy Wing Chun Pang, Thomas Anantharaman, Ernest T. Lam, Alex Hastie, Mark Borodkin.
Presenter affiliation: Bionano Genomics, San Diego, California. 150

**The influence of human inversions on gene expression, epigenetic changes and phenotypic variation**
Jon Lerga-Jaso, Marta Puig, Alejandra Delprat, Sergi Villatoro, Marina Laplana, Alba Vilella-Figuerola, Teresa Soos, Claudia Ramírez, Clara Vizuete, Olivier Delaneau, Tõnu Esko, Emmanouil T. Dermitzakis, Mario Cáceres.
Presenter affiliation: Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain. 151

**Evolutionary divergence of mammalian spermatogenesis—Single-cell comparative analyses across mouse, monkey, and human testes**
Xianing Zheng, Adrienne Shami, Qianyi Ma, Hailey Larose, Gabriel Manske, Sue S. Hammoud, Jun Z. Li.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 152

**Delving into chromatin on the ENCODE portal**
Khine Z. Lin, Idan Gabdank, Jason Hilton, Yunhai Luo, Jennifer Zamanian, Philip Adenekan, Ulugbek K. Baymuradov, Keenan Graham, Jennifer P. Jou, Jin W. Lee, Casey R. Litton, Zachary A. Myers, Emma O'Neill, Paul Sud, Forrest Tanaka, J. Seth Strattan, Benjamin C. Hitz, J. Michael Cherry.
Presenter affiliation: Stanford University School of Medicine, Palo Alto, California. 153

**Developing *Drosophila melanogaster* as a model for the genetic basis of early-stage reproductive isolation**
Matthew J. Lollar, John E. Pool.
Presenter affiliation: University of Wisconsin-Madison, Madison, Wisconsin. 154

**Less-is-more and more-is-less in integrating multi-scale 'omics with polygenic phenotype predictors**
Shengjie Lu, Chen Cao, Zhi Xiong, Xuewen Lu, Quan Long.
Presenter affiliation: University of Calgary, Calgary, Canada. 155

**SciApps.org—A cloud-based platform for reproducible bioinformatics workflows**
Zhenyuan Lu, Liya Wang, Peter Van Buren, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 156

**Putative non-coding regulatory driver discovery using unmatched tumor-normal samples**
N M R Lykoskoufis, H Ongen, E T. Dermitzakis.
Presenter affiliation: University of Geneva Medical School, Geneva, Switzerland. 157

**Somatic mutational profiling reveals the cellular history of normal human lymphocytes**
Heather Machado, Emily Mitchell, Nina Øbro, Henry Lee-Six, Megan Davies, Kourosh Saeb-Parsy, Dan Hodson, Anthony Green, David Kent, Peter Campbell.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 158

**Profiling immunoglobulin repertoires by RNA Sequencing across 8555 samples from 53 GTEx tissues**
Serghei Mangul, Noah Zaitlen, Eleazar Eskin.
Presenter affiliation: UCLA, Los Angeles, California. 159

**SigNet—Identifying mutational processes in cancer using neural networks**
Lasse Maretty, Søren Besenbacher.
Presenter affiliation: Aarhus University Hospital, Aarhus, Denmark. 160

**A custom bioinformatics pipeline for analyzing Strand-seq data**
Carl-Adam Mattsson, Victor Gurjev, Peter M. Lansdorp.
Presenter affiliation: Terry Fox Laboratory, Vancouver, BC, Canada. 161

**Developing a mouse model for identifying neoantigen-specific T cells for adoptive cell transfer**
Laura H. McCulloch, David M. Truong, Justin Mastroianni, Michelle Krogsgaard, Jef D. Boeke.
Presenter affiliation: NYU Langone Health, New York, New York. 162

**Isolating protein-occupied open chromatin regions**
Torrin McDonald, Rebecca Hurto, Peter Freddolino, Alan Boyle.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 163

**Identifying novel regulatory elements using RELICS, a statistical framework for the analysis of CRISPR regulatory screens**
Patrick C. Fiaux, Hsiuyi Chen, Ishika Luthra, Carolyn O'Connor, Graham McVicker.
Presenter affiliation: Salk Institute for Biological Studies, La Jolla, California. 164

**Population genomics of copy number variation in a natural population of maize wild relative teosinte**
Wenbin Mei, Simon Renny-Byfield, Anne Lorant, Arun S. Seetharam, John Doebley, Matthew B. Hufford, Jeffrey Ross-Ibarra.
Presenter affiliation: University of California, Davis, California. 165

**Decoding the transcriptional regulatory atlas of the maize leaf**
Maria K. Mejia-Guerra, Xiaoyu Tu, Jose A. Valdes Franco, Pinghua Li, Edward S. Buckler, Silin Zhong.
Presenter affiliation: Cornell University, Ithaca, New York. 166

**Multiple waves of peopling northeast Europe—An ancient DNA study through Mesolithic to the Middle Ages**
Lehti Saag, Kristiina Tambets, Mait Metspalu.
Presenter affiliation: Institute of Genomics, University of Tartu, Tartu, Estonia. 167

**Genomic analysis of contagious cancers of clams**
Marisa Yonemitsu, Brian Beal, Brian W. Davis, Michael J. Metzger.
Presenter affiliation: Pacific Northwest Research Institute, Seattle, Washington. 168

**Improving clinical assessment of whole genome sequencing data—The CAGI SickKids Challenges**
Stephen Meyn, Jesse Hunter, Rupa Udani.
Presenter affiliation: University of Wisconsin, Madison, Wisconsin; Hospital for Sick Children, Toronto, Canada. 169

**Homotypic cooperativity and collective binding are determinants of bHLH specificity and function**
Christian A. Shively, Jiayue Liu, Xuhua Chen, Kaiser Loell, Robi D. Mitra.
Presenter affiliation: Washington University School of Medicine in St. Louis, St Louis, Missouri. 170

**Impact of alternative polyadenylation on gene and protein expression levels in humans**
Briana E. Mittleman, Shane F. Warland, Kristen Patterson, Zheng Tony, Pott Sebastian, Gilad Yoav, Yang I. Li.
Presenter affiliation: University of Chicago, Chicago, Illinois. 171

**Genetic and environmental influences on protein variability in the human immune system**
Michael D. Morgan, Etienne Patin, Bernd Jagla, Milena Hasan, Massimo Mangino, Tim D. Spector, Lluis Quintana-Murci, John C. Marioni.
Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom. 172

**Recompleting the *Caenorhabditis elegans* genome**
Jun Yoshimura, Kazuki Ichikawa, Massa J. Shoura, Karen L. Artiles, Idan Gabdank, Lamia Wahba, Cheryl L. Smith, Mark L. Edgley, Ann E. Rougvie, Andrew Z. Fire, Shinichi Morishita, Erich M. Schwarz.
Presenter affiliation: The University of Tokyo, Tokyo, Japan. 173

**Prediction of neuronal phenotypes using RNA modifications and genomic dissection of cell compartments in single neurons**
Andrea B. Kohn, Leonid L. Moroz.
Presenter affiliation: University of Florida, St. Augustine, Florida; University of Florida, Gainesville, Florida. 174

**Sequencing of entire brains at the single-cell resolution—Principles of molecular classification and periodic system of neurons**
Leonid L. Moroz.
Presenter affiliation: University of Florida, St. Augustine/Gainesville, Florida. 175

**Connecting chromatin structure to regulatory function—Extending split-pool recognition of interactions by tag extension (SPRITE) to associate structural complexes with transcriptional regulation**
Rachel Murphy, Pranitha Vangala, Patrick McDonel, Sofia A. Quinodoz, Mitchell Guttman, Manuel Garber.
Presenter affiliation: University of Massachusetts Medical School, Worcester, Massachusetts. 176

**Genome-wide chromatin structural analysis reveals inter-chromosomal translocations in B-cell lymphoma cell line**
Luis A E. Nagai, Sung-Joon Park, Kenta Nakai.
Presenter affiliation: University of Tokyo, Kashiwanoha, Chiba, Japan. 177

**LINE-1 activation in cancer and its impact on cancer immunotherapy**
Fabio Navarro, George Gayed, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 178

**Examining the consistency between survey responses and genetic information of AncestryDNA customers**
Luong T. Nguyen, Natalie Telis, Harendra Guturu, Julie M. Granka, Jake Byrnes, Keith Noto, Eurie L. Hong, Ken G. Chahine, Catherine A. Ball.
Presenter affiliation: Ancestry.com DNA, San Francisco, California. 179

**Genomic predictions for selecting locally adapted beef cattle**
Sara M. Nilson, Troy N. Rowan, Robert D. Schnabel, Jared E. Decker.
Presenter affiliation: University of Missouri, Columbia, Missouri. 180

**Identifying subsets of cell-specific regulatory sequence features from heterogeneous tissue-specific chromatin state data using predictive gapped k-mer features**
Jin Woo Oh, Michael A. Beer.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 181

**TBrowse—A modular gene tree viewer for exploring functional conservation**
Andrew J. Olson, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 182

**Benchmarks for 'dark matter' of the human genome**
Nathan D. Olson, Justin Wagner, Jennifer McDaniel, Justin M. Zook, Genome In A Bottle Consortium.
Presenter affiliation: National Institute of Standards and Technology, Gaithersburg, Maryland. 183

**Combined DNase footprinting and motif matching model for identification of transcription factor binding sites**
Ningxin Ouyang, Alan P. Boyle.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 184

**Computational approaches to estimate the kinetics of mRNA biogenesis**
Nida Javeed, Kevin Fortier, Alexander Park, Leslie Torres Ulloa, Athma A. Pai.
Presenter affiliation: University of Massachusetts Medical School, Worcester, Massachusetts. 185

**Systems biology analysis of single-cell transcriptomics identifies hidden drivers in tumor microenvironment for metastasis of breast and liver cancers**
Qingfei Pan, Bensheng Ju, Liyuan Li, Yanyan Wang, Long Nguyen, Hongbo Chi, John Easton, Myriam Labelle, Liqin Zhu, Jiyang Yu.
Presenter affiliation: St. Jude Children's Research Hospital, Memphis, Tennessee. 186

**Macromap—A systematic map of condition-specific genetic effects on expression in immune response based in iPSC-derived macrophages**
Nikolaos I. Panousis, Andrew Knights, Maria Imaz, Alice Barnett, Lorena B. Vilarino, Anthi Tsingene, Celine Gomez, Daniel Gaffney.
Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom. 187

**Host-unmapped NGS reads shape the prevalence of microbial contamination**
Sung-Joon Park, Kenta Nakai.
Presenter affiliation: The University of Tokyo, Tokyo, Japan. 188

**Improved gkm-SVM regulatory element and TFBS detection incorporating multiple tissue/cell samples and data types**
Ayoti Patra, Dustin Shigaki, Michael A. Beer.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 189

**deTS—Tissue-specific enrichment analysis to decode tissue specificity**
Guangsheng Pei, Yulin Dai, Zhongming Zhao, Peilin Jia.
Presenter affiliation: The University of Texas Health Science Center at Houston, Houston, Texas. 190

**The landscape of interactions between germline and somatic variants in cancer**
Bernard Pereira, Catarina D. Campbell, Wendy Winckler.
Presenter affiliation: Novartis Institutes for Biomedical Research, Cambridge, Massachusetts. 191

**From gene co-expression to brain circuits—Prefrontal gene co-expression networks predict drug response in patients with schizophrenia**
Giulio Pergola, Pasquale Di Carlo, Andrew E. Jaffe, Marco Papalino, Qiang Chen, Thomas M. Hyde, Joel E. Kleinman, Joo Heon Shin, Antonio Rampino, Giuseppe Blasi, Alessandro Bertolino, Daniel R. Weinberger.
Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland; University of Bari Aldo Moro, Bari, Italy. 192

**Chromatin accessibility during adipocyte differentiation identifies regulatory elements at cardiometabolic trait GWAS loci**
H J. Perrin, K W. Currin, C K. Raulerson, S Vadlamudi, J P. Davis, M Wabitsch, P Pajukanta, M Laakso, K L. Mohlke.
Presenter affiliation: University of North Carolina, Chapel Hill, North Carolina. 193

**Characterization of activated ion-electron transfer dissociation (AI-ETD) reaction kinetics**
Trenton M. Peters-Clarke, Benton J. Anderson, Jean M. Lodge, Dain R. Brademan, Kevin L. Schauer, Michael S. Westphall, Joshua J. Coon.
Presenter affiliation: University of Wisconsin-Madison, Madison, Wisconsin. 194

**SPDI—Data model for variants and applications at NCBI**
J. Bradley Holmes, Eric Moyer, Lon Phan, Donna Maglott, Brandi Kattman.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 195

**Modeling genetic variants predisposing for schizophrenia in human neurons**
Olli Pietilainen, Ralda Nehme, Andrea Ganna, Mykyta Artomov, Aditi Trehan, Lindy Barrett, Aarno Palotie, Mark Daly, Kevin Eggan.
Presenter affiliation: Harvard University, Cambridge, Massachusetts; The Broad Institute, Cambridge, Massachusetts. 196

**Determining acquisition of pluripotency with single-cell transcriptomics**
Stefan J. Pietrzak, Khoa A. Tran, Nur Zafirah Zaidan, Alireza Siahpirani, Sunnie Grace McCalla, Gopal Iyer, Sushmita Roy, Rupa Sridharan.
Presenter affiliation: University of Wisconsin-Madison, Madison, Wisconsin. 197

**STREAM—Single-cell trajectories reconstruction, exploration and mapping of omics data**
Huidong Chen, Luca Albergante, Jonathan Y. Hsu, Caleb A. Lareau, Giosue` Lo Bosco, Jihong Guan, Shuigeng Zhou, Alexander N. Gorban, Daniel E. Bauer, Martin J. Aryee, David M. Langenau, Andrei Zinovyev, Jason D. Buenrostro, Guo-Cheng Yuan, Luca Pinello.
Presenter affiliation: Massachusetts General Hospital Research Institute and Harvard Medical School, Boston, Massachusetts. 198

**Transcriptional pathways of parturition at single-cell resolution**
R Pique-Regi, R Romero, A L. Tarca, E Sendler, Y Xu, O Erez, F Luca, S Hassan, N Gomez-Lopez.
Presenter affiliation: Wayne State University, Detroit, Michigan. 199

**Understanding methylation variation within natural lines in *Arabidopsis thaliana***
Rahul Pisupati, Magnus Nordborg.
Presenter affiliation: Gregor Mendel Institute, Vienna, Austria; Institute of Population Genetics, Vienna, Austria. 200

**The IOBIO architecture—High-performance genomic data analysis in the web browser**
Anders Pitman, Chase A. Miller, Yi Qiao, Gabor Marth, Tonya DiSera, Alistair Ward.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 201

**Inversion hotspots drive great-ape genome and gene evolution**
David Porubsky, Ashley D. Sanders, Shwetha C. Murali, David Gordon, Stuart Cantsilieris, Arvis Sulovari, Mario Ventura, Jan O. Korbel, Tobias Marschall, Evan E. Eichler.
Presenter affiliation: University of Washington, Seattle, Washington. 202

**Detection of local splicing disruptions in cancer**
Laurie Prelot, Andre Kahles, Natalie Davidson, Gunnar Rätsch.
Presenter affiliation: ETH Zurich, Zurich, Switzerland. 203

**Divergent neuronal DNA methylation patterns across human cortical development—Critical periods and a unique role of CpH methylation**
Amanda J. Price, Leonardo Collado-Torres, Nikolay A. Ivanov, Wei Xia, Emily E. Burke, Jooheon Shin, Ran Tao, Liang Ma, Yankai Jia, Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger, Andrew E. Jaffe.
Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland; Johns Hopkins School of Medicine (JHSOM), Baltimore, Maryland. 204

**Approaching the minimum copy number generation problem in cancer genomics**
Letu Qingge, Conor Nodzak, Xinghua Shi, Binhai Zhu.
Presenter affiliation: University of North Carolina at Charlotte, Charlotte, North Carolina. 205

**Discordant monozygotic twins pairs to complete the catalog of genes mutations which cause diseases**
Christophe Ambroise, Guillaume Charpentier, Jean-François Deleuze, Sylvia Franc, Michele Granier, Marion Hetzel, Paul Jabert, Francis Quétier, Guillem Rigaill, Béatrice Segurens, Marie-Victoire Sénat, Isabelle Sudre, Yves Ville.
Presenter affiliation: University of Evry, Evry, France. 206

**SESSION 7** ELSI PANEL and DISCUSSION

**The Use of Genetic Genealogy for Forensic Investigations**

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

**Panelists**

**Thomas F. Callaghan, Ph.D.,** FBI Laboratory
**James W. Hazel, Ph.D., J.D.,** Vanderbilt University Medical Center
**Sara Huston Katsanis, M.S.,** The Duke Initiative for Science & Society
**Ellen McRae Greytak, Ph.D.,** Parabon NanoLabs

The arrest of a suspect in the Golden State Killer case in April 2018 demonstrated the significant potential of combining genetic with genealogical data in cases lacking any direct matches of known offenders to DNA evidence. Since that time, dozens of suspects have been arrested after police authorized the upload of the SNP genetic profiles of crime scene samples to a public genealogical database and a genetic genealogist traced the family trees of distant relatives of the suspect, leading to a direct DNA match. Similar investigations into cases of unidentified human remains also have been pursued in this way. As law enforcement begins to combine genomic data with public genealogical databases, in what circumstances should such approaches be used? What are the boundaries that define legal use of public DNA databases? Are additional protections for genetic data or limits on law enforcement needed? Are there potential implications for research or health-related DNA data? This panel will explore this topic from a variety of perspectives including an overview of the technology, the laws and policies, and the ethical considerations.

**SESSION 8** EVOLUTIONARY NON-HUMAN GENOMICS

**Chairpersons:** **Kirk Lohmueller,** University of California, Los Angeles
**Joseph Thornton,** University of Chicago, Illinois

**Natural populations of foxes, wolves, and otters reveal the impact of demography on deleterious variation and the genetic basis of inbreeding depression**
Jacqueline Robinson, Annabel Beichman, Caitlin Brown, Jannikke Räikkönen, Robert Wayne, Kirk Lohmueller.
Presenter affiliation: UCLA, Los Angeles, California. 207

**Widespread haploid-biased gene expression in mammalian spermatogenesis associated with frequent selective sweeps and evolutionary conflict**
Kunal Bhutani, Katherine Stansifer, Simina Ticau, Lazar Bojic, Joanna Slisz, Claudia Cremers, Christian Roy, Jerry Donovan, Brian Fiske, Robin Friedman.
Presenter affiliation: Ohana Biosciences, Cambridge, Massachusetts. 208

**Associating molecular evolutionary rates with convergent metabolic traits in over 200 yeast genomes**
Wynn K. Meyer, Xing-Xing Shen, Chris T. Hittinger, Antonis Rokas, Nathan Clark.
Presenter affiliation: University of Pittsburgh, Pittsburgh, Pennsylvania. 209

**Genomic architecture and evolutionary conflict as drivers of expression patterns in a social supergene**
Carlos Martinez-Ruiz, Rodrigo Pracana, Eckart Stolle, Richard A. Nichols, Yannick Wurm.
Presenter affiliation: Queen Mary University London, London, United Kingdom. 210

**Alternative histories through the sequence space of an ancient protein—Evolution, epistasis, and chance**
Joseph Thornton.
Presenter affiliation: University of Chicago, Chicago, Illinois.

**Early origin of developmental enhancers in animals**
Emily S. Wong, Siew Z. Tan, Gaiti Federico, Victoria Garside, Ethan Scott, Edwina Mcglinn, Bernard Degnan, Mathias Francois.
Presenter affiliation: University of Queensland, St Lucia, Australia. 211

**Clusters of divergent enhancers in the adaptive evolution of natural stickleback populations**
Stanley J. Neufeld, Jonas Schwickert, Jukka-Pekka Verta, Muhua Wang, Malina John, Domenico Scionti, Felicity C. Jones.
Presenter affiliation: Max Planck Society, Tuebingen, Germany. 212

**Conserved transcriptional regulatory program brings about tissue differentiation during development**
Marina Ruiz-Romero, Cecilia C. Klein, Sílvia Pérez-Lluch, Alessandra Breschi, Amaya Abad, Emilio Palumbo, Roderic Guigó.
Presenter affiliation: Centre for Genomic Regulation (CRG), Barcelona, Spain. 213

FRIDAY, May 10—9:00 AM

**SESSION 9** GENOME ENGINEERING AND EDITING

**Chairpersons:** **Christoph Bock,** Austrian Academy of Sciences, Vienna
**Farren Isaacs,** Yale University, New Haven, Connecticut

**CRISPR single-cell sequencing—Toward functional biology in high throughput**
Christoph Bock.
Presenter affiliation: Austrian Academy of Sciences, Vienna, Austria; Medical University of Vienna, Vienna, Austria; Max Planck Society, Saarbrücken, Germany. 214

**Pairwise and higher-order genetic interactions during the evolution of a tRNA**
Júlia Domingo, Guillaume Diss, Ben Lehner.
Presenter affiliation: EMBL-CRG Systems Biology Research Unit, Barcelona, Spain; Universitat Pompeu Fabra, Barcelona, Spain. 215

**Enhancer function of ultraconserved elements does not require perfect sequence conservation**
Valentina Snetkova, Quan T. Pham, Momoe Kato, Stella Tran, Ingrid Plajzer-Frick, Catherine S. Pickle, Anne N. Harrington, Jennifer A. Akiyama, Veena Afzal, Diane E. Dickel, Axel Visel, Len A. Pennacchio.
Presenter affiliation: Lawrence Berkeley National Laboratory, Berkeley, California. 216

**Human pancreatic islet 3D chromatin architecture provides insights into the genetics of type 2 diabetes**
Inês Cebola, Irene Miguel-Escalada, Silvia Bonàs-Guarch, Goutham Atla, Julen Mendieta-Esteban, Biola M. Javierre, Irene Farabella, Marc A. Marti-Renom, Peter Fraser, Jorge Ferrer.
Presenter affiliation: Imperial College London, London, United Kingdom. 217

**Precise editing at DNA replication forks enables multiplex genome engineering in eukaryotes**
Farren Isaacs.
Presenter affiliation: Yale University, New Haven, Connecticut. 218

**Mitochondrial evolution among diverse yeast species and synthetic mitochondrial genomics tools for functional characterization**
John F. Wolters, Dana A. Opulente, Jacek Kominek, Abbe L. LaBella, Xing-Xing Shen, Xiaofan Zhou, Clete Kurtzman, Julie Trolle, Sudarshan Pinglay, Jef D. Boeke, Antonis Rokas, Chris T. Hittinger.
Presenter affiliation: UW-Madison, Madison, Wisconsin. 219

**CRISPResso2—Allele-specific quantification and visualization of genome editing events from CRISPR/Cas9, Cas12A (CPF1) and base editors**
Kendell Clement, Holly Rees, Matthew Canver, Jason M. Gehrke, Rick Farouni, Jonathan Y. Hsu, Mitchel A. Cole, David R. Liu, Keith Joung, Daniel E. Bauer, Luca Pinello.
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts. 220

**Big brains—High-throughput enhancer deletions and the genetics of becoming human**
Evan Geller, James P. Noonan.
Presenter affiliation: Yale University, New Haven, Connecticut. 221

FRIDAY, May 10—12:00 PM

**Meet the Speakers Lunch**

**SESSION 10** POSTER SESSION III

**Analysing the joint effect of common and rare variants on human traits, using whole-genome sequencing data and a gene-based approach**
Nima Rafati, Marcin Kierczak, Julia Höglund, Mathias Rask-Andersen, Weronica Ek, Diana Ekman, Torgny Karlsson, Åsa Johansson.
Presenter affiliation: Science for Life Laboratory, Uppsala University, Uppsala, Sweden. 222

**Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism**
Gokul Ramaswami, Daniel H. Geschwind.
Presenter affiliation: UCLA, Los Angeles, California. 223

**Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks**
Helen Zhu, Liis Uuskula-Reimand, Keren Isaev, Lina Wadi, Azad Alizada, Shimin Shuai, Vincent Huang, Dike Aduluso-Nwaobasi, Marta Paczkowska, Diala Abd-Rabbo, Oliver Ocsenas, Minggao Liang, Jared T. Simpson, Mathieu Lupien, Lincoln D. Stein, Paul C. Boutros, Michael D. Wilson, Juri Reimand.
Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada; University of Toronto, Toronto, Canada. 224

**Psychosocial experiences and their genetic interactions are associated with transcriptional dysregulation in asthma**
Justyna A. Resztak, Allison K. Farrell, Henriette E. Mair-Meijers, Adnan Alazizi, Samuele Zilioli, Roger Pique-Regi, Richard B. Slatcher, Francesca Luca.
Presenter affiliation: Wayne State University, Detroit, Michigan. 225

**Developmental roles of frizzled receptors in butterfly embryos**
Teomie S. Rivera-Miranda, Joseph Hanly, Jose J. Hermina-Perez, Christopher R. Day, Riccardo Papa, Arnaud Martin.
Presenter affiliation: University of Puerto Rico Rio Piedras, San Juan, PR, -. 226

**Genetic association analysis of type 1 diabetes in > 60,000 ancestrally diverse subjects refines association signals and identifies novel risk loci**
Catherine C. Robertson, Jamie R. Inshaw, Suna Onengut-Gumuscu, Wei-Min Chen, John A. Todd, Stephen S. Rich.
Presenter affiliation: University of Virginia, Charlottesville, Virginia. 227

**The impact of modern human specific sites on human phenotypes**
Christopher R. Robles, Sriram Sankararaman.
Presenter affiliation: UCLA, Los Angeles, California. 228

**Population-specific variants of interest for carrier screening in diverse US hispanics**
Juan L. Rodriguez-Flores, Juan C. Martinez-Cruzado, Robert A. Paulino-Ramirez, Philip L. Leopold, Ronald G. Crystal.
Presenter affiliation: Weill Cornell Medical College, New York, New York. 229

**Tissue-specific evolution of promoters and enhancers in mammals is associated with transposable elements**
Maša Roller, Ericca Stamper, Diego Villar, Aisling Redmond, Louise Harewood, Duncan T. Odom, Paul Flicek.
Presenter affiliation: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom. 230

**Predisposition towards malignant transformation in the human distal fallopian tube**
Ian Rose, Mallikarjun Bidarimath, Heiko Schultz, Blaine Harlan, Andrew K. Godwin, Andrea Flesken-Nikitin, Alexander Y. Nikitin.
Presenter affiliation: Cornell University, Ithaca, New York. 231

**An empirical P-value method for detecting cancer cell of origin from single cell RNA-seq with an application to lung cancer**
Mahboubeh Rostami, Ronald G. Crystal, Jason G. Mezey.
Presenter affiliation: Weill Cornell Meidcal College, New York, New York. 232

**exceRpt—A comprehensive analytic platform for extracellular RNA profiling**
Joel Rozowsky, Robert Kitchen, Timur Galeev, Jonathan Park, James Diao, Jonathan Warrell, William Thistlethwaite, Oscar Murillo, Sai Subramanian, Matt Roth, Aleksander Milosavljevic, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 233

**FLEA-ChIP enables reproducible immunoprecipitation from 100 cells**
Marina Ruiz-Romero, Cecilia C. Klein, Roderic Guigó, Sílvia Pérez-Lluch.
Presenter affiliation: Centre for Genomic Regulation (CRG), Barcelona, Spain. 234

**Identification of local copy number aberrations in cancer genomes using nanopore sequencer**
Yoshitaka Sakamoto, Ayako Suzuki, Yutaka Suzuki.
Presenter affiliation: The University of Tokyo, Chiba, Japan. 235

**Whole exome sequencing of canine osteosarcoma as a model for the corresponding human disease**
S Sakthikumar, I Elvers, J Kim, M L. Arendt, R Thomas, J Johnson, R Swofford, J Alföldi, E Axelsson, G Couto, M E. Pettersson, J Meadows, J F. Modiano, M Breen, M Kierczak, K Forsberg-Nilsson, V D. Marinescu, K Lindblad-Toh.
Presenter affiliation: Uppsala University, Uppsala, Sweden; Broad Institute, Cambridge, Massachusetts. 236

**Determining growth patterns and driver effects from mutational frequencies in 994 individual tumors**
Leonidas Salichos, William Meyerson, Jonathan Warrell, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 237

**Developmental dynamics of lncRNAs across mammalian organs and species**
Ioannis Sarropoulos, Ray Marin, Margarida Cardoso-Moreira, Henrik Kaessmann.
Presenter affiliation: Center for Molecular Biology (ZMBH), Heidelberg, Germany. 238

**Investigating atherosclerosis progression through single-cell transcriptional profiling of immune cells of the atherosclerotic plaque**
A Sartori, K Thanopoulou, C Borel, M Manioudaki, I Galani, E Andreakos, E T. Dermitzakis.
Presenter affiliation: University of Geneva, Geneva, Switzerland. 239

**Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation**
Thomas A. Sasani, Brent S. Pedersen, Ziyue Gao, Lisa Baird, Molly Przeworski, Lynn B. Jorde, Aaron R. Quinlan.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 240

**Retrotransposon activity in frontotemporal dementia**
Natalia Savytska, Kevin Menden, Tenzin Nyima, Margherita Francescatto, Patrizia Rizzu, Peter Heutink.
Presenter affiliation: German Centre for Neurodegenerative Diseases (DZNE), Tuebingen, Germany. 241

**Population genetics of *in vitro* neural differentiation**
Daniel Seaton, Julie Jerber, Daniel Pearce, Florian Merkle, Oliver Stegle, Daniel Gaffney.
Presenter affiliation: EMBL-EBI, Hinxton, United Kingdom. 242

**Dissemination of bioinformatics tools and workflows for reusability, findability, understandability through truwl.com**
Karl B. Sebby, Tim Thompson.
Presenter affiliation: xD Bio Inc., Whitefish, Montana. 243

**Application of CNVnator to analyze copy number alterations in cancer**
Shobana Sekar, Milovan Suvakov, Alexej Abyzov.
Presenter affiliation: Mayo Clinic, Rochester, Minnesota. 244

**A machine-learning method to detect non-coding regulatory variants in cancer genomes using ATAC-seq and ChIP-seq**
Arko Sen, Selene Tyndale, Graham McVicker.
Presenter affiliation: Salk Institute of Biological Studies, La Jolla, California. 245

**Multi-omics modules construction by non-parametric detection of conservation in features' ranks correlations in multiple cancers**
Sarun Sereewattanawoot, Ayako Suzuki, Masahide Seki, Yutaka Suzuki.
Presenter affiliation: The University of Tokyo, Chiba, Japan. 246

**Enzymatic Methyl-seq uncovers the regulatory backbone of the *Cannabis sativa* genome**
Brittany Sexton, Louise Williams, V.K. Chaithanya Ponnaluri, Keerthana Krishnan, Luo Sun, Yvonne Helbert, Lei Zhang, Christine Sumner, Fiona J. Stewart, Bradley W. Langhorst, Timothy Harkins, Kevin J. McKernan, Eileen T. Dimalanta, Theodore B. Davis.
Presenter affiliation: New England Biolabs, Inc., Ipswich, Massachusetts. 247

**Germline and somatic variant calling with NovaSeq™ 6000 2x250bp reads**
Minita Shah, Molly Johnson, Kanika Arora, Rashesh Sanghvi, Marta Byrska-Bishop, Wayne E. Clarke, Uday Evani, Kshithija Nagulapalli, Michael C. Zody, Soren Germer, Jade Carter, Giuseppe Narzisi, Nicolas Robine.
Presenter affiliation: New York Genome Center, New York, New York. 248

**Identifying survival-associated biomarkers from >17,000 cancer genomes**
Jason M. Sheltzer.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 249

**FTX-1821 effectively treat FSHD at its root cause, and specifically repress expression of DUX4 program genes**
Ning Shen, Alejandro Rojas, Erin Valentine, Peter Rahl, Anthony Accorsi, Joseph Maglio, Alan Robertson, Richard Barnes, Steven Kazmirski, Lorin A. Thompson, Owen Wallace, Diego Cadavid, Lucienne Renco, Angela Cacace, Aaron Chang.
Presenter affiliation: Fulcrum Therapeutics, Cambridge, Massachusetts. 250

**Identifying transcription factor binding site motifs from ENCODE epigenomic datasets by mapping position weight matrices to gapped k-mers**
Dustin Shigaki, Ayoti Patra, Michael A. Beer.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 251

**Natural selection on mutations affecting human complex traits**
Yuval B. Simons, Guy Sella.
Presenter affiliation: Stanford University, Stanford, California. 252

**Improving and using a chromosome-scale assembly of the enormous (32 Gb) axolotl genome**
Jeramiah J. Smith, Nataliya Timoshevskaya, Vladimir A. Timoshevskiy, Melissa C. Keinath, S. Randal Voss, AGC.
Presenter affiliation: University of Kentucky, Lexington, Kentucky. 253

**Genome reprogramming over eons and embryogenesis—Lessons from the sea lamprey (*Petromyzon marinus*)**
Jeramiah J. Smith, Cody Saraceno, Courtney K. Waterbury, Vladimir A. Timoshevskiy, Nataliya Timoshevskaya, SLGC.
Presenter affiliation: University of Kentucky, Lexington, Kentucky. 254

***Drosophila melanogaster* RNA expression patterns in flies selected for long and short night sleep duration**
Caetano Souto-Maior, Yanzhu Lin, Yazmin L. Serrano, Susan T. Harbison.
Presenter affiliation: National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland. 255

**Tracing the evolution of two transmissible cancers in Tasmanian devils**
Maximilian R. Stammnitz, Young Mi Kwon, Kevin C. Gori, Adrian Baez-Ortega, Tracy Wang, Rodrigo Hamede, Hannah V. Siddle, Mathew J. Garnett, Fengtang Yang, Zemin Ning, Elizabeth P. Murchison.
Presenter affiliation: University of Cambridge, Cambridge, United Kingdom. 256

**Mobile element insertions and associated structural variants occur early in longitudinal breast cancer samples**
Cody J. Steely, Julie E. Feusier, Yi Qiao, Gabor T. Marth, Lynn B. Jorde.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 257

**Gramene subsites—Pangenome browsers for crops**
Joshua C. Stein, Sharon Wei, Kapeel Chougule, Marcela Tello-Ruiz, Andrew Olson, Yinping Jiao, Bo Wang, Ivar Meijs, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 258

**The ENCODE data analysis pipelines are portable workflows for functional genomics**
J Seth Strattan, Ulugbek K. Baymuradov, Idan Gabdank, Keenan Graham, Ben C. Hitz, Otto A. Jolanki, Jin W. Lee, Vanessa Sochat, Anshul Kundaje, J Michael Cherry.
Presenter affiliation: Stanford University, Stanford, California. 259

**The impact of rare genetic variation on diverse transcriptomic signatures**
Benjamin J. Strober, Jonah Einson, Nicole Ferraro, Xin Li, Tuuli Lappalainen, Pejman Mohammadi, Stephen Montgomery, Alexis Battle.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 260

**Evolutionary analysis of sequence divergence of *cis*-regulatory modules associated with changes in chromatin and gene expression in brain tissue of primates**
Abusaid Shaimardanov, Dmitry Svetlichnyy.
Presenter affiliation: National Research University Higher School of Economics, Moscow, Russia. 261

**Human papillomavirus and the landscape of secondary genetic alterations in oral cancers**
Maura L. Gillison, Keiko Akagi, Weihong Xiao, Bo Jiang, Robert K. Pickard, Jingfeng Li, Benjamin J. Swanson, Amit D. Agrawal, Mark Zucker, Birgit Stache-Crain, Anne-Katrin Emde, Heather M. Geiger, Nicolas Robine, Kevin R. Coombes, David E. Symer.
Presenter affiliation: University of Texas MD Anderson Cancer Center, Houston, Texas. 262

**Assessing the accuracy of pipelines for variant calling, copy number calling, and subclonal deconvolution of heterogeneous tumour genomes**
Georgette Tanner, David R. Westhead, Alastair Droop, Lucy F. Stead.
Presenter affiliation: University of Leeds, Leeds, United Kingdom. 263

**Assessment of the impact on transcriptional output of rare fine-mapped intronic variants with large effect sizes on haematological traits in the UK Biobank cohort**
Manuel Tardaguila, Eve Coomber, Dragana Vuckovic, Nicole Soranzo.
Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom. 264

**Streamlining accessibility and computability of large-scale genomic datasets with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL)**
James Taylor, on behalf of the AnVIL project team.
Presenter affiliation: Johns Hopkins, Baltimore, Maryland. 265

**Familial language history links with population structure and admixture patterns**
Natalie Telis, Hamutal Arbel, Elena P. Sorokin, Ahna R. Girshick, Miao Zhang, J M. Granka, E L. Hong, K G. Chahine, C A. Ball.
Presenter affiliation: Ancestry.com LLC, San Francisco, California. 266

**Community curation strategies for improving genomic annotations and informing automated gene prediction**
Marcela K. Tello-Ruiz, Cristina F. Marco, Kapeel Chougule, Andrew Olson, Liya Wang, Doreen Ware, David Micklos.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 267

**Dual triage to identify poorly annotated transcripts in maize—The missing link in community curation**
Marcela K. Tello-Ruiz, Cristina F. Marco, Kapeel Chougule, Andrew Olson, Liya Wang, Doreen Ware, David A. Micklos.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 268

**Development of sequence variation graphs and graph-based software for genomics studies**
Devin Locke, Vladimir Semenyuk, Alexey Dolgoborodov, Amit Jain, John Browning, Serhat Tetikol, Ivan Johnson, Deniz Turgut, Ozem Kalay, Duygu Kabakci.
Presenter affiliation: Seven Bridges Genomics, Charlestown, Massachusetts. 269

**A new method to distinguish functional expression from background noise in transcriptomes**
Ammon Thompson, Michael R. May, Brian R. Moore, Artyom Kopp.
Presenter affiliation: UC Davis, Davis, California. 270

**Single-cell isoform RNA sequencing in mouse brain reveals isoform expression across cell types**
Hagen U. Tilgner.
Presenter affiliation: Weill Cornell Medicine, New York City, New York. 271

**DifCover—A computational pipeline for identifying structural variants in complex vertebrate genomes**
Nataliya Timoshevskaya, Vladimir Timoshevskiy, Melissa Keinath, Jeramiah Smith.
Presenter affiliation: University of Kentucky, Lexington, Kentucky. 272

**Fundamental biological processes are enriched in genetic sex differences**
Michela Traglia, Lauren A. Weiss.
Presenter affiliation: University of California, San Francisco, San Francisco, California. 273

**Searching for fertility genes using associations with cardiovascular disease**
Holly Trochet, Jean-Christophe Grenier, Marie-Julie Favé, Philip Awadalla, Julie Hussin.
Presenter affiliation: Université de Montréal, Montréal, Canada. 274

**A novel, population-robust, small-cohort risk stratification method identifies patients with 13-fold risk for cardiovascular disease**
Alex M. Tseng, Yosuke Tanigawa, Cole A. Deisseroth, Gill Bejerano.
Presenter affiliation: Stanford University, Stanford, California. 275

**"Semibulk" RNA-Seq analysis for revealing local cellular environments**
Issei Tsuchiya, Satoi Nagasawa, Seki Masahide, Yuta Kuze, Yutaka Suzuki.
Presenter affiliation: The University of Tokyo, Chiba, Japan. 276

**Mathematical modeling of aneuploid conception risk**
Katarzyna M. Tyc, Rajiv C. McCoy, Karen Schindler, Jinchuan Xing.
Presenter affiliation: Rutgers, The State University of New Jersey, Piscataway, New Jersey. 277

**Rare and common variants in the genetic architecture of intellectual disability in Northern Finland**
Lea Urpa, Mitja Kurki, Elmo Saarentaus, Olli Pietiläinen, Eija Hämäläinen, Padhraig Gormley, Minna Torniainen-Holm, Veikko Salomaa, Dennis Lal, Sini Kerminen, Matti Pirinen, Jaana Suvisaari, Elisa Rahikkala, Riikka Keski-Filppula, Merja Rauhala, Satu Korpi-Heikkilä, Jonna Komulainen-Ebrahim, Heli Helander, Päivi Vieira, Jukka S Moilanen, Jarmo Körkkö, Outi Kuismin, Mark J Daly, Aarno Palotie.
Presenter affiliation: Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. 278

**Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank**
Cristopher Van Hout, Ioanna Tachmazidou, Joshua Backman, Joshua Hoffman, Bin Ye, Ashutosh Pandey, Claudia Gonzaga-Jauregui, Nilanjana Banerjee, Suganthi Balasubramanian, David Ledbetter, Jeffrey Reid, Matthew Nelson, Jonathan Marchini, John Overton, Robert Scott, Goncalo Abecasis, Laura Yerges-Armstrong, Aris Baras.
Presenter affiliation: Regeneron Genetics Center, Tarrytown, New York. 279

**Tumor shedding into cell-free DNA (cfDNA) is associated with high-mortality cancers**
Oliver Venn, Earl Hubbell, Onur Sakarya, Christopher Chang, Meredith Miller, Kristan Steffen, Eric T. Fung, Anne-Renee Hartman, Tara Maddala, Alex M. Aravanis.
Presenter affiliation: GRAIL, Inc., Menlo Park, California. 280

**Large scale genomic testing for discovery in adult medicine**
Eric Venner, David Murdock, Tsung-Jung Wu, Sara Kalla, Victoria Yi, Yunyun Jiang, Kimberly Walker, Jianhong Hu, Viktoriya Korchina, Christie Kovar, Stacey Pereira, Christie Ballantyne, Mullai Murugan, Ginger Metcalf, Eric Boerwinkle, Donna M. Muzny, Richard Gibbs.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 281

**Leveraging evolutionary drivers of genetic variation reveals complexity underlying diseases across eMERGE epistasis study of 100,000 samples**
Shefali S. Verma, Pankhuri Singhal, Steven J. Schrodi, Sarah Pendergrass, Daniel Schaid, Iftikhar Kullo, David Fasel, Murray Brilliant, Ozan Dikilitas, Chunhua Weng, Hakon Hakonarson, Patrick Sleiman, Scott Hebbring, Alexander T. Frase, Scott Dudek, Anurag Verma, Marylyn D. Ritchie.
Presenter affiliation: University of Pennsylvania, Philadelphia, Pennsylvania. 282

**Homozygous *Ret* loss-of-function leads to precocious differentiation and changes in cell fate of enteric neural crest cells**
Elizabeth Vincent, Sumantra Chatterjee, Gabrielle Cannon, Dallas Auer, Holly Ross, Aravinda Chakravarti, Loyal Goff.
Presenter affiliation: Johns Hopkins School of Medicine, Baltimore, Maryland. 283

**Non-additive genetic effects on gene expression modulated by cell heterogeneity in whole blood**
A Viñuela, A A. Brown, A Martinez-Perez, N I. Panousis, A Ziyatdinov, M Sabater-Lleal, A Hamsten, J C. Souto, A Buil, J M. Soria, E T. Dermitzaki.
Presenter affiliation: University of Geneva, Geneva, Switzerland. 284

**Why details matter—The role of alternative splicing in glioblastoma cancer stem cells**
Kristoffer Vitting-Seerup, Yi C. Lim, Albin Sandelin, Petra Hamerlik.
Presenter affiliation: The Danish Cancer Society, Copenhagen, Denmark; University of Copenhagen, Copenhagen, Denmark. 285

**Mutational signatures are shaped by DNA damage and repair**
Nadezda Volkova, Bettina Meier, Peter J. Campbell, Anton Gartner, Moritz Gerstung.
Presenter affiliation: EMBL-EBI, Cambridge, United Kingdom. 286

**Genetic control of expression and splicing in developing human brain informs disease mechanisms**
Rebecca L. Walker, Gokul Ramaswami, Christopher Hartl, Nicholas Mancuso, Michael J. Gandal, Luis de la Torre-Ubieta, Bogdan Pasaniuc, Jason L. Stein, Daniel H. Geschwind.
Presenter affiliation: University of California, Los Angeles, Los Angeles, California. 287

**Haplotyping using full-length transcript sequencing in maize F1 hybrid reveals allele-specific expression**
Bo Wang, Elizabeth Tseng, Primo Baybayan, Kevin Eng, Michael Regulski, Yinping Jiao, Andrew Olson, Liya Wang, Kapeel Chougule, Peter Van Buren, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 288

**Access to MaizeCODE data**
Liya Wang, Zhenyuan Lu, Xiaofei Wang, Doreen Ware, Thomas Gingeras, MaizeCODE Consortium.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 289

**A web tool for physician-driven clinical diagnostic genomic analysis**
Tony Di Sera, Alistair Ward, Matt Velinder, Aditya Ekawade, Gabor Marth.
Presenter affiliation: University of Utah, Salt Lake City, Utah; Frameshift Genomics, Boston, Massachusetts. 290

**Functional characterization of programmatically eliminated genes in the sea lamprey**
Courtney KM Waterbury, Christine Brainson, Jeramiah J. Smith.
Presenter affiliation: University of Kentucky, Lexington, Kentucky. 291

**An investigation of DNA methylation correlation structure on the Illumina 450k methylation array**
Sarah H. Watkins, Matthew Suderman, Debbie A. Lawlor, Nicholas J. Timpson, Josine L. Min, Tom R. Gaunt.
Presenter affiliation: University of Bristol, Bristol, United Kingdom. 292

**Enzymatic methyl-seq—Next generation methylomes**
Louise Williams, V K Chaithanya Ponnaluri, Brittany Sexton, Lana Saleh, Katherine Marks, Mala Samaranayake, Bradley W. Langhorst, Zhiyi Sun, Thomas C. Evans, Jr, Romualdas Vaisvila, Eileen T. Dimalanta, Theodore B. Davis.
Presenter affiliation: New England Biolabs, Ipswich, Massachusetts. 293

**Mapping genome-wide DNA methylation patterns in gliomas in context of IDH gene mutation status and REST transcription factor binding**
Bartosz Wojtas, Michal J. Dabrowski, Agata Dziedzic, Michal Draminski, Rafal Guzik, Karolina Stepniak, Bartlomiej Gielniewski, Tomasz Czernicki, Pawel Nauman, Bartosz Czapski, Wieslawa Grajkowska, Katarzyna Kotulska, Bozena Kaminska.
Presenter affiliation: Nencki Institute of Experimental Biology of Polish Academy of Sciences, Warsaw, Poland. 294

**Assessment of bias and miRNA isoform quantifications of widely-used small RNA sequencing methods reveals major differences**
Carrie Wright, Anandita Rajpurohit, Emily E. Burke, Courtney Williams, Leonardo Collado-Torres, Martha Kimos, Nicholas J. Brandon, Alan J. Cross, Andrew E. Jaffe, Daniel R. Weinberger, Joo Heon Shin.
Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland; AstraZeneca, Cambridge, Massachusetts. 295

**Predicting differential expression of drought-associated genes in maize**
Travis Wrightsman, Jeremy Pardo, Jacob Washburn, Hai Wang, Edward Buckler.
Presenter affiliation: Cornell University, Ithaca, New York. 296

**Caste- and pesticide-specific effects of neonicotinoid pesticide exposure on gene expression in bumblebees**
Thomas J. Colgan, Andres Arce, Richard Gill, Yannick Wurm.
Presenter affiliation: Queen Mary University of London, London, United Kingdom. 297

**Building a comprehensive ENCODE long read transcriptome pipeline**
Dana Wyman, Gabriela Balderrama-Gutierrez, Shan Jiang, Brian Williams, Barbara Wold, Ali Mortazavi.
Presenter affiliation: University of California, Irvine, Irvine, California. 298

**Issues in experimental design in enhancer–promoter interaction prediction**
Wang Xi, Michael A. Beer.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 299

**Single cell profiling on siRNA induced c-Myc silencing of human glioblastoma**
Bingqing Xie, Anindita Basu, Susan Olalekan, Natalia Maltsev, Eric Gauchat, Heather Eckart, Janine Hensel, Olga Volpert.
Presenter affiliation: The University of Chicago, Chicago, Illinois. 300

**Analysis of lung cancer splicing abnormality using Nanopore-sequencing**
Liu Xu, Miho Oka, Masahide Seki, Ayako Suzuki, Yutaka Suzuki.
Presenter affiliation: Laboratory of Systems Genomics, Kashiwa, Japan. 301

**Detection of circulating miRNA levels in large cohort schizophrenia**
Qi Xu.
Presenter affiliation: Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. 302

**Likelihood model to estimate strength of selection and genomic architecture underlying a trait using GWAS summary statistics**
Alexander T. Xue, Yi-Fei Huang, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 303

**Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns**
Xiaofei Yang, Lin Gao, Shihua Zhang.
Presenter affiliation: Xi'an Jiaotong University, Xi'an, China. 304

**Motif-Raptor—Cell type specific prioritization of transcription factor and genetic variant based on GWAS summary statistics**
Qiuming Yao, Paolo Ferragina, Yakir Reshef, Daniel Bauer, Luca Pinello.
Presenter affiliation: Boston Children's Hospital, Boston, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts. 305

**MoMI-G—Modular Multi-scale Integrated Genome Graph Browser**
Toshiyuki T. Yokoyama, Sakamoto Yoshitaka, Masahide Seki, Yutaka Suzuki, Masahiro Kasahara.
Presenter affiliation: The University of Tokyo, Chiba, Japan. 306

**Epigenetics of type 2 diabetes in the Qatari population**
Noha A. Yousri, Omar Albagha, Steven C. Hunt.
Presenter affiliation: Weill Cornell Medicine-Qatar, Doha, Qatar. 307

**Germline genetic variation drives the somatic landscape of tumors**
Noah Zaitlen, Serghei Mangul, Alexander Gusev.
Presenter affiliation: UCLA, Los Angeles, California. 308

**The Ensembl post-GWAS analysis pipeline**
Daniel R. Zerbino, Verena Zuber, Seongwon Hwang, Gareth Peat, Michael Nuhn, José C. Marugán, William Jones, William Newell, Oliver Stegle, Ian Dunham.
Presenter affiliation: European Molecular Biology Laboratory, Hinxton, United Kingdom; Open Targets, Hinxton, United Kingdom. 309

**Deep learning-augmented RNA-seq analysis of transcript splicing**
Zijun Zhang, Zhicheng Pan, Yi Ying, Zhijie Xie, Samir Adhikari, John Phillips, Russ P. Carstens, Douglas L. Black, Yingnian Wu, Yi Xing.
Presenter affiliation: University of California, Los Angeles, Los Angeles, California. 310

**Genome-wide interrogation of human positive regulatory elements**
Nanxiang (Samuel) Zhao, Kayvan Najarian, Harm Derksen, Alan P. Boyle.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 311

**Comprehensive identification and functional analysis of cis regulatory elements and 3D architecture of the pig genome**
Yunxia Zhao, Ye Hou, Mingyang Hu, Yueyuan Xu, Yu Luan, Tingting Liu, Hongbo Yang, Feng Yue, Xinyun Li, Shuhong Zhao.
Presenter affiliation: Huazhong Agricultural University, Wuhan, China. 312

**Profiling of chromatin accessibility and identification of general cis-regulatory mechanisms that control two ocular lens differentiation pathways**
Yilin Zhao, Deyou Zheng, Ales Cvekl.
Presenter affiliation: Albert Einstein College of Medicine, Bronx, New York. 313

**Novel molecular signatures for heart failure—A cohort study of mRNA, lncRNA, and microRNA expression**
Yin-Ying Wang, Xiumeng Hua, Qing Xiong, Peilin Jia, Jiangping Song, Zhongming Zhao.
Presenter affiliation: The University of Texas Health Science Center at Houston, Houston, Texas. 314

**Hybrid *de novo* and reference assisted genome assembly with MaSuRCA**
Aleksey V. Zimin.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 315

FRIDAY, May 10—4:30 PM

## GUEST SPEAKERS

**Molly Przeworski**
Columbia University

**"What drives the dependence of human germline mutation rates on sex, age, and time?"**

**Robert Waterston**
University of Washington

**"A transcriptional program for the *C. elegans* embryo at single cell resolution—My life with worms"**

FRIDAY, May 10

**BANQUET**

Cocktails 6:00 PM
*Sponsored by* **Thermo Fisher**

Dinner 6:45 PM

SATURDAY, May 11—9:00 AM

**SESSION 11** COMPLEX TRAITS AND MICROBIOME

**Chairpersons:** **Guy Sella,** Columbia University New York, New York
**Joao Xavier,** Memorial Sloan Kettering Cancer Center, New York, New York

**Should "important genes" account for greater heritability in complex traits?**
Guy Sella, Yuval B. Simons, Jonathan K. Pritchard.
Presenter affiliation: Columbia University, New York, New York. 316

**Poor portability of polygenic scores even within an ancestry group**
Hakhamanesh Mostafavi, Arbel Harpak, Dalton Conley, Jonathan K. Pritchard, Molly Przeworski.
Presenter affiliation: Columbia University, New York, New York. 317

**Quantifying the heritability of the gut microbiota using 13,780 longitudinal microbiome profiles**
Laura Grieneisen, Mauna Dasari, Trevor Gould, Johannes Bjork, Jean-Christophe Grenier, Vania Yotova, David Jansen, Neil Gottel, Jack Gilbert, Luis Barreiro, Susan C. Alberts, Jeanne Altmann, Jenny Tung, Elizabeth Archie, Ran Blekhman.
Presenter affiliation: University of Minnesota, Minneapolis, Minnesota. 318

**Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture**
Bogdan Pasaniuc, Kathryn S. Burch, Kangcheng Hou, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman.
Presenter affiliation: UCLA, Los Angeles, California. 319

**Microbiome-host interaction studied directly in humans**
Joao Xavier.
Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York. 320

**Upstream stimulatory factors are pleiotropic meta regulators**
Hanna M. Ollila, Nasa Sinnott-Armstrong, Kirsi-Marja Zitting, Tiina Paunio, Tarja Stenberg, Richa Saxena, Jonathan Pritchard.
Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts; Stanford University, Palo Alto, California. 321

**The genetic architecture and phenotypic prediction of blood cell production**
Dragana Vuckovic, on behalf of BCX Consortium.
Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom. 322

**Longitudinal analysis of the gut microbiome reveals dynamic changes in relation to medications and phenotypes**
Lianmin Chen, Sanzhima Garmaeva, Alexander Kurilshikov, Ranko Gacesa, Arnau Vich Vila, Rinse Weersma, Cisca Wijmenga, Alexandra Zhernakova, Jingyuan Fu.
Presenter affiliation: University Medical Center Groningen, Groningen, the Netherlands. 323

Abad, Amaya, 213
Abd-Rabbo, Diala, 224
Abecasis, Goncalo, 279
Abel, Haley J., 16, 52
Abyzov, Alexej, 17, 25, 244
Accorsi, Anthony, 250
Adams, David J., 94
Adamson, Aaron W., 18
Adenekan, Philip, 90, 153
Adhikari, Samir, 310
Aduluso-Nwaobasi, Dike, 224
Afzal, Veena, 216
Agarwal, Vikram, 19
Agrawal, Amit D., 262
Agrawal, Raag, 119
Agrawal, Saumya, 8
Ahituv, Nadav, 7
Ahmed, Musaddeque, 131
Akagi, Keiko, 262
Akawi, Nadia, 113
Akiyama, Jennifer A., 216
Akiyama, Masato, 53
Alazizi, Adnan, 86, 225
Albagha, Omar, 307
Albergante, Luca, 198
Albert, Frank W., 20, 42
Alberts, Susan C., 318
Aldam, Gabriel, 83
Alföldi, Jessica, 64, 236
Alizada, Azad, 224
Alonges, Michael, 13
Altmann, Jeanne, 318
Alver, Maris, 21
Alves-Fernandes, Noemia-Rita, 70
Amador-Noguez, Daniel, 61
Amador-Rios, Raziel, 65
Amatya, Shradha, 68
Ambroise, Christophe, 206
Amit, Ido, 95
Anantharaman, Thomas, 150
Andersen, Stig U., 48
Anderson, Benton J., 194
Andreakos, E, 239
Antonarakis, Stylianos, 75
Antonescu, Christina, 103
Arauna, Lara R., 22, 116
Aravanis, Alexander M., 148, 280
Arbel, Hamutal, 266
Arce, Andres, 297
Archie, Elizabeth, 318
Arciero, Elena, 23
Ardlie, Kristin, 12
Arendt, M L., 236
Argelaguet, Ricard, 24
Arkin, Adam, 146
Arnold, Christian, 34
Arora, Kanika, 248
Artiles, Karen L., 173
Artomov, Mykyta, 196
Aryee, Martin J., 198
Ashuach, Tal, 7
Aslett, Louis, 59
Atla, Goutham, 217
Auer, Dallas, 50, 283
Awadalla, Philip, 274
Axelsson, E, 236

Bachmann, Asger, 48
Backman, Joshua, 279
Bae, Taejeong, 25
Baez-Ortega, Adrian, 256
Bailey, Matthew H., 26
Baird, Lisa, 240
Balasubramanian, Suganthi, 279
Balderrama-Gutierrez, Gabriela, 298
Ball, Catherine A., 179, 266
Ballantyne, Christie, 281
Ballouz, Sara, 27, 101
Banerjee, Nilanjana, 279
Banks, Eric, 64
Banovich, Nicholas E., 28
Baras, Aris, 279
Barbazuk, William, 123
Barnes, Richard, 250
Barnett, Alice, 187
Barreiro, Luis, 56, 318
Barreiro, Rodrigo, 91
Barrett, Jeffrey, 108, 110

Barrett, Lindy, 196
Barthel, Floris P., 29
Bartom, Elizabeth T., 30
Baslan, Timour, 103
Basu, Anindita, 300
Batra, Sanjit S., 31
Battle, Alexis, 126, 260
Batzoglou, Serafim, 84
Bauduer, Frédéric, 87
Bauer, Daniel E., 198, 220, 305
Baur, Brittany A., 32
Baybayan, Primo, 123, 288
Baymuradov, Ulugbek K., 90, 138, 153, 259
Beal, Brian, 168
Beer, Michael A., 33, 181, 189, 251, 299
Beichman, Annabel, 207
Bejerano, Gill, 275
Benaglio, Paola, 71
Bender, Katharina, 6
Bennett, David A., 141
Berest, Ivan, 34
Bertolino, Alessandro, 192
Besenbacher, Søren, 160
Beveridge, Allan J., 60
Bhetariya, Preetida J., 35
Bhutani, Kunal, 208
Bidarimath, Mallikarjun, 231
Bielas, Stephanie, 47
Birney, Ewan, 10
Bjork, Johannes, 318
Bjornsson, Hans, 40
Black, Douglas L., 310
Blasi, Giuseppe, 192
Blekhman, Ran, 318
Blibaum, Daniel, 36, 76
Blumberg, Amit, 37
Blundell, Jamie R., 38
Bock, Christoph, 214
Boehnke, Michael, 52, 121
Boeke, Jef D., 162, 219
Boerwinkle, Eric, 122, 149, 281
Boix, Carles, 141
Bojic, Lazar, 208
Bonàs-Guarch, Silvia, 217
Bonder, Marc Jan, 45, 68
Bonn, Stefan, 70
Bono, Hidemasa, 39
Borel, C, 239
Borel, Christelle, 75, 107
Borodkin, Mark, 150
Boubli, Jean P., 84, 145
Bouffard, Gerard G., 143
Boukas, Leandros, 40
Boutros, Paul C., 131, 224
Boyle, Alan P., 47, 82, 163, 184, 311
Brademan, Dain R., 194
Brainson, Christine, 291
Brand, Harrison, 64
Brandon, Nicholas J., 295
Brazma, Alvis, 74, 94
Breen, M, 236
Breschi, Alessandra, 213
Brettner, Leandra M., 144
Breuss, Martin W., 41
Brilliant, Murray, 282
Brion, Christian, 20, 42
Bristow, Robert G., 131
Brooks, Angela N., 74
Brooks, Shelise, 143
Brotman, S M., 43
Brown, A A., 284
Brown, Caitlin, 207
Brown, Christopher D., 126
Browning, John, 269
Buckler, Edward, 166, 296
Buenrostro, Jason D., 198
Bui, Linh, 28
Buil, A, 284
Burch, Kathryn, 14, 319
Burchert, Magdalena, 92
Burke, Emily E., 5, 204, 295
Burns, Monika A., 73
Bussemaker, Harmen J., 119
Butterfield, Andrew, 26
Butterworth, Adam S., 93
Buyske, Steven, 16
Buzzo, Jose L., 91
Byrnes, Jake, 179
Byrska-Bishop, Marta, 248

Cacace, Angela, 250
Cáceres, Mario, 44, 147, 151
Cadavid, Diego, 250

Cai, Na, 45
Calabrese, Claudia, 74
Calvert, Fiona, 115
Camilleri-Robles, Carlos, 65
Cammack, Alex J., 3
Campbell, Catarina D., 191
Campbell, Peter J., 112, 158, 286
Cannon, Gabrielle, 283
Cantsilieris, Stuart, 202
Canver, Matthew, 220
Cao, Chen, 155
Capra, John A., 88
Cardoso-Moreira, Margarida, 238
Carlson, Jedidiah, 46
Carninci, Piero, 8
Carstens, Russ P., 310
Carter, Jade, 248
Cass, Richard W, 103
Cassar, Olivier, 22, 116
Castel, Stephane, 119
Castillo Lizardo, Melissa G., 70
Castro, Christopher, 47
Cavassim, Maria Izabel A., 48
Cebola, Inês, 217
Chahine, Ken G., 179, 266
Chakravarti, Aravinda, 50, 283
Chan, Winston, 81
Chan, Yingguang Frank, 49
Chang, Aaron, 250
Chang, Christopher, 280
Chang, Howard, 2
Chappell, James, 2
Charpentier, Guillaume, 206
Chatterjee, Sumantra, 50, 283
Chatzipli, Aikaterini, 112
Chen, Hsiuyi, 164
Chen, Huidong, 198
Chen, James, 33
Chen, Junjie, 51
Chen, Lei, 52
Chen, Lianmin, 323
Chen, Minhui, 53
Chen, Qiang, 192
Chen, Wei-Min, 227
Chen, Xuhua, 3, 170
Cheng, Xiaoheng, 54
Cherry, J Michael, 90, 138, 153, 259
Chhatriwala, Mariya, 68
Chi, Hongbo, 186
Chia, Shumei, 81
Chiang, Charleston, 53, 55
Chiang, Colby, 16
Chick, Joel M., 102
Chinnery, Patrick, 45
Chmatal, Lukas, 102
Choin, Jeremy, 22, 56, 116
Choudhary, Saket, 57
Chougule, Kapeel M., 58
Chougule, Kapeel, 58, 258, 267, 268, 288
Christ, Ryan, 59
Chundru, V. Kartik, 60
Chung, Mei-i, 28
Clark, Nathan, 209
Clark, Ryan L., 61
Clark, Stephen J., 24
Clarke, Declan, 98
Clarke, Wayne E., 248
Clement, Kendell, 220
Coban Akdemir, Zeynep, 122
Cohen, Tal, 62
Cole, Mitchel A., 220
Colgan, Thomas J., 297
Coll, Moises, 118
Collado-Torres, Leonardo, 63, 204, 295
Collins, Ryan L., 64
Comas, David, 87
Conceicao, Helena, 91
Conley, Dalton, 317
Conrad, Donald F., 12
Cookson, Mark, 129
Coomber, Eve, 264
Coombes, Kevin R., 262
Coon, Hilary, 85
Coon, Joshua J., 194
Corominas, Montserrat, 65
Cottingham, Bob, 146
Couto Alves, Alexessander, 66
Couto, G, 236
Couturier, Dominique L., 94
Cremers, Claudia, 208

Cristea, Mihaela C., 18
Cross, Alan J., 295
Crotty, William, 106
Crow, Megan, 101
Crystal, Ronald G., 229, 232
Cui, Hongzhu, 67
Cukuroglu, Engin, 74, 81
Cunniff, Christopher M., 49
Cuomo, Anna, 68
Currant, Hannah, 10
Currin, K W., 193
Cvekl, Ales, 313
Czapski, Bartosz, 294
Czernicki, Tomasz, 294

Da Cruz Paula, Arnaud, 103
Dabrowski, Michal J., 69, 294
Dai, Yulin, 190
Dalmia, Anupriya, 70
Daly, Mark J., 64, 106, 196, 278
Danesh, John, 45
Danko, Charles G., 37
D'Antonio, Matteo, 71
D'Antonio-Chronowska, Agnieszka, 71
Das, Indraniel, 16, 52
Dasari, Mauna, 318
DasGupta, Ramanuj, 81
David, Eyal, 95
David, Julianne K., 72
Davidson, Natalie, 203
Davies, Megan, 158
Davila-Velderrain, Jose, 141
Davis, Brian W., 168
Davis, J P., 193
Davis, Sean, 63
Davis, Theodore B., 247, 293
Dawe, Kelly, 58
Day, Christopher R., 226
de Groot, John F., 29
de Hoon, Michiel, 8
de la Torre-Ubieta, Luis, 287
de Lange, Titia, 112
Dean, Wendy, 24
Deary, Ian J., 60
Decker, Jared E., 79, 120, 180
Deep-Soboslay, Amy, 5
DeGiorgio, Michael, 54
Degnan, Bernard, 211
Deisseroth, Cole A., 275
del Rosario, Ricardo C., 73
delaBastide, Melissa, 105
Delaneau, Olivier, 75, 107, 151
Deleuze, Jean-François, 116, 206
Dellinger, Thanh H., 18
Delprat, Alejandra, 44, 151
Demas, Vasiliki, 148
Demircioglu, Deniz, 74
DePaolo, William R., 144
Derksen, Harm, 311
Dermitzakis, Emmanouil T., 75, 107, 151, 157, 239, 284
Di Carlo, Pasquale, 192
Di Sera, Tonya, 80, 201, 290
Diao, James, 233
Dickel, Diane E., 216
Dikilitas, Ozan, 282
Dimalanta, Eileen T., 123, 247, 293
Dindler  Rasmussen, Kasper, 34
Ding, Yuan Chun, 18
Diss, Guillaume, 215
Dobin, Alexander, 27, 36, 76
Docherty, Anna, 85
Doebley, John, 165
Dolgoborodov, Alexey, 269
D'Oliveira Albanus, Ricardo, 77
Domingo, Júlia, 215
Dong, Shengcheng, 47
Donnelly, Peter, 110
Donovan, Jerry, 208
Donovan, Margaret K., 71
Dostie, Josée, 8
Dougherty, Joseph D., 3
Draminski, Michal, 69, 294
Dréau, Andreea, 49
Droop, Alastair, 263
Drubin, Casey, 12
Druely, Todd E., 38
Dubé, Marie-Pierre, 92
Dudek, Scott, 282
Dukler, Noah E., 78
Dunham, Ian, 309
Durbin, Harly J., 79
Dutcher, Susan K., 16

Dziedzic, Agata, 69, 294

Easton, John, 186
Eaton, Zachary, 123
Ebling, Heather, 123
Eckart, Heather, 300
Edgley, Mark L., 173
Eggan, Kevin, 99, 106, 196
Eichler, Evan E., 202
Einson, Jonah, 260
Ek, Weronica, 222
Ekawade, Aditya, 80, 290
Ekman, Diana, 222
Ellis, Shannon E., 63
Elvers, I, 236
Ema, Masatsugu, 130
Emani, Prashant, 98
Emde, Anne-Katrin, 262
Eng, Christine, 81, 122
Eng, Kevin, 288
Eng, Nicholas, 33
Englund, Melissa, 82
Ennes, Felipe, 84, 145
Eran, Alal, 62
Erez, O, 199
Erny, Daniel, 95
Eskin, Eleazar, 9, 159
Esko, Tõnu, 21, 151
Evani, Uday, 248
Evans, Jr, Thomas C., 293

Fairley, Susan, 83
Farabella, Irene, 217
Farh, Kyle Kai-How, 84, 145
Farouni, Rick, 220
Farrell, Allison K., 225
Farrell, Andrew, 85
Fasching, Liana, 25
Fasel, David, 282
Favé, Marie-Julie, 274
Federico, Gaiti, 211
Feng, Guoping, 73
Fengler, Kevin, 58
Ferragina, Paolo, 305
Ferraro, Nicole, 260
Ferrer, Jorge, 217
Feusier, Julie E., 257
Fiaux, Patrick C., 164
Fields, Bryden, 48
Findley, A S., 86
Finkbeiner, Steve, 129
Fire, Andrew Z., 173
Firth, Helen V., 113
Fischer, Jonathan, 31
Fisher, Daniel S., 38
Fiske, Brian, 208
Fitzgerald, Tomas, 10
FitzPatrick, David R., 113
Flesken-Nikitin, Andrea, 231
Flicek, Paul, 83, 89, 230
Flores-Bello, André, 87
Flynn, Elise D., 119
Fong, Sarah L., 88
Fonseca, Nuno A., 74, 94
Forneris, Mattia, 6
Forsberg-Nilsson, K, 236
Fortier, Kevin, 185
Foster, Paul J., 10
Franc, Sylvia, 206
Francescatto, Margherita, 241
Francioli, Laurent C., 64
Francois, Mathias, 211
Frankish, Adam, 89
Franklin, Robin, 115
Frase, Alexander T., 282
Fraser, Michael, 131
Fraser, Peter, 217
Frazer, Kelly A., 71
Freddolino, Peter, 163
Freimer, Nelson B., 52
Friedman, Robin, 208
Froment, Alain, 56
Fu, Jingyuan, 323
Fung, Eric T., 280
Fungtammasan, Chai, 58
Furey, Terrence S., 140
Furlong, Eileen E., 6

Gabdank, Idan, 90, 138, 153, 173, 259
Gacesa, Ranko, 323
Gaffney, Daniel, 114, 115, 187, 242
Galani, I, 239
Galante, Pedro A., 91
Galeev, Timur, 233

Gallone, Guiseppe, 108
Gamache, Isabel, 92
Gandal, Michael J., 287
Ganel, Liron, 52
Ganesan, Shridar, 105
Ganna, Andrea, 196
Gao, Hong, 84
Gao, Lin, 304
Gao, Xian, 73
Gao, Ziyue, 240
Garber, Manuel, 176
Garcia-Bernardo, Jose, 68
Gardner, Eugene J., 93, 108
Garg, Manik, 94
Garmaeva, Sanzhima, 323
Garnett, Mathew J., 256
Garside, Victoria, 211
Gartner, Anton, 286
Gauchat, Eric, 300
Gaunt, Tom R., 292
Gauthier, Laura D., 64
Gayed, George, 178
Gehrke, Jason M., 220
Geiger, Heather M., 262
Geirsdottir, Laufey, 95
Gelfand, Ellen, 12
Geller, Evan, 221
Genovese, Giulio, 96, 99, 117
Georges, Stephanie J., 97
Germer, Soren, 248
Gerstein, Mark, 98, 178, 233, 237
Gerstung, Moritz, 286
Gerton, Jennifer L., 143
Geschwind, Daniel H., 98, 223, 287
Gessain, Antoine, 22, 116
Ghareghani, Maryam, 111
Ghosh, Sulagna, 99
Ghoussaini, Maya, 100
Giannuzzi, Giuliana, 75
Gibbs, Raphael J., 129
Gibbs, Richard, 122, 281
Gielniewski, Bartlomiej, 69, 294
Gilad, Y, 86
Gilbert, David M., 128
Gilbert, Jack, 318
Gill, Richard, 297
Gillis, Jesse, 27, 101
Gillison, Maura L., 262
Giner-Delgado, Carla, 44
Gingeras, Thomas, 289
Giri, Anamika, 134
Girshick, Ahna R., 266
Gleeson, Joseph, 41
Godfrey, Alexander K., 102
Godwin, Andrew K., 231
Goeke, Jonathan, 81
Goff, Loyal, 50, 283
Göke, Jonathan, 74
Gomez, Celine, 187
Gomez-Duran, Aurora, 45
Gomez-Lopez, N, 199
Gonzaga-Jauregui, Claudia, 279
Goodwin, Sara, 13, 105, 136
Gorban, Alexander N., 198
Gordon, David, 202
Gori, Kevin C., 256
Gormley, Padhraig, 278
Gottel, Neil, 318
Gould, Trevor, 318
Graham, Keenan, 90, 138, 153, 259
Graig, David W., 129
Grajkowska, Wieslawa, 294
Granier, Michele, 206
Granka, Julie M., 179, 266
Grant, Struan, 66
Green, Anthony, 158
Greenleaf, Will, 2
Greenleaf, William J., 2, 4
Grenier, Jean-Christophe, 92, 274, 318
Grieneisen, Laura, 318
Griffith, Malachi, 12
Gu, Mengting, 98
Guan, Jihong, 198
Guardia, Gabriela, 91
Guigó, Roderic, 65, 213, 234
Guillen, Katrin P., 26
Gularte Mérida, Rodrigo, 103
Guo, Cong, 104
Gurjev, Victor, 161
Guschanski, Katerina, 84, 145
Gusev, Alexander, 308
Gut, Ivo, 84, 145

Gut, Marta, 84, 145
Gutierrez, Austin, 28
Guttman, Mitchell, 176
Guturu, Harendra, 179
Guzik, Rafal, 69, 294
Gygi, Steven P., 102

Habermann, Chris A., 28
Haiman, Christopher A., 55
Hakim, Amy A., 18
Hakonarson, Hakon, 282
Hall, Ira M., 16, 52, 59
Haluck-Kangas, Ashley, 30
Hämäläinen, Eija, 278
Hamede, Rodrigo, 256
Hamerlik, Petra, 285
Hammond, Christopher J., 10
Hammoud, Sue S., 152
Hamsten, A, 284
Han, Ernest S., 18
Handsaker, Juliet E., 108
Handsaker, Robert E., 96, 99, 121
Hanly, Joseph, 226
Hanna, Courtney, 24
Hannigan, Brett, 58
Hansen, Adam W., 122
Hansen, Kasper, 40
Hansen, Matthew, 31
Hansen, Nancy, 143
Harbison, Susan T., 255
Harewood, Louise, 230
Harkins, Timothy, 123, 247
Harlan, Blaine, 231
Harmant, Christine, 22, 56, 116
Harpak, Arbel, 317
Harrington, Anne N., 216
Harris, Kelley, 46
Hartl, Christopher, 287
Hartman, Anne-Renee, 280
Hasan, Milena, 172
Hashimoto, Shinichi, 135
Hassan, S, 199
Hastie, Alex, 150
Havrilla, James, 40
Hayakawa, Takashi, 84, 145
He, Chuan, 125
He, Housheng H., 131
He, June, 3
He, Liang, 141
He, Meng Xiao, 124
He, Xin, 125
He, Yuan, 126
Hebbring, Scott, 282
Heger, Andreas, 110
Helander, Heli, 278
Helbert, Yvonne, 123, 247
Helin, Kristian, 34
Henry, Chris, 146
Hensel, Janine, 300
Hensley, John, 77
Hermina-Perez, Jose J., 226
Herrero, Javier, 127
Heskett, Michael B., 128
Hetzel, Marion, 206
Heutink, Peter, 70, 129, 134, 241
Hickey, Peter, 40
Hilton, Jason, 90, 153
Hirose, Naoki, 130
Hirsch, Candice, 58
Hittinger, Chris T., 209, 219
Hitz, Benjamin C., 90, 138, 153, 259
Hodson, Dan, 158
Hoeppner, Daniel J., 5
Hoffman, Joshua, 279
Hoffner O'Connor, Michelle, 140
Höglund, Julia, 222
Holmes, Chris, 59
Holmes, J. Bradley, 195
Hon, Chung Chau, 8
Hong, Eurie L., 179, 266
Hormozdiari, Farhad, 9
Hou, Kangcheng, 14, 319
Hou, Ye, 312
Houlahan, Kathleen E., 131
Howald, Cedric, 107
Howe, Edmund, 143
Howe, Kerstin, 143
Howson, Joanna, 45
Hsu, Jonathan Y., 198, 220
Hu, Jianhong, 281
Hu, Mingyang, 312
Hua, Xiumeng, 314
Huang, Hailiang, 106
Huang, Ni, 12

Huang, Vincent, 224
Huang, Xiaomeng, 15, 26, 97
Huang, Yi-Fei, 37, 78, 303
Hubbell, Earl, 148, 280
Hufford, Matthew, 58, 165
Huganir, Richard, 106
Hume, David A., 60
Hunt, Steven C., 307
Hunter, Jesse, 169
Hurles, Matthew E., 23, 93, 108, 113
Hurto, Rebecca, 163
Huse, Jason T., 29
Hussin, Julie, 92, 132, 274
Hutter, Carolyn M., 133
Hwang, Seongwon, 309
Hyde, Thomas M., 5, 192, 204
Hysi, Pirro G., 10

Ichikawa, Kazuki, 173
Iles, Mark, 23
Illarionova, Anastasia, 134
Imaz, Maria, 187
Inoue, Fumitaka, 7
Inshaw, Jamie R., 227
Isaacs, Farren, 218
Isaev, Keren, 224
Ishigaki, Kazuyoshi, 53
Ito, Masayoshi, 8
Ivanov, Nikolay A., 204
Iwabuchi, Sadahiro, 135
Iyer, Gopal, 197
Iyer, Shruti, 136
Izquierdo, David, 44

Jabert, Paul, 206
Jaddoe, Vincent, 66
Jaffe, Andrew E., 5, 63, 192, 204, 295
Jaffe, Jake, 106
Jagla, Bernd, 172
Jain, Amit, 269
Jakobsson, Mattias, 116
Jakubosky, David, 71
Jamshidi, Arash, 148
Jansen, David, 318
Järvelin, Marjo-Riitta, 66
Javeed, Nida, 185
Javierre, Biola M., 217
Jerber, Julie, 114, 242
Jhangiani, Shalini, 122
Jia, Peilin, 137, 190, 314
Jia, Yankai, 204
Jiang, Bo, 262
Jiang, Shan, 298
Jiang, Yunyun, 281
Jiao, Yinping, 58, 258, 288
Johansson, Åsa, 222
John, Malina, 212
Johnson, Ivan, 269
Johnson, J, 236
Johnson, Kevin C., 29
Johnson, Matthew, 106
Johnson, Molly, 248
Johnson, Ruth D., 14
Jolanki, Otto, 138, 259
Jones, Felicity C., 212
Jones, William, 309
Joo, Jong Wha, 9
Jorde, Lynn B., 240, 257
Jou, Jennifer P., 90, 153
Joung, Keith, 220
Ju, Bensheng, 186

Kabakci, Duygu, 269
Kaessmann, Henrik, 238
Kahles, Andre, 74, 203
Kalay, Ozem, 269
Kalita, C A., 86
Kalla, Sara, 281
Kals, Mart, 21
Kamatani, Yoichiro, 53
Kamide, Yu, 135
Kaminska, Bozena, 69, 294
Kanchi, Krishna L., 16
Kane, Liam T., 123
Kaplanis, Joanna, 113
Kapourani, Chantriolnt--Andreas, 24
Karczewski, Konrad, 64, 99
Karlin-Neumann, George, 44
Karlsson, Torgny, 222
Kasahara, Masahiro, 306
Kashin, Seva, 96, 99, 121
Kasukawa, Takeya, 8
Kathiresan, Sekar, 64

Kathiria, Arwa, 2
Kato, Momoe, 216
Kattman, Brandi, 195
Kawai, Jun, 130
Kawaji, Hideya, 130
Kazmirski, Steven, 250
Keane, Thomas M., 139
Keinath, Melissa, 253, 272
Keith, Benjamin P., 140
Kellis, Manolis, 141
Kelly, Derek, 31
Kelsey, Gavin, 24
Kent, David, 158
Keren-Shaul, Hadas, 95
Kerin, Matthew, 142
Kerminen, Sini, 278
Keski-Filppula, Riikka, 278
Khandelwal, Garima, 127
Khavari, Paul, 2
Khaw, Peng T., 10
Khawaja, Anthony P., 10
Khayat, Michael, 122
Khera, Amit V., 64
Kierczak, Marcin, 222, 236
Kim, April, 106
Kim, Daniel S., 2
Kim, Hyunsung J., 148
Kim, J, 236
Kimos, Martha, 295
Kindermans, Martin, 74
Kingan, Sarah, 123
Kirov, George, 93
Kitchen, Robert, 233
Kitzman, Jacob, 77
Klein, Cecilia C., 65, 213, 234
Kleinman, Joel E., 5, 192, 204
Kliebhan, Margaret, 20
Knights, Andrew, 187
Knowles, James, 98
Knudsgaard, Per, 148
Ko, A, 43
Kohn, Andrea B., 174
Kominek, Jacek, 219
Komulainen-Ebrahim, Jonna, 278
Kondo, Naoto, 8
Kopp, Artyom, 270
Korbel, Jan O., 111, 202
Korchina, Viktoriya, 281
Koren, Sergey, 143
Korkin, Dmitry, 67
Körkkö, Jarmo, 278
Koronacki, Jacek, 69
Korpi-Heikkilä, Satu, 278
Kotulska, Katarzyna, 294
Kousi, Maria, 141
Kovar, Christie, 281
Krapohl, Eva, 110
Kreimer, Anat, 7
Kremitzki, Milinn G., 143
Krienen, Fenna, 73
Krishnan, Keerthana, 123, 247
Krogsgaard, Michelle, 162
Kropski, Jonathan A., 28
Krueger, Christel, 24
Krueger, Felix, 24
Kuchina, Anna, 144
Kucka, Marek, 49
Kuderna, Lukas F., 84, 145
Kuismin, Outi, 278
Kullo, Iftikhar, 282
Kumar, Vivek, 146
Kumari, Sunita, 146
Kumasaka, Natsuhiko, 115
Kundaje, Anshul, 2, 4, 259
Kurilshikov, Alexander, 323
Kurki, Mitja, 278
Kurokawa, Saki, 135
Kurtzman, Clete, 219
Kuze, Yuta, 276
Kwon, Young Mi, 256
Kyono, Yasuhiro, 77

Laakso, Markku, 43, 52, 193
LaBella, Abbe L., 219
Labelle, Myriam, 186
Lage, Kasper, 106
Lagunas, Tomas, 3
Lal, Dennis, 278
Lam, Ernest T., 150
Lander, Eric S., 16
Lanfear, D E., 86
Langenau, David M., 198
Langhorst, Bradley W., 247, 293
Lansdorp, Peter M., 161
Laplana, Marina, 147, 151

Lappalainen, Tuuli, 119, 260
Lareau, Caleb A., 198
Larena, Maximilian, 116
Larose, Hailey, 152
Larson, David E., 16, 52
Larson, Matthew H., 148
Lawlor, Debbie A., 292
Layer, Ryan M., 16, 149
Le Marchand, Loic, 55
Ledbetter, David, 279
Lee, Jin W., 90, 138, 153, 259
Lee, Joyce, 150
Leek, Jeffrey T., 63
Lee-Six, Henry, 158
Lehmann, Kjong, 74
Lehner, Ben, 215
Lemee, Laure, 22
Lemmon, Zachary, 13
Leopold, Philip L., 229
Lerga-Jaso, Jon, 44, 151
Leshchyk, Anastasia, 67
Li, He, 122
Li, Jingfeng, 262
Li, Jun Z., 152
Li, Liyuan, 186
Li, Pinghua, 166
Li, Wenzheng, 57
Li, Xiaoting, 119
Li, Xin, 260
Li, Xinyun, 312
Li, Yang I., 171
Liang, Minggao, 224
Lim, Yi C., 285
Lin, Khine Z., 90, 153
Lin, Meng, 55
Lin, Pin S., 18
Lin, Yanzhu, 255
Lindblad-Toh, K, 236
Lindsay, Tina G., 143
Lippman, Zachary, 13
Litton, Casey R., 90, 153
Liu, Biao, 123
Liu, David R., 220
Liu, Jianing, 58
Liu, Jiayue, 170
Liu, Ruize, 106
Liu, Shuang, 98
Liu, Tingting, 312
Llaca, Victor, 58
Lo Bosco, Giosue`, 198
Locke, Adam E., 52
Locke, Devin, 269
Lodge, Jean M., 194
Loell, Kaiser, 170
Loh, Po-Ru, 96, 117
Lohmueller, Kirk, 207
Lollar, Matthew J., 154
Long, Quan, 155
Loose, Matthew, 143
Lopez, Marie, 56
Lorant, Anne, 165
Lou, Blanpain, 5
Lowy, Ernesto, 83
Lu, Shengjie, 155
Lu, Xuewen, 155
Lu, Zhenyuan, 146, 156, 289
Luan, Yu, 312
Luca, Francesca, 86, 199, 225
Lunter, Gerton, 110
Luo, Kaixuan, 125
Luo, Yunhai, 90, 153
Lupien, Mathieu, 224
Lupski, James R., 122
Luthra, Ishika, 164
Lutz, Sheila, 20, 42
Ly, Lam-Ha, 81
Lykoskoufis, Nikolaos M., 107, 157

Ma, Liang, 204
Ma, Qianyi, 152
MacArthur, Daniel G., 64
Machado, Heather, 158
Maciejowski, John, 112
MacNamara, Aidan, 104
Maddala, Tara, 280
Maden, Sean K., 72
Maduro, Valerie, 143
Mägi, Reedik, 21
Maglio, Joseph, 250
Maglott, Donna, 195
Mair-Meijers, Henriette E., 225
Majumdar, Arunabha, 319
Malolepsza, Edyta, 106
Maltsev, Natalia, 300
Mancuso, Nicholas, 287, 319

Mangino, Massimo, 172
Mangul, Serghei, 159, 308
Manioudaki, M, 239
Manske, Gabriel, 152
Marchini, Jonathan, 142, 279
Marco, Cristina F., 267, 268
Maretty, Lasse, 160
Marin, Ray, 238
Marinescu, V D., 236
Marinov, Georgi K., 4
Marioni, John C., 24, 68, 172
Marioni, Riccardo E., 60
Marks, Katherine, 293
Marques-Bonet, Tomas, 84, 145
Marschall, Tobias, 111, 202
Marth, Gabor T., 15, 26, 35, 80, 85, 97, 201, 257, 290
Martin, Arnaud, 226
Martin, Hilary C., 23, 93, 113
Martinez, Iker, 68
Martinez-Cruzado, Juan C., 229
Martinez-Perez, A, 284
Martinez-Ruiz, Carlos, 210
Marti-Renom, Marc A., 217
Marugán, José C., 309
Masahide, Seki, 276
Mason, Dan, 23
Massingham, Tim, 105
Mastroianni, Justin, 162
Mathys, Hansruedi, 141
Matise, Tara C., 16
Matsumoto, Mitsuyuki, 5
Mattsson, Carl-Adam, 161
May, Michael R., 270
McCalla, Sunnie Grace, 197
McCarroll, Steven A., 73, 96, 99, 117, 121
McCarthy, Davis J., 68
McCarthy, Mark I., 66
McCombie, W. Richard, 13, 136
McCoy, Rajiv C., 277
McCulloch, Laura H., 162
McDaniel, Jennifer, 183
McDonald, Torrin, 82, 163
McDonel, Patrick, 176
Mcglinn, Edwina, 211
McKernan, Kevin J., 123, 247
McRae, Allan F., 60
McVean, Gilean, 110
McVicker, Graham, 164, 245
Meadows, J, 236
Mei, Wenbin, 165
Meier, Bettina, 286
Meiers, Sascha, 111
Meijs, Ivar, 258
Mejia-Guerra, Maria K., 166
Melin, Amanda D., 84, 145
Melo, Curtis J., 73
Menden, Kevin, 241
Mendieta-Esteban, Julen, 217
Mendoza-Revilla, Javier, 22, 116
Merkle, Florian, 99, 114, 242
Metcalf, Ginger, 281
Metspalu, Mait, 167
Metzger, Michael J., 168
Meyer, Wynn K., 209
Meyerson, William, 237
Meyn, Stephen, 169
Mezey, Jason G., 232
Micklos, David, 267, 268
Miga, Karen H., 143
Miguel-Escalada, Irene, 217
Miller, Chase, 80, 201
Miller, Meredith, 280
Miller, Thiago, 91
Milosavljevic, Aleksander, 233
Min, Josine L., 292
Mirauta, Bogdan, 45
Mishmar, Dan, 62
Mitchell, Emily, 158
Mitchell, Richard N., 102
Mitra, Robi D., 3, 170
Mittleman, Briana E., 171
Modiano, J F., 236
Moeskjær, Sara, 48
Mohammadi, Pejman, 119, 260
Mohammadi, Shahin, 141
Mohammed, Hisham, 24
Mohlke, K L., 43, 193
Moilanen, Jukka S., 278
Montgomery, Grant W., 60
Montgomery, Stephen, 260
Moore, Brian R., 270
Mordechai, Chen, 62
Morgan, Michael D., 172
Morishita, Shinichi, 173

Moroz, Leonid L., 174, 175
Morris, Samantha A., 3
Mortazavi, Ali, 138, 298
Moskalik, Anzhela D., 29
Moslemi, Camous, 48
Mostafavi, Hakhamanesh, 317
Moudgil, Arnav, 3
Mouguiama-Daouda, Patrick, 56
Moyer, Eric, 195
Mullikin, James, 143
Murakawa, Yasuhiro, 130
Murali, Shwetha C., 202
Murchison, Elizabeth P., 256
Murdock, David, 281
Murillo, Oscar, 233
Murphy, Rachel, 176
Murugan, Mullai, 122, 281
Muzny, Donna M., 16, 122, 281
Myers, Zachary A., 90, 153

Nacu, Eugene, 106
Nagai, Luis A E., 177
Nagasawa, Satoi, 276
Nagulapalli, Kshithija, 248
Nagy, Reka, 110
Najarian, Kayvan, 311
Nakai, Kenta, 177, 188
Nalls, Mike, 129
Nandi, Tannistha, 74
Naqvi, Sahin, 102
Narzisi, Giuseppe, 248
Nauman, Pawel, 294
Navarro, Fabio, 98, 178
Neale, Benjamin M., 16, 64
Nehme, Ralda, 196
Nellore, Abhinav, 72
Nelson, Matthew, 104, 279
Neufeld, Stanley J., 212
Neuhausen, Susan L., 18
Newell, William, 309
Nguyen, Long, 186
Nguyen, Luong T., 179
Nichols, Jennifer, 24
Nichols, Richard A., 210
Nielsen, Rasmus, 11
Niemi, Mari, 93
Nikitin, Alexander Y., 231
Nikopensius, Tiit, 21
Nilson, Sara M., 180
Ning, Zemin, 256
Nishizaki, Sierra, 82
Nodzak, Conor, 205
Noonan, James P., 221
Nordborg, Magnus, 200
Noto, Keith, 179
Nuhn, Michael, 309
Nyima, Tenzin, 70, 241

Øbro, Nina, 158
O'Connor, Carolyn, 164
Ocsenas, Oliver, 224
Odom, Duncan T., 230
Oh, Jin-Woo, 33, 181
Oka, Miho, 301
Okada, Tomoyo, 103
Okada, Yukinori, 53
Olalekan, Susan, 300
Ollila, Hanna M., 321
Olson, Andrew, 182, 258, 267, 268, 288
Olson, Nathan D., 183
O'Neill, Emma, 90, 153
Onengut-Gumuscu, Suna, 227
Ongen, Halit, 107, 157
Opulente, Dana A., 219
Ouyang, Ningxin, 184
Overton, John, 279
Oyharçabak, Bernard, 87

Pacheco, Sarai, 44
Paczkowska, Marta, 224
Page, David C., 102
Pai, Athma A., 185
Pajukanta, P, 43, 193
Palla, Giovanni, 34
Palotie, Aarno, 52, 196, 278
Palumbo, Emilio, 213
Pan, Qingfei, 186
Pan, Wenying, 148
Pan, Zhicheng, 310
Pandey, Ashutosh, 279
Pang, Andy Wing Chun, 150
Panousis, Nikolaos I., 187, 284
Papa, Riccardo, 226
Papalino, Marco, 192
Papula, Alana, 38

Pardo, Jeremy, 296
Park, Alexander, 185
Park, Jonathan, 233
Park, Sung-Joon, 177, 188
Parker, Stephen C.J., 77
Parts, Leopold, 108
Pasaniuc, Bogdan, 14, 287, 319
Patel, Praveen, 10
Patin, Etienne, 22, 56, 116, 172
Pato, Carlos N., 121
Pato, Michele T., 121
Patra, Ayoti, 33, 189, 251
Patterson, Kristen, 171
Paulino-Ramirez, Robert A., 229
Paunio, Tiina, 321
Pearce, Daniel, 114, 242
Peat, Gareth, 309
Pedersen, Brent S., 26, 149, 240
Pei, Guangsheng, 137, 190
Pena, Maria, 101
Pendergrass, Sarah, 282
Pennacchio, Len A., 216
Pereira, Bernard, 191
Pereira, Stacey, 281
Pérez-Lluch, Sílvia, 213, 234
Pergola, Giulio, 192
Perrin, H J., 193
Perry, George H., 56
Peter, Marcus E., 30
Peters-Clarke, Trenton M., 194
Petrossian, Natalie, 106
Pettersson, M E., 236
Pham, Quan T., 216
Phan, Lon, 195
Phillippy, Adam M., 143
Phillips, John, 310
Pickard, Robert K., 262
Pickle, Catherine S., 216
Pietilainen, Olli, 196, 278
Pietrzak, Stefan J., 197
Pimentel, Monica, 148
Pinello, Luca, 198, 220, 305
Pinglay, Sudarshan, 219
Pintacuda, Greta, 106
Pipes, Lenore, 11
Pique-Regi, Roger, 86, 199, 225
Pirinen, Matti, 278
Pisupati, Rahul, 200
Pitman, Anders, 201
Plagnol, Vincent, 110
Plajzer-Frick, Ingrid, 216
Ponnaluri, V.K. Chaithanya, 247, 293
Pool, John E., 154
Poon, Yeuk G., 38
Porubsky, David, 111, 202
Posey, Jennifer E., 122
Pracana, Rodrigo, 210
Prelot, Laurie, 203
Prendergast, James G., 60
Press, Maximilian, 123
Price, Amanda J., 204
Prinz, Marco, 95
Pritchard, Jonathan K., 316, 317, 321
Przeworski, Molly, 240, 317
Puig, Marta, 44, 147, 151
Putzbach, William E., 30

Qi, Zongtai, 3
Qiao, Yi, 15, 26, 97, 201, 257
Qingge, Letu, 205
Qiu, Maguanyun, 125
Quach, Helene, 56
Quétier, Francis, 206
Quinlan, Aaron R., 40, 149, 240
Quinodoz, Sofia A., 176
Quintana-Murci, Lluis, 22, 56, 87, 116, 172

Rabbie, Roy, 94
Rafati, Nima, 222
Rahikkala, Elisa, 278
Rahl, Peter, 250
Rahmanian, Sorena, 138
Räikkönen, Jannikke, 207
Raj, Arjun, 1
Rajpurohit, Anandita, 295
Ramachandran, Sohini, 124
Ramakrishnan, Srividya, 13
Ramaswami, Gokul, 223, 287
Ramilowski, Jordan, 8
Ramírez, Claudia, 151
Rampino, Antonio, 192
Ramu, Avinash, 12
Rask-Andersen, Mathias, 222

Rätsch, Gunnar, 74, 203
Rauhala, Merja, 278
Raulerson, C K., 43, 193
Razmara, Ashkaun, 63
Real, Aline, 107
Redmond, Aisling, 230
Rees, Holly, 220
Reeves, Catherine, 16
Regan, Jack F., 44
Regier, Allison, 52
Rego, Fernanda O., 91
Regulski, Michael, 288
Reid, Jeffrey, 279
Reik, Wolf, 24
Reimand, Juri, 224
Reisman, Charles A., 10
Renco, Lucienne, 250
Renny-Byfield, Simon, 165
Reshef, Yakir, 305
Resztak, Justyna A., 225
Rey, Guillaume, 75
Reyes-Palomares, Armando, 34
Reymond, Alexandre, 75
Rhéaume, Éric, 92
Rhie, Arang, 143
Rhodes, K, 86
Rich, Stephen S., 227
Richards, A L., 86
Richards, William, 85
Rigaill, Guillem, 206
Ripatti, Samuli, 52
Risca, Viviana, 2
Ritchie, Marylyn D., 282
Rivera-Miranda, Teomie S., 226
Rizzu, Patrizia, 241
Robertson, Alan, 250
Robertson, Catherine C., 227
Robine, Nicolas, 248, 262
Robinson, Jacqueline, 207
Robles, Christopher R., 228
Rockweiler, Nicole B., 12
Rodriguez-Flores, Juan L., 229
Rodriguez-Nieto, Salvador, 70
Rogers, Jeffrey, 84, 145
Rojas, Alejandro, 250
Rokas, Antonis, 209, 219
Roller, Maša, 230
Romero, R, 199
Roos, Christian, 84, 145
Rose, Ian, 231
Rosenfeld, Jeffrey A., 105
Rosenfeld, Jill, 122
Ross, Holly, 283
Ross-Ibarra, Jeffrey, 165
Rostami, Mahboubeh, 232
Roth, Matt, 233
Rougvie, Ann E., 173
Rowan, Troy N., 79, 120, 180
Roy, Christian, 208
Roy, Sushmita, 32, 197
Roychowdhury, Tanmoy, 17
Royo, José Luis, 147
Rozowsky, Joel, 233
Rubin, Adam, 2
Rugg--Gunn, Peter, 24
Ruiz-Romero, Marina, 213, 234

Saag, Lehti, 167
Saarentaus, Elmo, 278
Sabater-Lleal, M, 284
Sabo, Aniko, 122
Saeb-Parsy, Kourosh, 158
Saito, Takeshi, 5
Sakamoto, Yoshitaka, 235
Sakarya, Onur, 280
Sakthikumar, S, 236
Salaberria, Jasone, 87
Saleh, Lana, 293
Salerno, William J., 16
Salichos, Leonidas, 237
Salomaa, Veikko, 278
Samaranayake, Mala, 293
Samocha, Kaitlin E., 113
Sanchez, Rocio, 92
Sandelin, Albin, 285
Sanders, Ashley D., 111, 202
Sanghvi, Rashesh, 248
Sanguinetti, Guido, 24
Sankararaman, Sriram, 14, 228, 319
Santos, Felipe, 91
Saraceno, Cody, 254
Sarropoulos, Ioannis, 238
Sartori, A, 239
Sasaki, Erika, 130
Sasaki, Takayo, 128

Sasani, Thomas A., 240
Savytska, Natalia, 70, 241
Saxena, Richa, 321
Schaid, Daniel, 282
Schatz, Michael C., 13
Schauer, Kevin L., 194
Scherer, Sandra, 26
Schierup, Mikkel H., 48, 118
Schindler, Karen, 277
Schlessinger, David, 53
Schnabel, Robert D., 79, 120, 180
Schrodi, Steven J., 282
Schultz, Heiko, 231
Schwarz, Erich M., 173
Schwickert, Jonas, 212
Scionti, Domenico, 212
Scott, Ethan, 211
Scott, Robert, 279
Seaton, Daniel, 68, 114, 242
Sebastian, Pott, 171
Sebat, Jonathan, 41
Sebby, Karl B., 243
Sebert, Sylvain, 66
Sedlazeck, Fritz, 13, 122, 149
Seelig, Georg, 144
Seetharam, Arun S., 165
Segurens, Béatrice, 206
Sekar, Shobana, 244
Seki, Masahide, 246, 301, 306
Sella, Guy, 252, 316
Semenyuk, Vladimir, 269
Sen, Arko, 245
Sendler, E, 199
Sénet, Marie-Victoire, 206
Sereewattanawoot, Sarun, 246
Serrano, Yazmin L., 255
Serras, Florenci, 65
Service, Susan K., 52
Sethupathy, Praveen, 140
Sexton, Brittany, 123, 247, 293
Shah, Minita, 248
Shahamatdar, Sahar, 124
Shaimardanov, Abusaid, 261
Shami, Adrienne, 152
Sharma, Kamal, 106
Sharp, Kevin, 142
Sheikh, Shehzad Z., 140
Sheltzer, Jason M., 249
Shen, Ning, 250
Shen, Xing-Xing, 209, 219
Shen, Yawei, 103
Shendure, Jay, 19
Shenoy, Archana, 148
Sheridan, Eamonn, 23
Shi, Huwenbo, 319
Shi, Xinghua, 51, 205
Shi, Xu, 98
Shiah, Yu-Jia, 131
Shigaki, Dustin, 33, 189, 251
Shin, Jay W., 8
Shin, Joo Heon, 5, 192, 204, 295
Shiphony, Zohar, 4
Shively, Christian A., 170
Short, Patrick J., 108
Shoura, Massa J., 173
Shuai, Shimin, 224
Siahpirani, Alireza, 197
Siddle, Hannah V., 256
Sidore, Carlo, 53
Sieber, Karsten B., 104
Siepel, Adam, 37, 78, 303
Sifrim, Alejandro, 113
Sigalova, Olga, 6
Sikka, Timo T., 21
Sikora, Martin, 56
Silvert, Martin, 56
Simons, Yuval B., 252, 316
Simpson, Jared T., 224
Singer, Samuel, 103
Singh, Tarjinder, 106
Singhal, Pankhuri, 282
Singleton, Andrew, 129
Sinnott-Armstrong, Nasa, 321
Sivley, Mike, 110
Skov, Laurits, 118
Slatcher, Richard B., 225
Sleiman, Patrick, 282
Slisz, Joanna, 208
Small, Neil, 23
Smith, Andrew D., 57
Smith, Cheryl L., 173
Smith, Erin N., 71
Smith, Jeramiah J., 253, 254, 272, 291
Smith, Leslie, 128

Snetkova, Valentina, 216
Snyder, Michael, 2
Socci, Nicholas, 103
Sochat, Vanessa, 138, 259
Sokolowski, Dustin J., 63
Solomonson, Matthew, 64
Song, Jiangping, 314
Song, Yun S., 31
Soos, Teresa, 151
Soranzo, Nicole, 45, 264
Sorensen, Peter, 110
Soria, J M., 284
Sorokin, Elena P., 266
Souto, J C., 284
Souto-Maior, Caetano, 255
Soyk, Sebastian, 13
Spector, Tim D., 172
Spellman, Paul T., 128
Spencer, Chris, 59, 110
Sridharan, Rupa, 197
Srinivasan, Kaushik, 126
Srinivasan, Suhas, 67
Stache-Crain, Birgit, 262
Stammnitz, Maximilian R., 256
Stamper, Ericca, 230
Stansifer, Katherine, 208
Stapel, Carine, 24
Stead, Lucy F., 29, 263
Steele, Linda, 18
Steely, Cody J., 257
Steffen, Kristan, 280
Stegle, Oliver, 24, 45, 68, 74, 114, 242, 309
Stein, Jason L., 287
Stein, Joshua C., 58, 258
Stein, Lincoln D., 224
Steinsaltz, David, 59, 142
Stenberg, Tarja, 321
Stephens, Matthew, 125
Stepniak, Karolina, 69, 294
Stevens, Beth, 106
Stevens, Michael W., 107
Stevenson, David M., 61
Stewart, Fiona J., 247
Stitziel, Nathan O., 16, 52
Stoiber, Marcus, 105
Stolle, Eckart, 210
Stone, Matthew R., 64
Stoneking, Mark, 116
Strattan, J Seth, 90, 138, 153, 259
Strober, Benjamin J., 260
Stuart, Sarah, 148
Subramanian, Sai, 233
Sud, Paul, 90, 153
Suderman, Matthew, 292
Sudre, Isabelle, 206
Sulovari, Arvis, 202
Sumner, Christine, 247
Sun, Hanxiao, 55
Sun, Luo, 123, 247
Sun, Zhiyi, 293
Sundaram, Laksshman, 84
Surti, Urvashi, 143
Sutton, Vernon R., 122
Suvakov, Milovan, 244
Suvisaari, Jaana, 278
Suzuki, Ayako, 235, 246, 301
Suzuki, Harukazu, 8
Suzuki, Yutaka, 235, 246, 276, 301, 306
Svetlichnyy, Dmitry, 261
Swaffer, Matthew P., 4
Swanson, Benjamin J., 262
Switzenberg, Jessica, 82
Swofford, R, 236
Symer, David E., 262
Szczerbinska, Iwona, 81

Tachmazidou, Ioanna, 279
Taeger, Joachim, 134
Tajinda, Katsunori, 5
Talkowski, Michael E., 64
Tambets, Kristiina, 167
Tan, Patrick, 74
Tan, Siew Z., 211
Tanaka, Forrest, 90, 153
Tanigawa, Yosuke, 275
Tanner, Georgette, 29, 263
Tao, Ran, 5, 204
Tao, Shu, 18
Tarca, A L., 199
Tardaguila, Manuel, 264
Tardif, Jean-Claude, 92
Taylor Tavares, Ana Lisa, 113
Taylor, James, 265

Taylor-Weiner, Amaro, 124
Telis, Natalie, 179, 266
Tello-Ruiz, Marcela K., 258, 267, 268
Tetikol, Serhat, 269
Thanopoulou, K, 239
Thayer, Matthew J., 128
Thistlethwaite, William, 233
Thomas, R, 236
Thompson, Ammon, 270
Thompson, Lorin A., 250
Thompson, Reid F., 72
Thompson, Tim, 243
Ticau, Simina, 208
Tilgner, Hagen U., 271
Timoshevskaya, Nataliya, 253, 254, 272
Timoshevskiy, Vladimir A., 253, 254, 272
Timpson, Nicholas J., 292
Tishkoff, Sarah, 31
Todd, John A., 227
Tony, Zheng, 171
Torniainen-Holm, Minna, 278
Torres Ulloa, Leslie, 185
Toyonaga, Takahiko, 140
Traglia, Michela, 273
Tran, Khoa A., 197
Tran, Stella, 216
Trehan, Aditi, 196
Trembath, Richard, 23
Trochet, Holly, 274
Trolle, Julie, 219
Truong, David M., 162
Tsai, Li-Huei, 141
Tseng, Alex M., 275
Tseng, Elizabeth, 288
Tsingene, Anthi, 187
Tsismentzoglou, Theofanis, 23
Tsuchiya, Issei, 276
Tu, Xiaoyu, 166
Tung, Jenny, 318
Turgut, Deniz, 269
Tyc, Katarzyna M., 277
Tyndale, Selene, 245

Udani, Rupa, 169
Ukaigwe, Joy, 5
Urpa, Lea, 278
Usher, Christina, 121
Uuskula-Reimand, Liis, 224

Vaccarino, Flora, 25
Vadlamudi, S, 193
Vaisvila, Romualdas, 293
Valdes Franco, Jose A., 166
Valentine, Erin, 250
Vallier, Ludovic, 68
Van Allen, Eliezer, 124
Van Buren, Peter, 156, 288
van der Knaap, Esther, 13
Van Eck, Joyce, 13
van Heel, David A., 23
Van Hout, Cristopher, 279
Van Keuren-Jensen, Kendall, 129
Vangala, Pranitha, 176
Varn, Frederick S., 29
Vasek, Michael J., 3
Velinder, Matt, 80, 290
Venn, Oliver, 280
Venner, Eric, 281
Ventura, Mario, 202
Venturelli, Ophelia S., 61
Verdu, Paul, 56
Verhaak, Roel G., 29
Verma, Anurag, 282
Verma, Shefali S., 282
Verta, Jukka-Pekka, 212
Vich Vila, Arnau, 323
Vieira, Päivi, 278
Vilarino, Lorena B., 187
Vilella-Figuerola, Alba, 151
Vilhjalmsson, Bjarni, 48
Villar, Diego, 230
Villatoro, Sergi, 147, 151
Ville, Yves, 206
Vincent, Elizabeth, 50, 283
Viñuela, A, 284
Visel, Axel, 216
Visscher, Peter M., 60
Vitting-Seerup, Kristoffer, 285
Vizcaya-Molina, Elena, 65
Vizuete, Clara, 151
Volkova, Nadezda, 286
Volpert, Olga, 300

Voss, S. Randal, 253
Vuckovic, Dragana, 264, 322

Wabitsch, M, 193
Wadi, Lina, 224
Wagner, Justin, 183
Wahba, Lamia, 173
Wakabayashi, Mark T., 18
Walker, Kimberly, 281
Walker, Rebecca L., 287
Wallace, Owen, 250
Wang, Bo, 58, 258, 288
Wang, Daifeng, 98
Wang, Hai, 296
Wang, Harold, 64
Wang, Liwen, 122
Wang, Liya, 156, 267, 268, 288, 289
Wang, Muhua, 212
Wang, Tracy, 256
Wang, Xiaofei, 289
Wang, Xingang, 13
Wang, Yanyan, 186
Wang, Yin-Ying, 314
Wangler, Michael F., 122
Ward, Alistair, 80, 201, 290
Ward, M C., 86
Warden, Charles D., 18
Ware, Doreen, 58, 146, 156, 182, 258, 267, 268, 288, 289
Warland, Shane F., 171
Warrell, Jonathan, 98, 233, 237
Washburn, Jacob, 296
Waterbury, Courtney K., 254, 291
Watkins, Sarah H., 292
Watson, Caroline J., 38
Wayne, Robert, 207
Weale, Mike, 110
Weersma, Rinse, 323
Wei, Sharon, 58, 258
Weinberger, Daniel R., 5, 192, 204, 295
Weiss, Lauren A., 273
Welm, Alana, 26
Welm, Bryan E., 26
Wen, Jia, 51
Wen, X, 86
Weng, Chunhua, 282
Westhead, David R., 263
Westphall, Michael S., 194
Whelan, Christopher, 121
Wieder, Cecilia, 127
Wijmenga, Cisca, 323
Wilczynski, Sharon, 18
Wilkens, Lynne, 55
Wilkinson, Michael N., 3
Williams, Brian, 298
Williams, Courtney, 295
Williams, Louise, 247, 293
Wilson, Michael D., 63, 224
Winckler, Wendy, 191
Wojtas, Bartosz, 69, 294
Wold, Barbara, 298
Wolters, John F., 219
Won, Hyejung, 98
Wong, Emily S., 211
Wong, Wing H., 38
Wood, Jo, 143
Wray, Naomi R., 60
Wright, Carrie, 295
Wright, John, 23
Wrightsman, Travis, 296
Wu, Catherine J., 109
Wu, Tsung-Jung, 281
Wu, Yingnian, 310
Wu, Yue, 319
Wurm, Yannick, 210, 297
Wyman, Dana, 138, 298

Xavier, Joao, 320
Xi, Wang, 33, 299
Xia, Wei, 204
Xiang, Yunlong, 24
Xiao, Weihong, 262
Xie, Bingqing, 300
Xie, Wei, 24
Xie, Zhijie, 310
Xing, Jinchuan, 277
Xing, Yi, 310
Xiong, Qing, 314
Xiong, Zhi, 155
Xu, Liu, 301
Xu, Qi, 302
Xu, Y, 199
Xu, Yueyuan, 312

Xue, Alexander T., 303

Yagi, Ken, 8
Yampara-Iquise, Helen, 79
Yang, Fengtang, 256
Yang, Hongbo, 312
Yang, Qi, 10
Yang, Xiaofei, 304
Yao, Qiuming, 305
Yasuzawa, Kayoko, 8
Ye, Bin, 279
Yerges-Armstrong, Laura, 279
Yi, Victoria, 281
Ying, Yi, 310
Yip, Chi Wai, 8
Yoav, Gilad, 171
Yokoyama, Toshiyuki T., 306
Yonemitsu, Marisa, 168
Yonova-Doing, Ekaterina, 45
Yosef, Nir, 7
Yoshimura, Jun, 173
Yoshitaka, Sakamoto, 306
Yotova, Vania, 318
Young Greenwald, William W., 71
Young, Adam, 115
Young, Alice, 143
Young, Andrew L., 38
Young, J Peter W., 48
Yousri, Noha A., 307
Yu, Jiyang, 186
Yuan, Guo-Cheng, 198
Yue, Feng, 312

Zack, Travis, 124
Zaidan, Nur Zafirah, 197
Zaitlen, Noah, 159, 308
Zamanian, Jennifer, 90, 153
Zaugg, Judith, 6, 34
Zaurín, Roser, 147
Zazhytska, Mariana, 75
Zerbino, Daniel R., 309
Zhang, Hanrui, 119
Zhang, Lei, 123, 247
Zhang, Miao, 266
Zhang, Qiangge, 73
Zhang, Shihua, 304
Zhang, Shilu, 32
Zhang, Yu, 104
Zhang, Zijie, 125
Zhang, Zijun, 310
Zhao, Nanxiang (Samuel), 311
Zhao, Shuhong, 312
Zhao, Xuefang, 64
Zhao, Yilin, 313
Zhao, Yixin, 37
Zhao, Yunxia, 312
Zhao, Zhongming, 137, 190, 314
Zheng, Deyou, 313
Zheng, Xianing, 152
Zhernakova, Alexandra, 323
Zhong, Silin, 166
Zhou, Shuigeng, 198
Zhou, Xiaofan, 219
Zhou, Yiqi, 148
Zhu, Binhai, 205
Zhu, Helen, 224
Zhu, Liqin, 186
Zilioli, Samuele, 225
Zimin, Aleksey V., 315
Zinovyev, Andrei, 198
Zitting, Kirsi-Marja, 321
Ziyatdinov, A, 284
Zody, Michael C., 16, 248
Zook, Justin M., 183
Zuber, Verena, 309
Zucker, Mark, 262

Arjun Raj

University of Pennsylvania, Dept. of Bioengineering, Philadelphia, PA

Cancer is a disease that originates from single cells, and the treatment of cancer also is a problem of single cells: anti-cancer therapies can often kill the vast majority of tumor cells but a few rare cells remain and grow despite treatment. Often, it is thought that the underlying basis for the behavior of these rare cells is a genetic difference. However, we and others have shown that non-genetic differences may be a key driver of rare, drug resistant cells, yet the precise molecular nature of these differences often remains mysterious. We here describe the development of a cellular “time machine” that allows us to link the ultimate cellular fate to the initial cellular state on a single cell basis, thus revealing markers for pre-resistant cells in the population. Further, we use genetic screening technologies to elucidate the pathways that control the formation of these rare cells and discuss their therapeutic implications.

# DECODING REGULATORY DNA SEQUENCE AND REGULATORY NETWORKS IN KERATINOCYTE DIFFERENTIATION

Daniel S Kim[1], Viviana Risca[2], Arwa Kathiria[3], Adam Rubin[4], James Chappell[3], Howard Chang[4], Michael Snyder[3], Will Greenleaf[3], Paul Khavari[4], Anshul Kundaje[3,5]

[1]Stanford University, Biomedical Data Sciences, Stanford, CA, [2]The Rockefeller University, New York, NY, [3]Stanford University, Genetics, Stanford, CA, [4]Stanford University, Dermatology, Stanford, CA, [5]Stanford University, Computer Science, Stanford, CA

Cell state transitions involve cascades of transcription factors (TFs) driving dynamic chromatin landscapes to produce new transcriptional programs. Epidermal differentiation, in which a self-renewing progenitor keratinocyte becomes a terminally differentiated keratinocyte, is a robust system that can address fundamental questions about the dynamic combinatorial logic of regulation. To answer these questions, genomic profiling of chromatin state (ATAC-seq and ChIP-seq on H3K27ac, H3K4me1, and H3K27me3) and transcriptional state (3' RNA-seq) was captured at 12 hour intervals across 6 days of in vitro differentiation of primary keratinocytes. We inferred epigenetic and transcriptional trajectories across time to elucidate dynamically coordinated modules of regulatory elements and genes. We then used convolutional neural networks (CNNs) to learn predictive DNA sequence drivers of chromatin dynamics. To extract motifs and combinatorial motif sets (grammars) from the CNNs, we used backpropagation-based methods to derive dynamic nucleotide-level importance scores in regulatory elements across time. These importance scores were used in conjunction with epigenetic trajectories and a motif database to extract grammars that are predictive of each trajectory. The grammars are enriched for motifs of known TFs in skin. Using the transcriptional trajectories, we infer upstream TFs that could bind these grammars, and we also determine gene sets that are likely controlled by each grammar. The inferred TFs match closely with known master TFs in skin (including TP63, ZNF750, GRHL3, JUNB, and CEBPA/B), and the gene sets are functionally enriched, suggesting that grammars coordinate activation of cellular functions during differentiation. Notably, for any specific cellular function, multiple grammars are involved, suggesting a highly redundant and interconnected logic that produces a precise cascade of gene module activations. We also find that disease heritability for skin-specific conditions is enriched in grammars. We validate these findings by combinatorial TF knockdown followed by ATAC-seq to confirm additive and synergistic effects on accessibility in grammar instances. To validate effects on activity, Massively Parallel Reporter Assay experiments are in progress, using endogenous and motif-mutated grammar instances. In summary, we implicate a landscape of TFs that act on the dynamic skin epigenome, we characterize how they interact on DNA sequence to control gene modules with phenotypic relevance, and finally we describe a dynamic core regulatory circuitry governing skin differentiation. This analysis framework provides a generalizable approach to dissecting dynamic maps of combinatorial regulation encoded in DNA sequence driving cell state changes.

# JOINT ASSAY OF SINGLE CELL RNA-SEQ AND TRANSCRIPTION FACTOR BINDING

Arnav Moudgil[1,2,3], Michael N Wilkinson[1,2], Xuhua Chen[1,2], June He[1,2], Alex J Cammack[4], Michael J Vasek[1,6], Tomas Lagunas[1,6], Zongtai Qi[1,2], Samantha A Morris[1,5], Joseph D Dougherty[1,6], Robi D Mitra[1,2]

[1]Washington University in St. Louis School of Medicine, Genetics, St. Louis, MO, [2]Washington University in St. Louis School of Medicine, Edison Family Center for Genome Sciences and Systems Biology, St. Louis, MO, [3]Washington University in St. Louis School of Medicine, Medical Scientist Training Program, St. Louis, MO, [4]Washington University in St. Louis School of Medicine, Neurology, St. Louis, MO, [5]Washington University in St. Louis School of Medicine, Developmental Biology, St. Louis, MO, [6]Washington University in St. Louis School of Medicine, Psychiatry, St. Louis, MO

Single cell RNA-seq (scRNA-seq) identifies cell types in heterogeneous mixtures and complex tissues but cannot explain the underlying regulation driving specific cell states. To better understand mechanism, multi-modal single cell techniques have emerged to connect transcriptome to other genomic data, such as genotype, CpG methylation, and chromatin accessibility. There is no method, however, that can link scRNA-seq to transcription factor (TF) binding in those same cells. Here we report the development of self-reporting transposons (SRTs) and their subsequent use in single cell calling cards (scCC), a novel assay for the simultaneous identification of transcriptome and TF binding in single cells. We show that the genomic locations of SRTs can be mapped using either cellular RNA or DNA, but transposon recovery is more efficient with RNA. Next, we use the *piggyBac* transposase, which has a strong affinity for the bromodomain protein BRD4, to identify BRD4 binding sites solely from SRT localization. We then present a method to recover SRTs from scRNA libraries (scCC), which leads to concomitant labeling of cell types and identification of TF binding sites within those populations. Using scCC, we map BRD4 binding *in vitro* and *in vivo* in the mouse cortex. Finally, we show that fusing *piggyBac* to the TF SP1 redirects insertions to SP1 binding sites, thus demonstrating that this method can be used to study potentially any TF *in situ*.

# LONG-RANGE SINGLE-MOLECULE MAPPING REVEALS THE MULTIKILOBASE-SCALE ORGANIZATION OF ACCESSIBLE CHROMATIN

Georgi K Marinov[1], Zohar Shiphony[1], Matthew P Swaffer[2], Anshul Kundaje[1,3], William J Greenleaf[1,4,5]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]Stanford University, Department of Biology, Stanford, CA, [3]Stanford University, Department of Computer Science, Stanford, CA, [4]Stanford University, Department of Applied Physics, Stanford, CA, [5]Chan Zuckerberg Biohub, Chan Zuckerberg Biohub, San Francisco, CA

Gene regulation in most eukaryotes is carried out through the action of sequence-specific regulatory factors at specific locations in the genome. The association of these proteins with regulatory elements is accompanied by reduced nucleosomal occupancy, and, accordingly, higher sensitivity to enzymatic digestion of DNA. This property, the basis of assays such as DNAse-seq and ATAC-seq, has long been a highly useful proxy for the identification of regulatory elements, such as promoters, enhancers and insulators, but important fundamental questions are left unanswered by such approaches. Regulatory elements are often located at significant genomic distances from each other, yet they act in concert to influence gene expression. However, enzymatic digestion techniques followed by short-read sequencing technologies can only measure the state of each such element individually. The state of chromatin at medium-range scales has until now been impossible to characterize, and the extent and patterns of coordination of regulatory activity along the genome are not understood. We address these questions by developing a novel method for mapping chromatin accessibility on the multikilobase scale. Our approach (SMAC-seq, or **S**ingle- **M**olecule long-read **A**ccessible **C**hromatin mapping **seq**uencing assay) is based on using a combination of treating intact chromatin with 5-methylcytosine (5mC) and $N^6$-methyladenosine ($m^6A$) methyltransferases followed by a nanopore sequencing-based readout of methylated bases and the corresponding accessibility state of individual DNA molecules. It allows the assessment of patterns of coaccessibility between distally located regulatory elements, the quantification of the distribution of chromatin states within a population of individual chromatin molecules at varying length scales, and the measurement of chromatin accessibility in previously ``invisible'' to short reads repetitive regions of the genome. It also provides a high-resolution strand-specific view of protection and accessibility patterns associated with protein occupancy on DNA. We present the insights derived from its application to the genomes of the yeast *Saccharomyces cerevisiae* and the fly *Drosophila melanogaster*, and also discuss the future scope and prospects of single-molecule long read-based approaches to the study of the physical genome.

# CELL TYPE-SPECIFIC GENETIC REGULATION OF EXPRESSION IN THE GRANULE CELL LAYER OF THE HUMAN DENTATE GYRUS

Andrew E Jaffe[1,2], Daniel J Hoeppner[3], Takeshi Saito[3], Blanpain Lou[1], Joy Ukaigwe[1], Emily E Burke[1], Ran Tao[1], Katsunori Tajinda[3], Amy Deep-Soboslay[1], Joo Heon Shin[1], Joel E Kleinman[1,4], Daniel R Weinberger[1,4], Mitsuyuki Matsumoto[3], Thomas M Hyde[1,4]

[1]Lieber Institute for Brain Development, Translational Sciences, Baltimore, MD, [2]Johns Hopkins Bloomberg School of Public Health, Mental Health, Baltimore, MD, [3]Astellas Research Institute, Neuroscience, San Diego, CA, [4]Johns Hopkins School of Medicine, Psychiatry and Behavioral Sciences, Baltimore, MD

Extensive effort has been spent over the past ten years to more fully characterize the human brain transcriptome within and across cell types, and to better understand changes in RNA expression associated with brain development and aging, developmental or psychiatric brain disorders, and local genetic variation. Here we survey gene expression in the two extremes from postmortem human brain: 1) profiling bulk tissue from hundreds of subjects and 2) hundreds of thousands of cells/nuclei from few individuals by deeply profiling gene expression from individual cell populations, by using laser capture microdissection (LCM). We performed LCM followed by RNA-seq (LCM-seq) to profile the transcriptional landscape of the granule cell layer of the dentate gyrus in human hippocampus, and contrasted to sequencing data from bulk tissue. This LCM-seq approach identified widespread cell type-specific aging and genetic effects in the DG-GCL that were either missing or directionally discordant in corresponding bulk hippocampus RNA-seq data from largely the same subjects. We identified 1337 genes with expression that only associated with age across the lifespan in the DG-GCL that were enriched for diverse neuronal processes. We further identified ~9 million SNP-feature eQTL pairs in the DG-GCL, of which 15% were not even marginally significant ($p > 0.05$) in bulk hippocampus. Using these eQTL maps, we identified novel schizophrenia-associated genes and their features that were completely missed in bulk brain tissue, including associations to GRM3 and CACNA1C which are classic therapeutic targets for this debilitating disorder. We lastly found a small number of genes differentially expressed in the DG-GCL in patients with schizophrenia, bipolar disorder or major depression compared to neurotypical individuals that were largely missed in bulk tissue. These results together highlight the importance, and biological resolution, of exploring cell type-specific gene expression levels using targeted sampling strategies like LCM, and can complement the growing sets of homogenate and single cell/nuclei datasets being generated on human brain tissue.

# ROBUSTNESS OF GENE EXPRESSION TO STOCHASTIC, ENVIRONMENTAL AND GENETIC VARIATION IS ENCODED IN THE GENOME

Olga Sigalova[1], Mattia Forneris[1], Katharina Bender[1], Judith Zaugg*[2], Eileen E Furlong*[1]

[1]European Molecular Biology Laboratory, Genome Biology, Heidelberg, Germany, [2]European Molecular Biology Laboratory, Structural and Computational Biology, Heidelberg, Germany

Gene expression occurs in an intrinsically noisy environment. A population of genetically identical cells within a controlled environment experiences intrinsic noise from the stochasticity of biochemical reactions as well as extrinsic noise due to differences in local cellular environments and cell state. Yet, living systems have a remarkable capacity to give rise to robust and highly reproducible phenotypes. A striking example is embryonic development, where developmental programs seem to be buffered to withstand variation in their genetic and environmental conditions without exhibiting major changes to embryonic development – a phenomenon that Waddington coined canalization (Waddington 1942). However, the mechanisms underlying such robustness to environmental and genetic variation remains poorly understood.

Here, we investigate the mechanism(s) underlying robustness and variation in gene expression during embryonic development. For this, we studied expression variation across across genotypes, in a population of developing embryos sampled at three phenotypically different stages of embryogenesis. To dissect the regulatory mechanisms that best discriminate between genes with high and low expression variation, we devised random forest models using features associated with diverse properties of promoters, enhancers and 3'UTRs (totalling 1205 features). Overall, the model performed well and predicted expression variation with an R2 of 0.45. The most important features for this prediction are related to promoter architecture. Expression variation is generally lower for genes with essential functions. Low-variable genes are characterized by high regulatory complexity at their TSSs, in combination with other features including RNA polymerase II pausing, enhancer complexity and post-transcriptional regulation. Interestingly, these properties identified from bulk data have previously been associated with buffering intrinsic noise in single-cell studies. This suggests that robustness of gene expression is encoded in the DNA. The model also predicts new sequence features potentially involved in buffering noise, which we are currently testing in single cells.

Taken together, our results indicate that regulatory mechanisms that have evolved to buffer intrinsic noise are the same as (or highly correlated with) mechanisms that buffer sources of extrinsic noise, including genetic and environmental variation. This provides an important link between intrinsic gene regulatory mechanisms and robustness within developmental programs.

# MASSIVELY PARALLEL CHARACTERIZATION OF REGULATORY DYNAMICS DURING NEURAL INDUCTION

Anat Kreimer[1,2], Fumitaka Inoue[2], Tal Ashuach[1], Nadav Ahituv[2], Nir Yosef[1]

[1]University of California, Berkeley, Electrical Engineering and Computer Sciences, Berkeley, CA, [2]University of California San Francisco, Bioengineering and Therapeutic Sciences, San Francisco, CA

The temporal interplay between gene regulation and gene expression during cell differentiation remains largely unknown. Using neural induction as a model, we set out to decipher these dynamics. We performed RNA-seq, ChIP-seq (H3K27ac, H3K27me3) and ATAC-seq on human embryonic stem cells at seven early neural differentiation time points (0-72 hours). We found that DNA accessibility precedes H3K27ac, which is then followed by gene expression changes. Using massively parallel reporter assays (~2,500 sequences tested at all seven time-points), we further show that temporal enhancers correlate with H3K27ac. Development of a prioritization method that incorporated all genomic data identified key transcription factors (TFs) involved in temporal function, several of which were functionally validated to be important and novel neural induction regulators. Combined, our results provide a resource of genes and regulatory elements that orchestrate neural induction and illuminate the temporal framework that is needed to obtain this differentiation.

# FUNCTIONAL ANNOTATION OF LONG NON-CODING RNAS BASED ON THEIR SPATIAL LOCALIZATION

Saumya Agrawal[1], Jordan Ramilowski[1], Chi Wai Yip[1], Chung Chau Hon[1], Kayoko Yasuzawa[1], Masayoshi Ito[1,2], Takeya Kasukawa[1], Naoto Kondo[1], Ken Yagi[1], Harukazu Suzuki[1], Josée Dostie[3], Jay W Shin[1], Michiel de Hoon[1], Piero Carninci[1]

[1]RIKEN, Division of Genomic Medicine, Center for Integrative Medical Sciences, Yokohama, Japan, [2]RIKEN, Preventive Medicine and Diagnosis Innovation Program, Wako, Japan, [3]McGill University, Department of Biochemistry, Montreal, Canada

Identification of the regulatory role of long noncoding RNAs (lncRNAs) is challenging as the information regarding their targeted genes is largely unknown. As a result, comprehensive functional annotation of lncRNAs is still lacking and therefore for a long time they were thought to be nonfunctional "junk RNA". Studies have shown that ncRNAs can act as regulators for genes that are nearby in three-dimensional space but are several kilobases away in linear genomic space. Therefore, as part of the FANTOM6 project, we have used chromosome conformation data to identify the genomic regions spatially close to the lncRNA transcribing regions in different cell types. Subsequently, we have used the information regarding spatially closed regions to identify the potential targets of the lncRNAs in different human cells. We have generated a high-resolution genomic interaction map using newly produced Hi-C data from different cell types to identify the potential interacting partners of the lncRNAs such as protein-coding genes, other lncRNAs, microRNAs or enhancers in human cells. To determine the dynamic nature of the lncRNA, we have also identified their common or cell type-specific interacting partners. Currently, we are exploring different properties such as population conservation, evolutionary conservation, RNA binding proteins, transcription factor binding sites, and specific genomic patterns of lncRNA and their interacting partners to annotate the function of the lncRNAs. Overall, we will discuss how the three-dimensional localization of lncRNAs in the nuclei plays a critical role in determining their function.

LEVERAGING ALLELEIC HETEGENEITY IN ASSOCIATION STUDIES

Eleazar Eskin[1], Farhad Hormozdiari[2], Jong Wha Joo[3]

[1]University of California, Los Angeles, Computational Medicine, Los Angeles, CA, [2]Harvard, Epidemiology, Boston, MA, [3]Dongguk University-Seoul, Computer Science and Engineering, Seoul, South Korea

Over 100,000 variants have been implicated in human traits through genome-wide association studies (GWAS). Virtually all reported GWAS identify associations by measuring the correlation between a single variant and a phenotype of interest. Recently, several studies reported that at many risk loci, there may exist multiple causal variants, a phenomenon referred to as allelic heterogeneity. Naturally, for a locus with multiple causal variants with small effect sizes, the standard association test is underpowered to detect the associations. Alternatively, an approach considering effects of multiple variants simultaneously may increase statistical power. Counterintuitively, most approaches that consider multiple variants in association studies find fewer associations than the single SNP association test. This is due to the fact that most multiple variant methods assume a structure of allelic heterogeneity which is very different from what is observed in genetic studies. In this work, we propose a new statistical method, Model-based Association test Reflecting causal Status (MARS), that tries to find an association between variants in risk loci and a phenotype, considering multiple variants at each locus. One of the main advantages of MARS is that it only requires the existing summary statistics to detect associated risk loci. Thus, MARS is applicable to any association study with summary statistics, even though individual level data is not available for the study. Utilizing extensive simulated data sets, we show that MARS increases the power of detecting true associated risk loci compared to previous approaches that consider multiple variants, while robustly controls the type I error. Applied to data of 44 tissues provided by the Genotype- Tissue Expression (GTEx) consortium, we show that MARS identifies more eGenes compared to previous approaches in most of the tissues; e.g. MARS identified 16% more eGenes than the ones reported by the GTEx consortium. Moreover, applied to Northern Finland Birth Cohort (NFBC) data, we demonstrate that MARS effectively identifies association loci with improved power (56% of more loci found by MARS) in GWAS studies compared to the standard association test.

# HUMAN GENETICS OF COMPLEX RETINAL ANATOMY REVEALED BY LARGE SCALE IMAGING

Hannah Currant[1], Tomas Fitzgerald[1], Pirro G Hysi[2], Charles A Reisman[3], Qi Yang[3], Christopher J Hammond[2], Peng T Khaw[4], Paul J Foster[4], Praveen Patel[4], Anthony P Khawaja[4], Ewan Birney[2]

[1]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute, Cambridge, United Kingdom, [2]Kings College London, Diabetes, London, United Kingdom, [3]Topcon Healthcare Solutions, Research and Development, Oakland, NJ, [4]University College London, NIHR Biomedical Research Centre at Moorfields Eye Hospital Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom

Human anatomy has been used to diagnose and measure aspects of humans from the birth of medical science. The advent of imaging technologies has both broadened the range of anatomy that can be measured and provided more detail for many human organs, often at a cellular scale. Imaging technology has been used to provide single measurements of humans in genetic studies, but more sophisticated use of imaging for human genetics has been hindered due to an excess of diversity in measurement procedures, the need for human based annotation of the resulting images, and a lack of data availability at scale.

The UK Biobank Eye and Vision Consortium provided the resources to image ~65,000 individuals in the last phase of the UK Biobank phenotype collection, using both fundus imaging and Optical Coherence Tomography (OCT). OCT is able to measure the reflectivity of tissue layers in the retina, and able to distinguish specific single cell layers in the living retina. The OCT technique was used on both left and right eyes and focused on the macula region, the central area of the retina responsible for highest acuity vision. We used automated OCT image analysis software to derive a rich set of phenotypes for each individual. These phenotypes can be studied individually, but as the dimensionality of the measurement increases, single phenotype associations become both more complex to interpret and the resulting tests become less powerful. We have therefore used dimensionality reduction to provide a more compact representation of the variation in the retinal anatomy. In the first study of this scale on the human eye (and one of the largest studies of any human organ) we are able to robustly associate components of this variation in anatomy to human genetics and other epidemiological factors. Interestingly some dimensions of variation show strong genetic effects whereas others do not.

Overall, the resulting genetic associations have overlap with many other eye phenotypes and eye related diseases, as well as other traits. This study immediately suggests a number of clinically relevant hypotheses in human eye biology and is a model for human organ intermediate phenotyping at scale.

# A RAPID PHYLOGENY-BASED METHOD FOR ACCURATE COMMUNITY PROFILING OF LARGE-SCALE METAGENOMIC DATASETS

Lenore Pipes, Rasmus Nielsen

University of California-Berkeley, Department of Integrative Biology, Berkeley, CA

In the past ten years, metagenomics has promised to enable the characterization of the full diversity of unknown environments. This has led to the publication of over 80 taxonomic classification methods with the majority being either alignment-based or kmer-based approaches. These approaches use simple sequence similarity metrics to identify the origin of metagenomic sequences, and have been favored over phylogenetic-based methods as a trade-off between speed and precision to handle the rapidly growing reference databases of genome markers and the increasingly large amounts of sequencing data. Furthermore, even though a more informative type of classification is preferred, the current implementations of phylogenetic-based methods are impractical for the sheer volume of data routinely desired as input. We have developed the first phylogeny-based taxonomic classification method designed to truly enable the use of modern-day reference databases and high throughput sequencing data. Our approach relies on a multiple sequence alignment and tree of the reference sequences which is recursively split into subtrees and re-aligned until the new multiple sequence alignments achieve a heuristic quality threshold. State probabilities for all internal nodes for each subtree are computed using a maximum likelihood approach for ancestral sequence reconstruction, and can be saved as a reference data structure for future analyses. We then find the most likely subtree(s) that the query sequence originated from using the user's choice of a functional integration of either a short-read mapper or BLAST. Finally, a Needleman-Wunsch alignment is performed on the most similar sequences in the reference database and probability scores are computed by summing across the alignment for all of the nodes in the subtree. Comparisons with competing phylogeny-based methods show speed-ups over 1000x with a vastly reduced memory footprint while maintaining the same or better precision. Additionally, we use both simulated and biological data to show that our method outperforms the most popular alignment-based and kmer-based tools with comparable speed. Finally, our simulations with left out reference sequences from the database demonstrate that our method can recover the correct taxonomic path even when a species is unknown, a realistic issue with reference databases and a severe limitation of current methods. Our method represents a substantial improvement over current classification approaches by correctly assigning to the putative origin in the tree thus reducing the rate of misclassifications.

# WHOLE-BODY MAPS OF POSTZYGOTIC MUTATIONS IN HUMANS

Nicole B Rockweiler[1], Avinash Ramu[1], Ellen Gelfand[2], Casey Drubin[1], Ni Huang[1], Malachi Griffith[1,3], Kristin Ardlie[2], The GTEx Consortium[2], Donald F Conrad[1,4]

[1]Washington University School of Medicine, Department of Genetics, St. Louis, MO, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [4]Division of Genetics, ONPRC, Oregon Health & Science University, Portland, OR

**Background**: A key but largely unanswered question in biology is: how do normal cells accumulate postzygotic mutations to transform into diseased cells?
**Methods**: Here, we begin to answer this question by estimating the burden of postzygotic mutations and the mutational processes in normal tissues. We developed a method for calling postzygotic mutations from single sample RNA-seq data that first uses a position-specific error model and then applies a more sensitive multisample step to refine the list of mutations. This framework can detect postzygotic mutations with variant allele frequencies as low as 0.01%—approximately 2 orders of magnitude below the average sequencing error rate. We ran the algorithm on ~17,000 samples across 53 diverse tissues from 964 donors from the NIH's Genotype-Tissue Expression (GTEx) project.
**Results**: We detected ~130,000 postzygotic mutations with a median of 0.1 mutations per expressed Mb per tissue biopsy. Amplicon-seq validation on genomic DNA yielded a validation rate of 88%. Using power simulations, we estimate a biopsy contains ~50 expressed postzygotic mutations. Tissues showed remarkable variation in mutation burden and mutational signatures. Mutational signatures were strand specific, associated with donor age, or characteristic of known mutagenic exposures in biologically relevant tissues, e.g., high rates of ultraviolet light-associated CC>TT and C>T mutations in exposed skin samples. In order to study the developmental origin of the mutations, we constructed a model of human development relating all 53 tissues through a multifurcating tree. After mapping the postzygotic mutation callset onto this tree, we observed changes in mutation processes throughout development, and estimate >70% of multi-tissue postzygotic mutations occur before gastrulation. Finally, we measured the functional impact of the mutations by assessing signatures of selection and clinical significance. We conclude that, while most observed mutations are selectively neutral, the seeds of age-related disease are observable in healthy tissues, as 4% of the donors carry mutations that would be considered pathogenic in a clinical setting.
**Significance**: This work provides an innovative framework for calling postzygotic mutations as well as the most extensive catalog of postzygotic mutations in normal human tissues and thus, expands the foundation for understanding mutational processes in disease.

# 100 GENOMES IN 100 DAYS: THE STRUCTURAL VARIANT LANDSCAPE IN TOMATO GENOMES

Michael C Schatz[1,2], Sebastian Soyk[1], Xingang Wang[1], Michael Alonge[2], Zachary Lemmon[1], Srividya Ramakrishnan[2], Sara Goodwin[1], Fritz Sedlazeck[3], W. Richard McCombie[1], Esther van der Knaap[4], Joyce Van Eck[5], Zachary Lippman[1,6]

[1]Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [3]Baylor College of Medicine, Human Genome Sequencing, Houston, TX, [4]University of Georgia, Department of Horticulture, Athens, GA, [5]Boyce Thompson Institute, Plant Breeding and Genetics Section, Ithaca, NY, [6]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Structural variations (SVs) are known to be major drivers of quantitative variation in many species. Short-read sequencing has proven valuable for SNP discovery, but lacks power for more complex SVs. Long-read sequencing has the potential to illuminate this large reservoir of hidden variation, but has previously been too costly and time-intensive to apply at scale. Addressing this critical gap, we have optimized the Oxford Nanopore PromethION to produce read lengths averaging over 30kb and yields up to 130Gbp per flowcell, making it fast and affordable to sequence large numbers of samples.

We have used this technology to sequence and understand SVs in nearly 100 varieties of tomato. Here we discuss 3 important questions: (1) How to select the samples for sequencing that will capture the largest amount of variation; (2) How to rapidly sequence, basecall, and manage the data for 100 genomes; and (3) How to identify SVs across the population using a combination of de novo assembly and read-mapping approaches. Each sample shows ten to forty thousand variants, including many within gene sequences that were not detectable using short-reads. Furthermore, for some variants we have developed CRISPR-based mutants showing novel fruit morphologies and inflorescence branching phenotypes. We believe this approach will become the new gold standard for SV analysis in all species.

# DISSECTING THE GENETIC ARCHITECTURE OF COMPLEX TRAITS THROUGH LOCAL POLYGENICITY USING SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES

Ruth D Johnson[1], Kathryn Burch[2], Kangcheng Hou[3], Bogdan Pasaniuc[4,5,6,2], Sriram Sankararaman[1,5,6,2]

[1] University of California, Los Angeles, Department of Computer Science, Los Angeles, CA, [2] University of California, Los Angeles, Bioinformatics Interdepartmental Program, Los Angeles, CA, [3]College of Computer Science and Technology, Zhejiang University, Hangzhou, China, [4]David Geffen School of Medicine, Department of Pathology and Laboratory Medicine, Los Angeles, CA, [5]David Geffen School of Medicine, Department of Human Genetics, Los Angeles, CA, [6]David Geffen School of Medicine, Department of Computational Medicine, Los Angeles, CA

A major goal of studying complex traits is investigating how many genetic variants contribute to the phenotypic variation of a trait as well as the extent to which these variants are spread throughout the genome. Genome-wide polygenicity estimates provide an estimate of the total number of causal SNPs, but this estimate does not provide insight into the distribution of these causal SNPs across the genome. Additionally, to further dissect the genetic architecture of complex traits, we need to stratify SNPs beyond the binary classification of causal and non-causal.

To address this, we estimate and leverage local polygenicity, defined as the proportion of causal SNPs (p) within a region, in combination with local heritability estimates to provide an estimate of the expected per causal SNP variance contribution within a region. This quantity not only provides information about the proportion of SNPs in a region that are causal, but also the magnitude of their effect. This allows us to uncover regions with high effect SNPs that may contribute to a trait, thus enabling a principled way to prioritize regions of the genome for further follow-up analyses.

To estimate local polygenicity, we introduce BEAVR (Bayesian EstimAtion of Variants in Regions), a probabilistic approach to estimating local polygenicity from GWAS summary statistics, while accounting for linkage disequilibrium. Here we present a fully Bayesian generative model coupled with a highly scalable inference algorithm that leverages the natural sparsity of most traits and scales approximately linearly with the number of SNPs analyzed.

We divide the genome into 6MB regions and apply our method to estimate the local polygenicity for 22 traits from the UK Biobank (N=290K), ranging from anthropomorphic, autoimmune, cardiovascular, and respiratory related traits. We find that anthropomorphic traits harbor the most regions with high local polygenicity ($p > 5\%$). Next, we compute local heritability estimates for each region and compute the expected per causal SNP variance contribution. We identify an average of 53 regions of the genome that have a 2X enrichment of expected per causal SNP variance for a single trait and an average of 35 regions with a 3X enrichment. Finally, we find that many of these enriched regions are shared between two or more traits, such as autoimmune related diseases.

**SCBAYES**: A COMPUTATIONAL METHOD TO STUDY TUMOR SUBCLONE-SPECIFIC GENE EXPRESSION AND CHROMATIN ACCESSIBILITY USING SINGLE CELL RNA SEQUENCING AND SINGLE CELL ATAC SEQUENCING IN COMBINATION OF BULK DNA SEQUENCING

Yi Qiao[1,2], Xiaomeng Huang[1,2], 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

Although critical for understanding the underlying cancer biology, the study of tumor cell phenotype and epigenetic state from bulk gene expression (RNA-seq) and chromatin accessibility (ATAC-seq) data is often challenging because tumor samples are typically mixtures of different cell populations due to normal tissue contamination, immune cell infiltration, and tumor subclonality. These cell populations may have vastly divergent transcriptomic and epigenetic characteristics (this is especially true for metastatic cancer samples, where cancer tissue and surrounding normal tissue are of different organ of origin), resulting in mixtures of signals that are often impossible to deconvolute.

Here we present ***scBayes***, a computational method that allows one to characterize the transcriptional behavior and/or epigenetic state corresponding to each tumor subclone present in a tumor sample, or multiple serial or multisite samples from a given cancer patient. This tool starts with genomically defined subclones reconstructed from bulk DNA sequencing data, e.g. using our own published ***SubcloneSeeker*** algorithm. ***scBayes*** then layers single cell gene expression (scRNA-seq) and/or chromatin accessibility (scATAC-seq) data on this genomic subclone "backbone" to assign individual cells to one of the tumor subclones or normal cell compartment, facilitating subclone-specific expression or epigenetic analysis. This allows (1) the identification of tumor expression or epigenetic signals without normal tissue interference; (2) the comparison between tumor and normal cells from the same sample; (3) the comparison of expression and epigenetic state across distinct tumor subclones; and (4) the study of transcriptional / epigenetic plasticity through the comparison of the cell states of the same genomically defined subclone across different time points or metastatic sites.
***scBayes*** achieves cell-to-subclone assignment by assessing the presence of somatic variants (CNVs or SNVs) that define cancer subclones in each individual cell from the single cell sequencing reads. Because per-cell sequencing coverage is sparse, and therefore variant dropout frequent, scBayes utilizes a Bayesian probabilistic framework to calculate the likelihood of a particular cell originating from each subclone (including the normal clone, defined as having no somatic mutations), and makes a probabilistic assignment, maximizing the utilization of the sparse single cell sequencing data for subclone assignment. We successfully applied this method to multiple longitudinal and metastatic cancer patient datasets, representing both bulk WES and WGS DNA sequencing, as well as single cell datasets collected using the 10X Chromium and Fluidigm C1 platforms. Our algorithmic approach enables comparative gene expression and pathway analysis, and open chromatin accessibility assessment, across subclones. scBayes is open source and freely available at *https://github.com/yiq/scBayes*.

# MAPPING AND CHARACTERIZATION OF STRUCTURAL VARIATION IN 17,795 DEEPLY SEQUENCED HUMAN GENOMES

Haley J Abel*[1,2], David E Larson*[1,2], Colby Chiang[2], Indraniel Das[2], Krishna L Kanchi[2], Ryan M Layer[8], Benjamin M Neale[4], William J Salerno[5], Catherine Reeves[6], Steven Buyske[9], Tara C Matise[7], Donna M Muzny[5], Michael C Zody[6], Eric S Lander[4], Susan K Dutcher[1,2], Nathan O Stitziel[2,1,3], Ira M Hall[2,1,3]

[1]Washington University, Department of Genetics, St. Louis, MO, [2]Washington University, McDonnell Genome Institute, St. Louis, MO, [3]Washington University, Department of Medicine, St. Louis, MO, [4]Broad Institute, Cambridge, MA, [5]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [6]New York Genome Center, New York, NY, [7]Rutgers University, Department of Genetics, Piscataway, NJ, [8]University of Colorado, BioFrontiers Institute, Boulder, CO, [9]Rutgers University, Statistics and Biostatistics, Piscataway, NJ, [10]NHGRI Centers for Common Disease Genomics, Bethesda, MD

A key goal of whole genome sequencing (WGS) for human genetics studies is to interrogate all forms of variation, including single nucleotide variants (SNV), small insertion/deletion (indel) variants and structural variants (SV). However, tools and resources for the study of SV have lagged behind those for smaller variants. We have used our highly-scalable software toolkit (svtools) and a cloud-based pipeline to map and characterize SV in 17,795 deeply sequenced human genomes from common disease trait mapping studies sequenced under the CCDG program, supplemented with ancestrally diverse samples from the PAGE consortium (~950 Latinos) and Simons Genome Diversity Panel (263 individuals from 142 populations). The publicly released site-frequency information comprises the largest WGS-based SV resource to date and promises to be a valuable reference for the research and medical communities. On average, individuals carry 2.9 rare SVs that alter coding regions, which affect the dosage or structure of 4.2 genes and account for 4.0-11.2% of rare high-impact coding alleles. Based on a computational model, we estimate that SVs account for 17.2% of rare alleles genome-wide whose predicted deleterious effects are equivalent to loss-of-function (LoF) coding alleles; ~90% of such SVs are non-coding deletions (mean 19.1 per genome). We report 158,991 ultra-rare SVs and show that ~2% of individuals carry ultra-rare megabase-scale SVs, nearly half of which are balanced and/or complex rearrangements. Finally, we exploit this resource to infer the dosage sensitivity of genes and non-coding elements, revealing strong trends related to regulatory element class, conservation and cell-type specificity. This work will help guide SV analysis and interpretation in the era of WGS.

# CHROMATIN ORGANIZATION MODULATES THE ORIGIN OF HERITABLE STRUCTURAL VARIATIONS IN HUMAN GENOME

Alexej Abyzov, Tanmoy Roychowdhury

Mayo Clinic, Health Sciences Research, Rochester, MN

Structural variations (SVs) in the human genome originate from different mechanisms related to DNA repair, replication errors, and retrotransposition. Our analyses of 26,927 SVs from the 1000 Genomes Project revealed differential distributions and consequences of SVs of different origin, e.g. deletions from non-allelic homologous recombination (NAHR) are more prone to disrupt chromatin organization while processed pseudogenes can create accessible chromatin. Spontaneous double stranded breaks (DSBs) are the best predictor of enrichment of NAHR deletions in open chromatin. This evidence, along with strong physical interaction of NAHR breakpoints belonging to the same deletion suggests that majority of NAHR deletions are non-meiotic i.e. originate from errors during homology directed repair (HDR) of spontaneous DSBs. In turn, the origin of the spontaneous DSBs is associated with transcription factor binding in accessible chromatin revealing the vulnerability of functional, open chromatin. The chromatin itself is enriched with repeats, particularly fixed Alu elements that provide the homology required to maintain stability via HDR. Additionally, we observed a striking difference between distributions of fixed and variable Alus across genome compartments. Through co-localization of fixed Alus and NAHR deletions in open chromatin we hypothesize that old Alu expansion had a stabilizing role on the human genome.

# GENOMIC ANALYSIS OF GERMLINE AND SOMATIC VARIATION IN HIGH-GRADE SEROUS OVARIAN CANCER

Aaron W Adamson[1], Yuan Chun Ding[1], Mihaela C Cristea[2], Sharon Wilczynski[3], Mark T Wakabayashi[4], Ernest S Han[4], Thanh H Dellinger[4], Pin S Lin[4], Amy A Hakim[4], Linda Steele[1], Charles D Warden[5], Shu Tao[5], Susan L Neuhausen[1]

[1]Beckman Research Institute of City of Hope, Department of Population Sciences, Duarte, CA, [2]City of Hope National Medical Center, Department of Medical Oncology, Duarte, CA, [3]City of Hope National Medical Center, Department of Pathology, Duarte, CA, [4]City of Hope National Medical Center, Department of Surgery, Duarte, CA, [5]Beckman Research Institute of City of Hope, Integrative Genomics Core, Department of Molecular and Cellular Biology, Duarte, CA

High-grade serous ovarian cancers (HGSCs) display a high degree of complex genetic alteration involving numerous oncogenes and tumor suppressor genes. Germline mutations in *BRCA1* and *BRCA2* are associated with approximately 15% of HGSC cases and additional rarer genes of high to moderate risk (e.g., *BRIP1*, *PALB2*) have recently been identified. The objective of this study was to identify additional genes for HGSC associated with occurrence and disease-free survival. Using a targeted capture of 577 genes involved in the response to DNA damage and the PI3K/AKT/mTOR pathways, we conducted next-generation sequencing of DNA from matched blood and tumor tissue from 71 HGSC patients. Similar to other studies of HGSC, we found that approximately one-third of tumors had loss-of-function (LOF) germline (18/71, 25.4%) or somatic (8/71, 11.3%) mutations in one or more homologous recombination (HR) genes: *BRCA1*, *BRCA2*, *CHEK2*, *MRE11A*, and *PALB2*. As expected, we also observed LOF germline mutations in other Fanconi anemia genes as well as PI3K/AKT/mTOR pathway genes. Most of the tumors harbored somatic mutations in *TP53* (65/71, 91.5%). Of interest, we identified several germline and somatic protein-truncating mutations in genes involved in apoptosis not previously associated with HGSC. To identify copy number variants, we used the OncoScan array performing genome-wide analysis of somatic copy number variation (SCNV) and loss of heterozygosity (LOH) on DNA from tumors from 61 patients; we identified focal homozygous deletions in *BRCA2*, *MAP2K4*, *PTEN*, *RB1*, *SLX4*, *STK11*, and *NF1*. From our combined sequencing and SCNV analysis approach, we found that (28/71, 39.4%) of the HGSC patients in this study harbor pathogenic mutations in HR genes and would be candidates for therapy with PARP inhibitors. We also investigated whether SCNVs were associated with disease-free survival. Using GISTIC analysis, we identified three regions of high-amplitude amplification (containing *NOTCH3*, *ZNF536*, and *PIK3R2*) that were significantly associated with an increase in cancer recurrence and a reduction in overall survival. Further exploration of these regions is planned to assess their role in chemotherapy resistance.

# PREDICTING mRNA ABUNDANCE DIRECTLY FROM GENOMIC SEQUENCE USING DEEP CONVOLUTIONAL NEURAL NETWORKS

Vikram Agarwal[1], Jay Shendure[1,2,3]

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

Algorithms that accurately predict gene structure from primary sequence alone were transformative for annotating the human genome. Can we also predict the *expression levels* of genes based solely on genome sequence? Here we sought to apply deep convolutional neural networks towards this goal. Surprisingly, a model that includes only promoter sequences and features associated with mRNA stability explains 59% and 71% of variation in steady-state mRNA levels in human and mouse, respectively. This model, which we call Xpresso, more than doubles the accuracy of alternative sequence-based models, and isolates rules as predictive as models relying on ChIP-seq data. Xpresso recapitulates genome-wide patterns of transcriptional activity and predicts the influence of enhancers, heterochromatic domains, and microRNAs. Model interpretation reveals that promoter-proximal CpG dinucleotides strongly predict transcriptional activity. Looking forward, we propose the accurate prediction of cell type-specific gene expression based solely on primary sequence as a grand challenge for the field.

# CODING AND NONCODING DNA VARIANTS UNDERLIE *TRANS*-ACTING HOTSPOTS OF GENE EXPRESSION VARIATION

Sheila Lutz, Margaret Kliebhan, Christian Brion, Frank W Albert

University of Minnesota, Department of Genetics, Cell Biology, & Development, Minneapolis, MN

DNA variants that alter gene expression underlie variation in many important traits. In humans and other species, most heritable variation in gene expression arises from *trans*-acting loci located far from their target genes, typically on different chromosomes. Despite their profound effects, most *trans*-acting variants, and the mechanisms by which they alter gene expression, are unknown. In a cross between a laboratory and a wine strain of the yeast *Saccharomyces cerevisiae*, *trans*-regulatory variation is almost exclusively caused by dozens of "hotspot" loci that each influence mRNA levels of up to thousands of genes. We do not know the causal variants that produce most of these hotspots.

Here, we identified nucleotide variants underlying three *trans*-regulatory hotspots. We developed a fine-mapping strategy based on the CRISPR/Cas9 system. First, the locus is replaced or flanked with commonly used yeast selectable marker cassettes. Next, a single guide RNA targets these cassettes to swap them along with any intervening sequence for the desired alleles. We systematically narrowed each locus to the causal variant by engineering recombinant allelic series and quantifying allele effects on the expression of a green fluorescent protein-tagged gene affected by the locus in *trans*.

In one hotspot, we identified a causal missense variant in the glucose receptor *RGT2*. This variant is present in many global yeast isolates, and its effect depended strongly on the concentration of glucose in the culture medium.

In the remaining two hotspots, we identified causal variants that both alter the expression of genes involved in fatty acid metabolism. A rare missense variant resulted in reduced activity of Oaf1, a transcription factor that activates genes involved in fatty acid β-oxidation. A globally common noncoding variant in the promoter of *OLE1* (a fatty acid desaturase) influenced the expression of distant genes in *trans* and *OLE1* expression locally, suggesting that a *cis*-acting effect on *OLE1* is the causal mechanism. Indeed, experimental manipulation of *OLE1* expression showed that altering the levels of this essential metabolic enzyme changed the expression of other genes in *trans*. The *OAF1* and *OLE1* variants interacted in a non-additive fashion, such that the more active *OAF1* allele enabled the *OLE1* promoter variant to have a greater effect on gene expression in *trans*. Two additional hotspots affect mRNA levels of genes involved in fatty acid metabolism and are centered on *FAS1* & *FAA4*, which encode additional enzymes in fatty acid metabolism.

Our results revealed remarkable diversity in the molecular basis of *trans*-regulatory variation. The three fine-mapped hotspots variants included rare and common, as well as coding and non-coding variants in genes encoding a membrane sensor, a transcription factor and an essential enzyme.

# IDENTIFYING VERY RARE GENETIC VARIANTS IN DNA MICROARRAY DATA USING A POPULATION-SPECIFIC REFERENCE PANEL FOR IMPUTATION

Maris Alver[1,2], Timo T Sikka[1,2], Tiit Nikopensius[1], Mart Kals[1], Reedik Mägi[1], Tõnu Esko[1,3]

[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]Broad Institute, Cambridge, MA

While the ascertainment of individuals affected by rare clinically actionable disorders requires next generation sequencing (NGS) for accurate calling of carrier status, availability of these resources is not always feasible for large-scale cohorts due to cost. The Estonian population-based biobank, that holds DNA microarray-derived genotype information for 52,274 individuals, has built a population-specific reference panel from NGS data (n=2,244) for imputation. We aimed to investigate whether the utilisation of a custom reference panel for imputing missing genotypes enables the ascertainment of very rare genetic variant (VRGV) carriers in microarray data.

We set to determine carrier status for 20 VRGVs linked with clinically actionable disorders that were identified in heterozygous state in the NGS data (allele counts between 1-7, n=34) from the Global Screening Array-genotyped data of 33,277 individuals. To this end, five different approaches were applied, utilising the long-range haplotype approach as well as basic (EAGLE for phasing, BEAGLE for imputation) and read-aware phasing-based (SHAPEIT2) imputation (BEAGLE). As phasing of very rare alleles is largely random, we flipped the phased allele of very rare variants in the reference panel to the other haplotype, and additionally carried out the two imputation approaches using created dummy haplotypes.

Across five methods, we identified 1064 putative VRGV carriers who were Sanger sequenced for validation. Carrier status was confirmed for 187 individuals for 16 variants, yielding the true positive rate of 18% across all methods, 41% for the basic imputation and 23% for the long-range haplotype approach. While 163 of 187 (87%) true carriers were identified via the basic imputation method, allele switching with basic imputation and long-range haplotyping facilitated to uniquely identify carriers for two singleton (n=15) and three singleton variants (n=4), respectively, not identified with any other approach, confirming the random phasing of singleton variants.

These results demonstrate that the utilisation of a population-specific reference panel for imputing missing genotypes in whole-genome microarray data enables to ascertain carriers of clinically actionable disorders within a population-specific biobank, when NGS data is unavailable. While basic imputation allows to identify VRGV carriers, modification approaches will enhance the detection of singleton variants.

# GENOME-WIDE ANALYSES OF NI-VANUATU POPULATIONS REVEAL DIVERSE ANCESTRIES AND TEMPORAL DIFFERENCES IN ADMIXTURE PATTERNS

Lara R Arauna[1], Jeremy Choin[1], Javier Mendoza-Revilla[1], Christine Harmant[1], Laure Lemee[2], Olivier Cassar[3], Antoine Gessain[3], Etienne Patin[1], Lluis Quintana-Murci[1]

[1]Human Evolutionary Genetics, Institut Pasteur, Paris, France, [2]Biomics Platform, Institut Pasteur, Paris, France, [3]Epidemiology and Physiopathology of Oncogene Viruses, Institut Pasteur, Paris, France

The Vanuatu archipelago was peopled during the last human migration at the worldwide scale ~3,000 years ago (ya), concomitantly with the expansion of Austronesian (AN) languages and Lapita assemblies. Although languages spoken in Vanuatu are of AN origin, recent evidence suggests a predominantly Papuan-like (PL) genetic ancestry in some Vanuatu islands. Recent ancient DNA studies support the arrival of PL ancestry soon after AN expansions, likely from the Bismarck Archipelago. Yet, genetic studies of modern populations from Vanuatu —the country with the highest number of languages per capita and spreading through 65 islands— remain scarce. Furthermore, the temporal peopling of this archipelago, and the sources and extent of admixture processes in the region remain uncharacterized.

Here, we generated genome-wide SNP data from more than 1,000 Ni-Vanuatu DNA samples, originating from 30 islands. Despite their high cultural diversity, our data revealed an overall degree of genetic homogeneity, suggesting a common peopling process for the archipelago. When using haplotype-based methods (ChromoPainter, fineSTRUCTURE), however, we observed fine patterns of genetic structure that mirror geography. We found a PL component that is present at overall high proportions (~80%) but that varies according to geography, revealing differences in admixture proportions across islands. Using GLOBETROTTER, we estimated the admixture events between the PL and AN ancestry components to have occurred between 2,000 and 500 ya, depending on the different islands. Notably, our analyses revealed more complex patterns of ancestry sources than previously appreciated, with different AN sources contributing to the current patterns of diversity of Ni-Vanuatu. Finally, although a high degree of relatedness was generally observed within islands, supporting the sea as a general barrier to gene flow, we found that the northern Torres Islands show similar numbers of related individuals within and between islands, supporting a history of cultural exogamy.

In conclusion, our analyses revealed that the various populations inhabiting today the Vanuatu archipelago descend from different ancestry components, which vary today in intensity across islands, that have admixed to different extents since the early peopling of the region. Moreover, the fine-scale genetic structure of the region has also been shaped by geographic barriers and by different mating patterns. Together, this study sheds new light onto the complex peopling and admixture history of this largely neglected region of the world.

# FINE-SCALE POPULATION STRUCTURE AND HISTORY IN BRITISH-PAKISTANIS AND IMPLICATIONS FOR DISEASE RISK

Elena Arciero[1], Theofanis Tsismentzoglou[2], Dan Mason[3], Neil Small[4], Eamonn Sheridan[2], Richard Trembath [5], Matthew E Hurles[1], David A van Heel[6], John Wright[3], Mark Iles[2], Hilary C Martin[1]

[1]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [2]Leeds Institute for Data Analytics, Worsley Building, University of Leeds, Leeds, United Kingdom, [3]Bradford Institute for Health Research, Bradford, United Kingdom, [4]University of Bradford, Bradford, United Kingdom, [5]School of Basic and Biomedical Sciences, King's College London, London, United Kingdom, [6]Centre for Genomics and Child Health, Queen Mary University of London, London, United Kingdom

The Pakistani population is extremely diverse and contains many different ethnic subgroups as a result of geographic origin, social organisation in kinship networks (*biraderi*) and language. We have analysed genetic data from > 4000 British-Pakistani individuals recruited to the Born in Bradford and East London Genes & Health (ELGH) cohorts. We first used a combination of SNP-chip and exome sequence data to explore the demographic history and genetic structure in the British-Pakistani population. The Bradford Pakistanis originate mostly from either the Mirpur region of Azad Kashmir or from the Punjab, and are relatively homogeneous relative to the full spectrum of genetic variation in Pakistan (e.g. the highly bottlenecked Kalash and admixed Hazara and Makrani). Nonetheless, we found signatures of population substructure within them, with most of them clustering according to their *biraderi* group rather than geographical origin. There are 14 self-defined *biraderi* groups with different proportions of ancestral components, revealing different genetic affinity both to other Pakistani and Indian populations and between themselves. The Pathan are the most distinct group followed by the Jatt, Choudhry, Bains and Rajput. The Kashmiri, Syed, Sheikh, Mughal are more genetically similar. Analysis of these groups suggests differing degrees of population bottleneck in historical times, with Qasabi, Gujjar and Sheikh having the smallest estimated historical effective population sizes. Overall, British-Pakistani individuals from Born in Bradford show high levels of autozygosity, with on an average of ~3.9% of the genome autozygous, and ~25% of individuals having > 2% autozygosity. The different *biraderi* groups have a similar total length distribution of runs of homozygosity. Genetically, Pakistanis within ELGH mostly cluster with the Kashmiri, Syed and Mughal but contain few people from the Pathan, Jatt and Choudhry. We are currently investigating whether Mendelian-disease causing alleles have increased in frequency as a result of founder effects in the various groups, and to what extent endogamy versus consanguinity contributes to disease in ~800 developmental disorder patients with Pakistani ancestry. This study highlights how social organisation such as the *biraderi* kinship networks impact both the distribution of genetic variation and potentially the risk of rare genetic disorders.

# SINGLE CELL MULTI-OMICS PROFILING REVEALS A HIERARCHICAL EPIGENETIC LANDSCAPE DURING MAMMALIAN GERM LAYER SPECIFICATION

Ricard Argelaguet[2], Stephen J Clark[1], Hisham Mohammed[1], Carine Stapel[1], Christel Krueger[1], Chantriolnt-Andreas Kapourani[5], Yunlong Xiang[8], Courtney Hanna[1], Guido Sanguinetti[5], Felix Krueger[1], Wei Xie[8], Peter Rugg-Gunn[1,7], Gavin Kelsey[1], Wendy Dean[1], Jennifer Nichols[7], Oliver Stegle[2,3], John C Marioni[2,4,6], Wolf Reik[1,6]

[1]Babraham Institute, Epigenetics Programme, Cambridge, United Kingdom, [2]European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom, [3]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, [4]University of Cambridge, Cancer Research UK, Cambridge, United Kingdom, [5]University of Edinburgh, School of Informatics, Edinburgh, United Kingdom, [6]Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [7]University of Cambridge, Wellcome - MRC Cambridge Stem Cell Institute, Cambridge, United Kingdom, [8]Tsinghua University, THUPKU Center for Life Sciences, Beijing, China

Formation of the three primary germ layers during gastrulation is an essential step in the establishment of the vertebrate body plan. Recent studies employing single cell RNA-sequencing have identified major transcriptional changes associated with germ layer specification. Global epigenetic reprogramming accompanies these changes, but the role of the epigenome in regulating early cell fate choice remains unresolved, and the coordination between different epigenetic layers is unclear.
Here we used single-cell Nucleosome, Methylome and Transcriptome sequencing (scNMT-seq), to generate the first triple-omics map of chromatin accessibility, DNA methylation and RNA expression during the exit from pluripotency and the onset of gastrulation in mouse embryos. We find dynamic dependencies between the different molecular layers, with evidence for distinct modes of epigenetic regulation. The initial exit from pluripotency coincides with the establishment of a global repressive epigenetic landscape, followed by the emergence of local lineage-specific epigenetic patterns during gastrulation. Notably, cells committed to mesoderm and endoderm undergo widespread epigenetic rearrangements, driven by loss of methylation in enhancer marks and a concomitant increase of chromatin accessibility. In striking contrast, the epigenetic landscape of ectodermal cells is already established in the early epiblast. Hence, regulatory elements associated with each germ layer are either epigenetically primed or epigenetically remodelled prior to overt cell fate decisions during gastrulation. This study provides the molecular logic for the existence of a default pathway in embryonic development and for the hierarchical emergence of the primary germ layers.

# BEST PRACTICES FOR DISCOVERING MOSAIC VARIANTS IN WHOLE-GENOME SEQUENCING DATA

Taejeong Bae[1], Liana Fasching[2], Flora Vaccarino[2], Alexej Abyzov[1], Brain Somatic Mosaicism Network[3]

[1]Mayo Clinic, Department of Health Sciences Research, Rochester, MN, [2]Yale University, Child Study Center, New Haven, CT, [3]National Institutes of Health, National Institute of Mental Health, Rockville, MD

Somatic mutations accumulate during human life resulting in mosaicism, the existence of genetically different cell population in a single organism. Recent studies have shown that somatic mosaicism is common in healthy individuals and can contribute to many diseases including cancer and brain-related disorders. The precise detection of mosaic variants remains a challenging task due to their low frequency, the relatively high sequencing error rate of current technologies, the abundance of repeat regions in the genome, and the relatively short length of sequencing read. Most programs for somatic variant discovery have been developed in the context of cancer, comparing two paired tumor and non-tumor samples. However, such programs optimized to a paired sample approach may not be appropriate to calling mosaic variants in non-tumor samples. To explore possible analytical strategies optimal for mosaic variant discovery while correcting for common types of artifacts which arise during sequencing and bioinformatic analysis, we generated whole-genome sequencing data for bulk cortical brain tissue and dural fibroblasts with ~200X depth of coverage, as well as for NeuN+ and NeuN- fractions with ~30X depth of coverage from the same cortical brain tissue. Altogether, we called 1,069 mosaic variant candidates from the four whole-genome sequencing datasets using GATK, adopting a single sample approach, with their default options and a few simple filtering criteria. Interestingly, we called fewer candidates in the bulk brain tissue, despite a much higher depth of coverage, than in the NeuN+ and NeuN- fractions. We then conducted ultra-deep amplicon-seq validations for these initial candidates in DNA of the same cortical brain tissue to obtain true mosaic variants and false positives. This validation experiment confirmed 6 out of 8 candidates discovered in the bulk brain tissue as true mosaic variants. In contrast, we found that most of the candidates in the NeuN+ and NeuN- fractions were enriched toward germline variants. The validation also revealed that most of the candidates in the dural fibroblasts were the result of clonal expansion during culture. The most common type of artifact resulting in false positive was strand bias of reads. In conclusion, for effective mosaic variant discovery in bulk tissue sequencing data, we suggest high-depth bulk DNA sequencing with more than 200X depth of coverage, using GATK for single sample approach variant calling, and filters to remove false positive somatic calls of germline variants and artifacts due to strand bias of reads.

# STRATEGIES FOR DIFFERENTIAL OMICS OF RESPONDERS AND NON-RESPONDERS IN PATIENT-DERIVED TUMOR ORGANOIDS

Matthew H Bailey[1,2,3], Brent S Pedersen[3], Katrin P Guillen[2], Sandra Scherer[2], Yi Qiao[2], Xiaomeng Huang[1,3], Andrew Butterfield[2], Alana Welm[2], Bryan E Welm [1,3], Gabor T Marth[1,3]

[1]University of Utah, Eccles Institute of Human Genetics, Salt Lake City, UT, [2]Huntsman Cancer Institute, Department of Oncological Sciences, Salt Lake City, UT, [3]Center for Genetic Discovery, USTAR, Salt Lake City, UT

Organoids have recently emerged as a promising technique to study stem-cell differentiation, tissue development, and tumor growth in three-dimensional culture. Furthermore, because of their relative stability in 96/384-well plates, organoids are a suitable model for drug screening. Here, we propose strategies for assessing molecular fidelity and differential drug response of breast cancer organoids. Patient tumor samples collected from biopsies were implanted in mice to establish patient-derived xenografts (PDX) and, following successful engraftment, xenograft tumors were processed into organoids (PDxO). To study the molecular fidelity of the PDX and PDxO models, we evaluated multi-omic data, including SNP array, whole-exome sequencing (WES) and RNA-seq, collected at different stages of model development. We used 'somalier' (https://github.com/brentp/somalier) to assess relatedness metrics and track genomic alterations, to determine the effect of passaging and environment on molecular stability. Additionally, using a drug screen of more than 20 compounds, we identified 12 PDxOs with dichotomous response to titrated treatment regimens. We leveraged multi-omic and drug response datasets to prioritize likely gene candidates for follow-up functional validation and hypothesis testing. We believe that this approach, combining drug-screening with next-generation sequencing of patient-derived models, will lead to the identification of novel biomarkers and disease modifiers. Additionally, if performed rapidly, this strategy may inform treatment decisions and improve patient outcomes.

IS IT TIME TO CHANGE THE REFERENCE GENOME?

Sara Ballouz, Alexander Dobin, Jesse Gillis

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

A major intent behind the original sequencing of the human genome was to provide a reference for future analysis and this has been wildly successful. The human reference genome has shaped methods and data across modern genomics, as it functions as the backbone for all genomic data and databases. It provides a scaffold in genome assembly to a background for read sequence alignment. However, there are some issues with the current reference, in particular its nearness to a “type specimen”. Here, we asked whether updating the current reference genome to a consensus genome – a genome which contains variation representing the most common alleles of a population - would have a strong impact on current practices such as DNA alignment, RNA-seq alignment and variant calling. We first assessed the current reference and show it is a good model of a human but not a general reference. We then built a series of population specific consensus genomes and benchmarked their use in place of the current reference. We suggest that switching to a consensus reference offers important advantages over the current reference with few disadvantages.

# scRNA-SEQ IDENTIFIES CHANGES IN CELL TYPES, CELL STATES, AND EXPRESSION LEVELS IN PATIENTS WITH IDIOPATHIC PULMONARY FIBROSIS.

Nicholas E Banovich[1], Linh Bui[1], Austin Gutierrez[1], Chris A Habermann[2], Mei-i Chung[1], Jonathan A Kropski[2]

[1]Translational Genomics Research Institute (TGen), Integrated Cancer Genomics, Phoenix, AZ, [2]Vanderbilty Univeristy, Department of Medicine, Nashville, TN

Idiopathic pulmonary fibrosis (IPF) is the most common and severe form of interstitial lung disease. IPF occurs in older adults and affects over 50,000 Americans each year. Most IPF patients die from respiratory failure within five years of diagnosis. The current therapies target downstream disease mechanisms, and while they modestly slow the decline in lung function, they have not been shown to improve survival or quality of life for IPF patients. There is considerable heterogeneity of clinical outcomes among IPF patients, and we believe this is due to distinct mechanisms and programs involved in disease initiation that culminate in a common a pathology of end-stage lung fibrosis. As such, the development of novel treatments hinges on our ability to better understand and target "upstream" disease mechanisms. However, progress has held back by the limited study of the cell types and molecular changes initiating IPF pathogenesis. To this end, we have performed single cell RNA-sequencing (scRNA-seq) on 20 samples from explant lungs of patients with IPF (n = 10) and healthy controls (n = 6). To assess the disease progression, which generally presents with more severe fibrosis in the distal lung and spreads into the proximal lung, we have sampled two sites, one more fibrotic and one less fibrotic, from four of the IPF patients. In total, we are analyzing over 67,000 cells.

Using graph-based clustering and canonical correlation analysis of scRNA-seq we identified distinct clusters, representing the major known subtypes in the lung. We found an enrichment of fibroblasts, macrophages and T-cells in disease lungs compared to controls, and identified cell states specific to the diseased lung samples. Furthermore, we identified hundreds of differentially expressed genes in most cell types when comparing fibrotic to control lungs, some of which exhibited previously unknown cell type specificity. Additionally, we identified numerous differentially expressed genes when comparing the less fibrotic tissue to both more fibrotic tissue and control samples. Some of the expression differences provide key details on the molecular program of disease progression. For example, *CLDN18* (previously reported to be downregulated in IPF) is highly expressed in type 2 airway epithelial cells from the less fibrotic and control samples but at low levels in the more fibrotic samples. Finally, we have identified allele specific expression within individual cell types in both diseased and healthy samples.

# LONGITUDINAL MOLECULAR TRAJECTORIES OF DIFFUSE GLIOMA IN ADULTS

Floris P Barthel[1,2], Kevin C Johnson[1], Frederick S Varn[1], Anzhela D Moskalik[1], Georgette Tanner[3], The GLASS Consortium[1], Jason T Huse[4], John F de Groot[4], Lucy F Stead[3], Roel G Verhaak[1]

[1]The Jackson Laboratory, Department of Genomic Medicine, Farmington, CT, [2]VU Medical Center, Brain Tumor Center, Amsterdam, Netherlands, [3]Leeds Teaching Hospitals NHS Trust, Leeds Institute for Medical Research, Leeds, United Kingdom, [4]The University of Texas MD Anderson Cancer Center, Department of Translational Molecular Pathology, Houston, TX

Therapeutic options for adult patients with a diffuse glioma are currently limited to a multimodal approach consisting of debulking surgery, radiation therapy and chemotherapy. These gliomas invariably progress and with increasing resistance to treatment. To longitudinally characterize this evolutionary process, we have pooled a large number of DNA sequencing and clinical annotation datasets from glioma patient tumors collected across multiple time points. Patients with at least two tumors and control DNA were considered for inclusion, collecting a total of 627 tumor samples from 274 patients across two to six time points. Leveraging mutational profiles from patients from the three major diffuse glioma subgroups, we observed that treatment with temozolomide can result in a hypermutator phenotype in gliomas regardless of *IDH1* mutation and 1p/19 codeletion status, and that hypermutation is likely a consequence and not a cause of malignant progression. Acquired aneuploidy was frequently detected in recurrent gliomas characterized by *IDH1* mutation but without 1p/19q codeletion and further converged with acquired cell cycle alterations and poor outcomes. We found that the large majority of newly acquired events over time are random events resulting in no or undetectable selection, suggesting an important role for neutral tumor evolution in glioma progression. We grouped patients on the basis of their evolutionary profile and found that patients whose tumors followed a neutral evolutionary trajectory have relatively favorable outcomes compared to cases with evidence for clonal selection. We were unable to detect an association between evolution and chemo- or radio-therapy, suggesting that these treatments do not drive a selective process. Finally, we found evidence in support of immunoediting, suggesting that active immunosurveillance is ongoing in gliomas from adult patients. Our work comprehensively resolves the molecular trajectories under therapy and provides evolutionary insights that have the potential to guide treatment.

# EXPRESSION OF EXOGENOUS mRNA IN HUMAN CANCER CELL LINES LEADS TO RISC-MEDIATED CELL DEATH

William E Putzbach[1,2], Ashley Haluck-Kangas[2], Elizabeth T Bartom[3], Marcus E Peter[2,3]

[1]University of Illinois at Chicago, Biochemistry and Molecular Genetics, Chicago, IL, [2]Northwestern University, Medicine, Chicago, IL, [3]Northwestern University, Biochemistry and Molecular Genetics, Chicago, IL

We have previously reported (Putzbach et al., 2017) that the majority of shRNAs and siRNAs derived from the CD95/Fas ligand (CD95L) transcript sequence trigger cell death. The genes down-regulated after one of these toxic siRNA/shRNAs is expressed in the cell are enriched both for the presence of the corresponding seed match in their 3'UTRs, and for essential cellular functions. Consequently, we refer to this cell death phenomenon as DISE — **D**eath **I**nduced by **S**urvival gene **E**limination. DISE is dependent on the RISC machinery. In particular, Ago2 knockout blocks DISE, while Drosha knockout augments DISE. We now show that expression of an exogenous full length mRNA for CD95L can trigger DISE. We pulled-down Ago2 to identify RISC-bound short RNAs, and found that short fragments of the transcript are being loaded into the RISC complex. This is true both in Drosha wild-type (WT) and Drosha knockout (KO) cells, but the reads are more abundant in Drosha KO cells. In WT cells, the RISC complex is dominated by microRNA species (partially protecting them from DISE). The miRNAs are replaced by other short RNAs upon Drosha removal. Interestingly, the induced CD95L transcript is not the only mRNA we see processed into short fragments and loaded into the RISC. We see fragments of many other transcripts as well, all involved in cell growth, including CCND1 and MYC. It is not clear what role these RISC-loaded transcripts are playing in the cell, but at least in the case of CD95L, they appear to play a role in DISE. We hypothesize that under certain circumstances, mRNAs can be processed and loaded into the RISC to affect cell fate through the RNA interference mechanism.

# INFERRING AFRICAN POPULATION DIFFERENCES FROM GENE EXPRESSION

Sanjit S Batra[1], Jonathan Fischer[1], Matthew Hansen[3], Derek Kelly[3], Yun S Song[1,2,4], Sarah Tishkoff[3]

[1]University of California, Computer Science Division, Berkeley, CA, [2]University of California, Department of Statistics, Berkeley, CA, [3]University of Pennsylvania, Departments of Genetics and Biology, Philadelphia, PA, [4]Chan Zuckerberg Biohub, San Francisco, CA

Genetic variation influences gene expression, and this can be efficiently mapped to specific genomic regions and variants. We used gene expression to elucidate the detailed features of genetic variation underlying gene expression variation. Using expression data from African populations in Tanzania and Ethiopia, we characterized the genetic architecture of gene expression within and between these populations. Genotype and whole-blood gene expression were available in individuals with Southern Ethiopian (SET, n = 56), Northern Ethiopian (NET, n = 71), and Tanzanian (TZ, n = 34) ancestry. We performed clustering of individuals using PCA, following batch-effect correction, to demonstrate that gene expression correlates well with population structure. Further, the genes distinguishing these populations are enriched in immune response. We also perform expression quantitative trait loci (eQTL) mapping in each population and find cis-eQTLs putatively associated with Tuberculosis, previously known to be observed in West African populations.

# LEVERAGING PUBLIC EPIGENOMIC DATASETS TO EXAMINE THE ROLE OF REGULATORY VARIATION IN THE THREE-DIMENSIONAL ORGANIZATION OF THE GENOME

Brittany A Baur[1], Shilu Zhang[1], Sushmita Roy[1,2]

[1]University of Wisconsin-Madison, Wisconsin Institute for Discovery, Madison, WI, [2]University of Wisconsin-Madison, Biostatistics and Medical Informatics, Madison, WI

Regulatory sequences such as enhancers can regulate the expression level of a gene hundreds of kilobases away through chromosomal looping that can bring distal regulatory elements in three-dimensional proximity to target genes. High-throughput Chromosome Conformation Capture (3C) technologies that measure the three-dimensional proximity of genomic loci at high resolution (e.g. 5kb) exist for a few model cell lines due to sequencing costs. We have developed a computational framework, L-HiC-Reg, to generate in silico contact counts using one-dimensional chromatin mark and transcription factor signals. L-HiC-Reg is based on a Random Forests regression model, which takes as input one-dimensional signals such as histone modifications, chromatin accessibility and accessible sequence motifs of architectural proteins such as CTCF. We trained L-HiC-Reg models on high-resolution Hi-C datasets for five cell lines and found that L-HiC-Reg accurately predicts contact counts in a new test cell line. We applied the models to generate contact count predictions in 55 human cell lines and tissues from the ROADMAP epigenomics project with no available high-resolution Hi-C measurements. Our predictions, when aggregated to a lower resolution (e.g. 40kb), accurately recapitulate the measured low-resolution contact count, as well as identify larger units of 3D chromosome organization such as topologically associating domains. Significant interactions from our predicted counts are associated with increased gene expression and enriched for interactions from ChIA-PET experiments. We used our compendium of significant interactions to link regulatory SNPs associated with different complex traits from the NHGRI GWAS catalog to subnetworks of genes. Our subnetworks are enriched for processes that are relevant to the phenotype of interest (e.g. SNPs associated with immune diseases are enriched for immune-related biological processes). Taken together, L-HiC-Reg and associated analysis pipeline offers a powerful resource to study three-dimensional genome organization in a large number of cell lines and interpret the role of regulatory variation in complex phenotypes.

# CRITICAL ASSESSMENT OF COMPUTATIONAL METHODS TO PREDICT THE IMPACT OF ENHANCER VARIATION

Nicholas Eng[1], Dustin Shigaki[1,2], Ayoti Patra[1,2], James Chen[1], Jin-Woo Oh[1,2], Wang Xi[1,2], Michael A Beer[1,2]

[1]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

Most SNPs associated with common human disease are intergenic and contribute to disease susceptibility by altering enhancer activity. Several machine learning methods have recently been proposed to detect the disrupted regulatory sequences and to predict the impact of regulatory mutations, based on SVMs and Deep Neural Networks (DNN). While these methods have similar cross-fold validation rates, variability in feature detection and importance can lead to significantly different predictions for mutation impact. Here I will show that most human population variation arises from weak transcription factor binding site disruption, and that differences in machine learning approaches can dramatically affect the accuracy of the predictions. Then, I will describe how these methods can identify functionally conserved regulatory elements missed by conventional sequence alignment methods. Finally, we show that training on multiple datatypes, cell-types, and tissues provides the most predictive models of mutation impact in the CAGI5 Regulation Saturation MPRA experiment. Together, these results show that statistical learning from large functional datasets can more accurately determine the quantitative contribution of weak binding sites to enhancer function in their native cellular contexts.

# QUANTIFICATION OF DIFFERENTIAL TRANSCRIPTION FACTOR ACTIVITY AND MULTIOMICS-BASED CLASSIFICATION INTO ACTIVATORS AND REPRESSORS: diffTF

Ivan Berest*[1], Christian Arnold*[1], Armando Reyes-Palomares[1], Giovanni Palla[1], Kasper Dindler Rasmussen [2,3], Kristian Helin[2,3], Judith Zaugg[1]

[1]EMBL, SCB, Heidelberg, Germany, [2]University of Copenhagen, Biotech Research and Innovation Centre (BRIC), Copenhagen, Denmark, [3]University of Copenhagen, Novo Nordisk Foundation Center for Stem Cell Biology, Copenhagen, Denmark

Genetic variation may affect the activity of disease-associated gene regulatory elements by transcription factor (TF) binding disruption in active chromatin. Given recent developments in methods for assaying chromatin accessibility, we were able to use aggregated changes in chromatin accessibility in the proximity of putative binding sites of a TF as a readout for its "activity" genome-wide, which we define as changes in open chromatin as measured by ATAC-Seq, DNase-seq or ChIP-Seq for any chromatin mark that is known to be affected by TFs. Identified TF activity is an important readout of cellular signalling pathways and it allows us to assess regulatory differences across conditions. To this end, we introduce a widely-applicable genome-wide method, diffTF, to assess differential TF activity and in particular to determine whether a specific TF acts globally as a transcriptional repressor or activator (available at https://git.embl.de/grp-zaugg/diffTF). This is done by integrating any type of genome-wide chromatin accessibility data with RNA-Seq data and *in silico* predicted TF binding sites. We corroborated the classification of TFs into repressors and activators by three independent analyses based on enrichments of active/repressive chromatin states, correlation of TF activity with gene expression, and activator- and repressor-specific chromatin footprints around the TF binding sites. Here, we present two case studies using diffTF: First, we applied diffTF in to a large ATAC-Seq/RNA-Seq dataset comparing mutated and unmutated chronic lymphocytic leukemia samples, where we identified dozens of known (40%) and potentially novel (60%) TFs that are differentially active. We were also able to classify almost half of them as either repressor or activator. Second,we applied diffTF to a small ATAC-Seq/RNA-Seq data set comparing two cell types along the hematopoietic differentiation trajectory (multipotent progenitors – MPP – versus granulocyte-macrophage progenitors – GMP). Here we identified the known drivers of differentiation and found that the majority of the differentially active TFs are transcriptional activators. Overall, diffTF was able to recover the known TFs in both case studies, identifying TFs that have been less well characterized in the given condition, and provides a classification of the TFs into transcriptional activators and repressors.

# BIOINFORMATICS PIPELINE FOR DETECTION AND PRIORITIZATION OF VERY LOW-FREQUENCY SOMATIC CANCER MUTATIONS IN TUMOR-DERIVED CIRCULATING CELL-FREE DNA SEQUENCING DATA

Preetida J Bhetariya, Gabor T Marth

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

Somatic cancer mutation profiling in circulating cell-free DNA (ccfDNA) is a promising avenue for non-invasive biomarker development. With growing interest in circulating tumor DNA (ctDNA) testing, it is important to establish a sensitive and specific variant calling and variant prioritization bioinformatic pipeline, which will, in turn, facilitate personalized therapies. We have developed and optimized an in-house variant calling pipeline by combining consensus calls from three variant callers MuTect2, LoFreq and Strelka2.0. Our cancer variant prioritization workflow continues by incorporating three resources: functional mutations from the TCGA datasets, known cancer driver mutations from CHASMplus, and clinically actionable mutations from the DEPO database. We tested the pipeline with a dataset that contains tumor and paired normal whole-exome sequencing of circulating cell-free DNA from cancer patients (TCGA). We found tumor-specific mutations in corresponding ctDNA samples, with the observed mutation spectra highly concordant with the matched tumor tissues. For each patient, we used our pipeline to identify clinically actionable variants detectable in ctDNA sample. Along with ever-improving sequencing library protocols, the computational pipeline is ready to be integrated into high throughput sequencing-based ctDNA mutation profiling, both in research and clinical settings.

# MERCURY: HIGH QUALITY VISUALIZATION AND REPRODUCIBLE EXPLORATORY ANALYSIS OF scRNA-SEQ DATA

Daniel Blibaum, Alexander Dobin

Cold Spring Harbor Lab, Genome Center, Cold Spring Harbor, NY

Single cell RNA-seq (scRNA-seq) has recently become an extremely popular methodology for the analysis of biological systems, providing transcriptome profiles at single cell resolution. However, scRNA-seq data is challenging to analyse due to its high dimensionality and sparsity, making visual and supervised analysis of the data crucially important. This typically involves visualization of cell clustering, manual editing and definition of cell clusters/groups, and selection of cell groups for differential expression analysis. Though hundreds of tools exist for statistical analysis of scRNA-seq data, very few interactive processing and visualization solutions are available, and all have limitations in both their functionality and workflow. Here we present Mercury, a tool for high quality visualization and reproducible exploratory analysis of scRNA-seq data, that addresses the shortcomings of current visualization software and unifies the scRNA-seq analysis pipelines.

Mercury provides flexible and aesthetically pleasing 2D and 3D visualization of scRNA-seq data that's fully integrated with the popular Seurat and Scanpy statistical analysis libraries. It uses Unreal Engine 4, a modern game development engine, for ultra-high quality, performant rendering and a sophisticated user experience. Mercury's close integration with analysis tools allows the user to define a sequence of data processing steps, change their parameters, reprocess the data using Seurat or Scanpy as a backend and update the visualization, all in an intuitive user interface and without leaving the app. Mercury speeds up data exploration by providing a streamlined user experience and workflow designed specifically for the core analysis tasks of cell cluster creation/labeling and differential gene expression analysis. In addition to core workflow improvements, Mercury provides important analysis functions beyond most existing tools, including high quality 3D visualization, seamless hierarchical clustering, and integration of multiple datasets. Mercury also enhances reproducibility by tracking user actions, which can allow another researcher, or the users themselves, to replay the steps they took in their analysis. Mercury is fully open source and uses well documented file formats for interoperability between analysis pipelines. Mercury closes the gap in the current scRNA-seq analysis pipeline, providing a sophisticated, usable, and open source tool for reproducible visual scRNA-seq analysis.

# CHARACTERIZING RNA STABILITY GENOME WIDE THROUGH COMBINED ANALYSIS OF PRO-SEQ AND RNA-SEQ DATA

Amit Blumberg[1], Yifei Huang [1], Yixin Zhao[1], Charles G Danko[2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, The Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cornell University, The Baker Institute for Animal Health, Ithaca, NY

Eukaryotic transcriptional regulation balances transcription and RNA decay to regulate gene expression and maintain cellular homeostasis. RNA transcription is a dynamic process that is coupled to several critical events (e.g., 5' capping, RNA splicing, and transcription termination) in order to develop a fully functional mature RNA, while RNA decay occurs post-transcriptionally and determines steady-state expression levels. Here we combine traditional RNA-seq data with nascent RNA sequencing data generated using precision nuclear run-on sequencing (PRO-seq) to characterize the determinants of RNA stability genome-wide across various types of RNAs.

We found that transcriptional patterns differed across gene types, especially between protein coding (mRNA) and non-coding (lincRNA) genes. LincRNAs were found to be less stable than mRNAs even after controlling for transcription levels (PRO-seq signal). At protein-coding mRNAs, our approach provided good estimations of RNA stability as compared to published direct measurements of RNA half-life. We used Structural Equation Modeling (SEM) to evaluate features predictive of stability and found that gene GC content correlates positively with RNA transcription and negatively with RNA stability in both mRNA and lincRNA genes. Other features, such as the number of exons per transcript, have a positive correlation with mRNA transcription, but a negative correlation with lincRNA transcription. In addition, we find that the known polyadenylation motif and U1 splicing motif model explains the directionality of divergent promoters but does not explain RNA stability levels in both mRNA and lincRNA genes. In contrast, analysis of enhancer RNAs (eRNAs) suggests (1) the U1 splicing motif model does explain some differences in RNA stability, (2) stable eRNAs are enriched in GC content close to the transcription start site (TSS), and (3) the differences between stable and unstable transcripts become less apparent moving further downstream from the TSS.

Finally, we found that some epigenomic markers (i.e. H3K79me2) are positively correlated with RNA stability, while others have a negative correlation (i.e. H3K4me1) even after controlling for RNA transcription levels. Surprisingly, one of those three epigenomic markers, H3K79me2, exhibited a similar correlation to the stability of enhancer RNAs. Overall, by using careful controls, we find that stable non-coding transcripts are enriched for features related to the transcription start site as compared to mRNA.

THE FITNESS LANDSCAPE OF CLONAL HAEMATOPOIESIS

Caroline J Watson[1], Alana Papula[2], Yeuk G Poon[1], Wing H Wong[3], Andrew L Young[3], Todd E Druely[3], Daniel S Fisher[2], Jamie R Blundell[1]

[1]University of Cambridge, CRUK Cambridge Cancer Centre, Early Detection Programme, Cambridge, United Kingdom, [2]Stanford University, Applied Physics, Stanford, CA, [3]Washinton University School of Medicine, Department of Paediatrics, St. Louis, MS

Somatic mutations acquired in healthy tissues as we age are major determinants of cancer risk. Whether variants confer a clonal "fitness advantage" or rise to detectable frequencies by chance, however, remains largely unknown. Here, by combining blood sequencing data from >50,000 individuals, we reveal how mutation, genetic drift and fitness advantages combine to shape the genetic diversity of healthy blood (`clonal haematopoiesis'). By analysing the spectrum of variant allele frequencies we quantify fitness advantages for key pathogenic variants and genes and provide bounds on the number of haematopoietic stem cells. The remarkably wide variation in variant allele frequencies observed across individuals is driven by chance differences in the timing of mutation acquisition combined with strong cell-intrinsic fitness advantages. Contrary to the widely held view that clonal haematopoiesis is driven by ageing-related alterations in the stem cell niche, the variant allele frequency spectrum for all variants is consistent with age dependence being driven clonal outgrowth leading to increased detectability at older ages.

# AN INTEGRATED INDEX OF PUBLIC GENE EXPRESSION DATABASES AND ITS APPLICATION FOR META-ANALYSIS OF HYPOXIC TRANSCRIPTOMES

Hidemasa Bono

Research Organization of Information and Systems, Database Center for Life Science, Mishima, Japan

Gene expression data have been archived as microarray and RNA-seq data in two public databases, NCBI Gene Expression Omnibus (GEO) and EBI ArrayExpress. In 2018, DNA DataBank of Japan (DDBJ) recently started similar repository called Genomic Expression Archive (GEA; https://www.ddbj.nig.ac.jp/gea/ ) [1]. These databases are useful resource for the functional interpretation of genes. But, these databases have been separately maintained and these may lack RNA-seq data whose original sequence data is available in the sequence read archive (SRA).
In order to promote life science researches, National Bioscience Database Center (NBDC) in Japan Science and Technology Agency (JST) cares for public databases in Japan. In tight collaboration with NBDC, we Database Center for Life science (DBCLS) have constructed an index for those gene expression data repositories, called all of gene expression (AOE) to integrate publicly available gene expression data. By collecting gene expression data by RNA-seq from SRA, AOE also includes data not included in GEO and ArrayExpress. The web interface of AOE can graphically query data in addition to the application programming interface (API). AOE is freely available from http://aoe.dbcls.jp/.

Using AOE, we attempted meta-analysis of hypoxic transcriptome for public expression data which have been archived as microarray and RNA-seq data in public transcriptome databases [2]. We aimed to elucidate functional insights in hypoxic response in data-driven way. We manually curated possible pairs of transcriptome before and after hypoxic stress from microarray and RNA-seq data from various hypoxic conditions (oxygen concentration and duration of hypoxia) and cell lines. The number of experiments for all genes was counted and those were classified into three categories, which are up-regulated, down-regulated, and unchanged. Genes up-regulated in all records contained well-studied hypoxia responsive genes, while down-regulated genes were unfamiliar in hypoxia. Meta-analysis approach to public gene expression database can be a useful tool for selecting candidate genes from gene expression profiles of various experimental conditions. The data produced in this study are open and ready for use to all biologists.

We will also present another application to study publicly available transcriptomes of non-human organism.

1. Kodama Y et al. DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data. Nucleic Acids Res. 2019; 47(D1):D69-D73. https://doi.org/10.1093/nar/gky1002
2. Bono H. Meta-analysis of hypoxic transcriptomes from public databases. BioRxiv https://doi.org/10.1101/267310

CO-EXPRESSION PATTERNS DEFINE VARIATION-INTOLERANT CHROMATIN REGULATORS ASSOCIATED WITH NEUROPSYCHIATRIC PHENOTYPES.

Leandros Boukas[1], James Havrilla[2], Peter Hickey[3], Aaron Quinlan[2], Hans Bjornsson[1], Kasper Hansen[1]

[1]Johns Hopkins University, Institute of Genetic Medicine, Baltimore, MD, [2]University of Utah, Department of Human Genetics, Salt Lake City, UT, [3]The Walter and Eliza Hall Institute of Medical Research, Molecular Medicine Division, Melbourne, Australia

Coding variants in chromatin regulators (CRs) have emerged as causes of neuropsychiatric phenotypes and cancer. Many studies have yielded insights into the individual functions of these genes; however, despite their obvious functional similarities, CRs have never been studied collectively as a class. Here, we used protein domain annotations to define 295 human CR genes. We first observed that 170/295 are highly intolerant to coding loss-of-function variation (pLI>0.9 in ExAC, $p<0.001$, OR=7.7 vs other genes), with 102/170 having no current disease associations; CRs are even more variation-intolerant than transcription factors ($p<0.001$, OR=4.4 for enrichment in the pLI>0.9 category). Additionally, brain-specific enhancers surrounding CRs were enriched for the common-variant heritability signal of 5 neuropsychiatric traits (median enrichment = 5.05).
Subsequently, we observed that CRs have multiple protein domains. We derived a local, domain-specific constraint score using variation in gnomAD, and showed that 82% of the domains responsible for their epigenetic function are constrained. This provides direct genetic evidence that chromatin state dysregulation is mediating the disease phenotypes of these (rare coding or common regulatory) gene-dosage disrupting variants. Finally, leveraging the GTEx dataset, we searched for expression signatures that distinguish between variation-intolerant and variation-tolerant CRs. Absolute expression levels or tissue specificity could not explain this difference. However, we discovered a large subset of 74 CRs which are co-expressed within multiple tissues; this is not observed with random genes or transcription factors. The co-expressed subset almost exclusively contains variation-intolerant genes (pLI>0.9 for 72/74, $p<0.001$). It shows enrichment for CRs causing Mendelian diseases with neuropsychiatric features ($p=0.006$, OR=3.3), even when accounting for high pLI ($p=0.003$, OR=3), but not for those linked to cancer ($p=0.41$). Moreover, partitioning heritability for IQ and generalized epilepsy (chosen since low IQ and seizures are the most common neurological symptoms of such Mendelian disorders), reveals that only the co-expressed CRs show significant enrichment.
Our results provide a comprehensive genetic analysis of human chromatin regulators and uncover a systems-level property (co-expression) associated with their variation-intolerance and involvement in brain disease.

# ASSESSMENT OF HUMAN MALE GONADAL MOSAICISM REVEALS MUTATIONAL ORIGINS AND STRATIFIES RISK OF TRANSMISSION

Martin W Breuss[1,2], Jonathan Sebat[3,4], Joseph Gleeson[1,2]

[1]University of California, San Diego, Neurosciences, Howard Hughes Medical Institute, La Jolla, CA, [2]Rady Children's Institute for Genomic Medicine, San Diego, CA, [3]University of California, San Diego, Beyster Center for Genomics of Psychiatric Diseases, La Jolla, CA, [4]University of California, San Diego, Psychiatry, Cellular and Molecular Medicine, Pediatrics, La Jolla, CA

A newborn child harbors on average 40-80 *de novo* genomic alterations absent in the genome of either parent, contributing to phenotypes under purifying selection like autism spectrum disorder (ASD). Approximately 80% of such mutations arise on the paternal haplotype, and they double in number for every additional decade of age of the father. Here, we assessed male gonadal mosaicism (GM) of such variants through 200× whole genome sequencing of sperm in 8 nuclear families. We detected 2.8% (23/812) of non-maternally phased single nucleotide variants (SNVs) as mosaic in father's sperm, the majority of which were either undetectable or at appreciably lower allelic fractions (AFs) in father's blood. When assessing all GM, sperm contained on average ~15 detectable variants, with a mutational signature predicting an origin during embryogenesis of the father rather than during aging. Surprisingly, AFs for sperm exclusive variants reached up to ~15%, which has important implication for the formation of the human germ line. We extended these observations to structural variants (SVs) and short tandem repeat variation, finding similar patterns, suggesting that this phenomenon extends to all *de novo* mutations. Finally, we also assessed mosaicism for disease causing variants. Among 19 families with a child with ASD attributed to *de novo* mutations (SNVs and SVs), we identified paternal GM for the mutation in 4, illustrating the impact of this phenomenon on disease inheritance. Our data provides evidence of an embryological origin for most detectable sperm mosaicism, valuable information that can be used to stratify risk of transmission to future offspring.

# MAPPING *TRANS*-ACTING GENETIC VARIATION IN GENE EXPRESSION BY SIMULTANEOUS QUANTIFICATION OF mRNA AND PROTEIN LEVELS IN SINGLE CELLS

Christian Brion, Sheila Lutz, Frank W Albert

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

Heritable variation in gene expression is a key component of phenotypic diversity. Most of this variation arises from *trans*-acting variants, which influence distant genes that are typically located on different chromosomes. Unfortunately, *trans*-acting variation has been more difficult to study than "local" variation that often resides in *cis*-regulatory elements and perturbs the expression of neighboring genes. Previous work suggested that *trans*-acting variants can specifically affect mRNA levels ("expression quantitative trait loci"; eQTLs) or protein levels (protein QTLs, pQTLs) of a given gene. However, most prior comparisons had low statistical power, and all relied on eQTLs vs. pQTLs mapped at different times, with different experimental designs, and in different environmental conditions.

To overcome these limitations, we developed a new system for simultaneous quantification of mRNA and protein levels at the same time in single, live cells of the yeast *Saccharomyces cerevisiae*. We tagged genes of interest with a C-terminal green fluorescent protein (GFP) sequence followed by a CRISPR guide RNA sequence flanked by two ribozymes in the 3'UTR. After transcription, the ribozymes release the guide RNA, which directs a deactivated Cas9 protein fused to a VP64 domain to drive expression of an mCherry reporter. This reporter serves as a readout of mRNA abundance, while the GFP-tag allows measurement of protein levels in the same cell.

Using this reporter system, we mapped the genetic basis of expression variation for ten genes using bulk segregant analysis. We generated millions of recombinant progeny by crossing two genetically divergent yeast strains and used fluorescence-activated cell sorting to collect pools of thousands of recombinant cells with extreme mRNA or protein levels. Whole-genome sequencing of these pools revealed eQTLs and pQTLs with high statistical power.

Across the ten genes, we identified 90 loci affecting gene expression. By design, all of these loci are located distantly from their target gene and act in *trans*. About a quarter of the loci have a strong influence on protein level of a given gene but show no effect on its mRNA level, demonstrating that genetic variation can affect protein expression independently of mRNA. We also observed strong eQTLs that did not overlap a pQTL, as well as QTLs controlling both mRNA and protein levels of the same gene, but sometimes in opposite direction.

Our design combining high statistical power and essentially perfect environmental control revealed complex and distinct genetic influences in *trans* on mRNA vs. protein levels. These results have major implications for understanding the diversity and evolution of gene expression within and between species.

# COMPARISON OF GENE- AND EXON-BASED CIS-eQTLs IN HUMAN SUBCUTANEOUS ADIPOSE TISSUE

S M Brotman[1], C K Raulerson[1], A Ko[2], P Pajukanta[3], M Laakso[4], K L Mohlke[1]

[1]UNC, Dept Genetics, Chapel Hill, NC, [2]UCLA, Dept Medicine, Los Angeles, CA, [3]UCLA, Dept Human Genetics, Los Angeles, CA, [4]Univ Eastern Finland and Kuopio Univ Hospital, Inst Clinical Medicine, Internal Medicine, Kuopio, Finland

Gene and exon expression quantitative trait loci (eQTL) can help identify associations between genetic variants and gene expression levels. While exon-based eQTL analyses (exon-eQTL) enable variant effects on splicing to be detected, these analyses generate a higher testing burden, and both gene- and exon-eQTL analyses could be used to detect variant effects on overall gene expression level. Here, we compared the ability of these two approaches to detect variant effects on overall gene expression level in adipose tissue. We evaluated RNA-sequencing of 434 adipose tissue samples (average 45M PE reads) from METSIM participants and retained genes (or exons) with >5 counts in >25% of samples. We analyzed TMM-adjusted CPMs, adjusted for BMI and PEER factors (30 for 25,610 genes, 50 for 160,730 exons), and tested 7.8M variants (MAF>0.01) in cis-eQTL (<1 Mb) analyses. We compared gene- and exon-eQTL results at 1% FDR thresholds, acknowledging that linkage disequilibrium (LD) and non-independent exons in multi-exon genes could influence the equality of these thresholds. 16,082 genes were significant in the gene-eQTL analysis, 16,225 genes contained at least one exon significant in the exon-eQTL analysis, and 13,441 genes were significant in both analyses. Among genes detected by only one analysis at 1% FDR, most (81% of exon-eQTL; 68% of gene-eQTL) were detected in the other analysis at 5% FDR. We further characterized the genes detected by only one analysis. Of the 2,781 genes found by only the exon-eQTL analysis, 2,365 (85%) were protein-coding and the median exon count per gene was 8, whereas of the 2,638 genes found by only the gene-eQTL analysis, 1,332 (51%) were protein-coding and the median exon count per gene was 4. Notably, 388 (15%) of the genes found only in the gene-eQTL analysis did not contain sufficient reads for the exon analysis. Based on the difference in median exon count, we asked whether the number of exons per gene influenced detection of significant eQTLs in the exon-eQTL analysis. Of 23,921 genes with an exon-eQTL association, 7,225 genes consist of only one exon, and 2,871 (40%) of those were significant, while of the 16,696 genes consisting of at least two exons, 13,354 (80%) were significant. Thus, the gene- and exon-eQTL analyses detected variant associations with many of the same genes, however, exon-eQTL analyses better identified eQTLs in genes with more exons and gene-eQTL analyses better detected eQTLs in genes with the lowest read counts.

# DETERMINING THE IMPACT OF MISSED INVERSIONS IN THE HUMAN GENOME BY DROPLET DIGITAL PCR (ddPCR)

Marta Puig[1], Jon Lerga-Jaso[1], Carla Giner-Delgado[1], Sarai Pacheco[1], David Izquierdo[1], Alejandra Delprat[1], Jack F Regan[2], George Karlin-Neumann[2], Mario Cáceres[1,3]

[1]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra (Barcelona), Spain, [2]Bio-Rad Laboratories, Digital Biology Center, Pleasanton, CA, [3]Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

Despite the interest in characterizing all genomic variation, the presence of large repeats at the breakpoints hinders the analysis of many structural variants. Thus, it has not been possible to investigate their population distribution, functional consequences and association to phenotypic traits. Here we describe a novel linkage-based application of droplet digital PCR (ddPCR) for the validation and genotyping of polymorphic inversions mediated by inverted repeats (IRs) of up to 150 kb, which are typically missed by current methods, such as long-read sequencing technologies. We developed ddPCR assays for 20 inversions ranging from 6.8 to 742 kb and flanked by IRs between 6.3 and 134 kb, that were used to validate 12 predicted inversions. In addition, we have generated new experimental human population information across 95 individuals from Africa, Europe and East-Asia for 16 of them, including four previously-known inversions for which there are not direct high-throughput genotyping assays, like the well-characterized 17q21 inversion. Through comparison with previous data, independent replicates and different inversion breakpoints, we have demonstrated that the technique is highly accurate and reproducible. Most of the studied inversions are widespread, with global frequencies ranging between 0.04 and 0.50, and several have significant differences between continents. Moreover, all except two show clear signs of being recurrent, given the lack of association of the inversion alleles with SNPs or haplotypes. Finally, thanks to the generated genotypes, we have been able to check the effect of these inversions on gene expression, both validating gene expression differences reported previously for two inversions and finding new candidate associations. This work therefore makes it possible to screen these and other complex genomic variants quickly in a large number of samples for the first time, highlighting the importance of direct genotyping to assess their potential functional effects and clinical implications.

# COMMON MITOCHONDRIAL DNA VARIANTS IMPLICATED IN VARIATIONS IN MITOCHONDRIAL OXIDATIVE ACTIVITY

Na Cai[1,2], Aurora Gomez-Duran[3,4], Ekaterina Yonova-Doing[5], Marc Jan Bonder[2], Bogdan Mirauta[2], John Danesh[5], Joanna Howson[5], Oliver Stegle[2], Patrick Chinnery[3,4], Nicole Soranzo[1]

[1]Wellcome Sanger Institute, Human Genetics, Cambridge, United Kingdom, [2]European Bioinformatics Institute (EMBL-EBI), Statistical Genomics, Cambridge, United Kingdom, [3]University of Cambridge, MRC Mitochondrial Biology Unit, Cambridge, United Kingdom, [4]University of Cambridge, Clinical Neurosciences, Cambridge, United Kingdom, [5]University of Cambridge, Public Health and Primary Care, Cambride, United Kingdom

The human mitochondrial DNA (mtDNA) encodes 13 proteins and 22 RNAs involved in the synthesis of key subunits of the mitochondrial respiratory complexes (RCs). Little is known of the effects common mtDNA variations have on human traits and common diseases, as no association studies on the mtDNA have been performed at scale to date. In this study, we obtain mtDNA sequence variations in 8,000 healthy British blood donors in the INTERVAL study and assess their effects on 896 metabolites that capture a wide spectrum of biological processes and pathways. We find 14 mtDNA SNPs significantly associated with a single metabolite, N-formylmethionine (fMet). 13 of the SNPs are from Haplogroups Uk (top SNP mt.1811G from 16s mitochondrial ribosomal RNA RNR2, P value = $2.92 \times 10^{-10}$).

As fMet is the initiation amino acid for translation in the mitochondria, we probe the molecular mechanism behind this association using cohort-based transcriptomics and proteomics data, as well as cytoplasmic hybrid (cybrid) cellular models. We perform the first cis-eQTL analysis on mtDNA encoded genes, and find that mtDNA SNPs in Haplogroup Uk significantly and consistently down-regulate the expression of MT-ND3 (a subunit of mitochondrial respiratory Complex I) across 36 out of 41 tissues represented in the GTEx Consortium (5% study-wide FDR, $p < 10^{-8}$). Using both transcript and protein quantification in induced pluripotent cell (iPSC) lines in the HipSci Consortium, we find that iPSC lines with mtDNA Haplogroup Uk show a higher correlation between transcript and protein levels (highest Spearman's $\rho$ for ATP6 =0.87, P = 0.002). We verify all our findings using cybrid cell lines, and using Blue Native Polyacrylamide Gel Electrophoresis (BN-PAGE) we show that those with mtDNA Haplogroup Uk have lower levels of Complex I Super-Complexes and oxidative phosphorylation (OXPHOS) activity.

Taken together, our results demonstrate that mtDNA SNPs exert control over mitochondrial gene expression, and suggest higher fMet levels in Haplogroup Uk may enable greater local mitochondrial control over production and function of mitochondrial RCs, in particular Complex I.

# TRACKING THE INTERDISCIPLINARITY AND MISAPPROPRIATION OF GENOMICS RESEARCH THROUGH SOCIAL MEDIA

Jedidiah Carlson, Kelley Harris

University of Washington, Genome Sciences, Seattle, WA

Social media plays an increasingly important role in the discussion and dissemination of scientific literature, spawning the concept of "altmetrics"—measures of scholarly output based on social media references rather than literature citations. In addition to augmenting traditional scientometrics, altmetrics provide an unprecedented opportunity to learn about the audience(s) engaging with scientific publications. The need to understand these audience dynamics is particularly salient for the field of genomics which is frequently in the public spotlight and often has immediate relevance for society.

In this study, we selected 20 genomics papers that were published in the last two years and referenced on Twitter >100 times. Using Latent Dirichlet Allocation, we summarized each paper's audience as a composite of 12 "topics," represented by sets of keywords that frequently co-occur in users' Twitter biographies. The inferred topics typically follow intuitive semantic classifications. For example, the two most common audience topics for archaic admixture papers are defined by the keyword sets [genetics, phd, genomics, evolution, student] and [phd, archaeology, anthropology, student, human], suggesting a primary audience of evolutionary geneticists, archaeologists, and anthropologists. This granular differentiation of a paper's academic audience may encourage collaboration and provide a new altmetric for evaluating the interdisciplinary impact of a publication or journal.

Our approach also reveals detailed traits of the sociodemographic, occupational, and/or ideological makeup of non-academic lay audiences. For example, proof-of-concept papers demonstrating forensic identification of individuals using genetic databases are of interest to privacy advocates, genealogists, and law enforcement, suggesting these are potentially important stakeholder groups. Such information may prove useful in developing more effective public engagement and science communication strategies and helping quantify scholars' broader impacts.

Finally, we find that several papers have audience sectors aligned with white nationalism and other extreme right-wing movements, sometimes comprising over 50% of the total audience. These users show strong network homophily with prominent white nationalists and frequently use scientific research to bolster this ideology. Unexpectedly, the papers embraced by this audience are not restricted to topics typically associated with scientific racism (e.g., heritability of IQ), but span a range of research areas. We discuss how racist misappropriation of genomics research through social media is a non-trivial issue that the genomics community must address.

# IDENTIFYING *DE NOVO* NONCODING MUTATIONS ASSOCIATED WITH AUTISM SPECTRUM DISORDERS

Christopher Castro[1], Shengcheng Dong[1], Stephanie Bielas[2], Alan P Boyle[1,2]

[1]The University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]The University of Michigan, Department of Human Genetics, Ann Arbor, MI

Autism spectrum disorders (ASDs) are a group of neurodevelopmental conditions often characterized by difficulties with social interaction, communication, and behavior. While the genetic contribution to ASD risk has been extensively explored, most studies have restricted their focus to variants in protein-coding regions, which collectively account for only about 2% of the human genome. This limitation is in sharp contrast to the results of many genome-wide association studies that have revealed that the vast majority of disease-associated variants lie within the noncoding regions. Therefore, investigating the noncoding regions of individuals with ASD has the potential to uncover novel ASD risk-variants. In particular, de novo variants have been implicated in the genetic etiology of ASD. The rare nature of these uninherited variants makes them potentially more deleterious than common variants, as they are not subject to selective pressures.

Our aim is to annotate and prioritize noncoding de novo single nucleotide variants (SNVs) in regulatory regions of the genome in order to better understand their contributions to ASD risk. To identify predicted pathogenic de novo SNVs, we have analyzed whole genome sequence data from 1,918 affected families from the Simons Simplex Collection. We hypothesize that a subset of these SNVs disrupt the function of regulatory elements that ultimately impair expression of ASD risk genes. We use RegulomeDB for comprehensive annotation of known and predicted regulatory elements, including data from the Encyclopedia of DNA elements (ENCODE) consortium and the Roadmap Epigenomics Consortium. SNVs were prioritized according to the RegulomeDB heuristic scoring, which is based on their overlap with predicted functional elements. This project contributes to our understanding of the functions of noncoding regulatory elements and their contributions to ASD genetics and pathogenesis.

# THE GENOMIC ARCHITECTURE OF INTROGRESSION AMONG SIBLING SPECIES OF BACTERIA

Maria Izabel A Cavassim[1], Sara Moeskjær[2], Camous Moslemi[2], Bryden Fields[3], Asger Bachmann[1], Bjarni Vilhjalmsson [1], Mikkel H Schierup[1], J Peter W Young[3], Stig U Andersen[2]

[1]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark, [2]Aarhus University, Department of Molecular Biology, Aarhus, Denmark, [3]University of York, Department of Biology, York, United Kingdom

Gene transfer between bacterial species is an important mechanism for adaptation. For example, sets of genes that confer the ability to form nitrogen-fixing root nodules on host plants have frequently moved between *Rhizobium* species. It is not clear, though, whether such transfer is exceptional, or if frequent inter-species introgression is typical. To address this, we sequenced the genomes of 196 isolates of the *Rhizobium leguminosarum* species complex obtained from root nodules of white clover (*Trifolium repens*). Core gene phylogeny placed the isolates into five distinct genospecies that show high intra-genospecies recombination rates and remarkably different demographic histories. Most gene phylogenies were largely concordant with the genospecies, indicating that recent gene transfer between genospecies was rare. In contrast, very similar symbiosis gene sequences were found in two or more genospecies, suggesting recent horizontal transfer. The replication and conjugative transfer genes of the plasmids carrying the symbiosis genes showed a similar pattern, implying that introgression occurred by conjugative plasmid transfer. The only other regions that showed strong phylogenetic discordance with the genospecies classification were two small chromosomal clusters, one neighbouring a conjugative transfer system. Phage-related sequences were observed in the genomes, but appeared to have very limited impact on introgression. Introgression among these closely-related species has been very limited, confined to the symbiosis plasmids and a few chromosomal islands. Both introgress through conjugative transfer, but have been subject to different types of selective forces.

# A MAP OF MITOTIC RECOMBINATIONS IN BLOOM SYNDROME FAMILY TRIOS REVEALED BY DIRECT HAPLOTYPE TAGGING

Marek Kučka[1], Andreea Dréau[1], Christopher M Cunniff[2], Yingguang Frank Chan[1]

[1]Friedrich Miescher Laboratory of the Max Planck Society, Chan Lab, Tuebingen, Germany, [2]Weill Cornell Medicine, Department of Pediatrics, New York, NY

Mitosis is an essential process in the maintenance of genome integrity in healthy somatic tissues. Despite its importance, surprisingly little is known about the location and propensity of mitotic recombination in the human genome and its link to cancer. *Bloom Syndrome helicase* is a key gene that ensures proper segregation during mitosis. Patients suffering from Bloom Syndrome show high rates of loss of heterozygosity through mitotic recombination and develop many types of cancer at a young age.
We present here the first mitotic recombination map in human from healthy and Bloom Syndrome genomes, made possible by a low-cost, highly accurate linked-read barcoding strategy we have developed called "haplotagging". We will first show the ~99.9% accurate molecular phasing results against the reference human sample GM12878. We then apply haplotagging to family trios of Bloom Syndrome patients. A key feature of haplotagging is the ease of applying > 21 million barcodes per sample. This barcode diversity results in unique tagging of single haplotypes and outperforms commercial microfluidic alternatives by avoiding barcode collision. Using haplotagging on primary blood leukocytes in four families with Bloom Syndrome patients and their unaffected, carrier parents, we can detect rare mitotic recombinant molecules, from which we generate a genome-wide profile of mitotic recombination rate from single individuals. Data from representative Bloom Syndrome patients show efficient recovery of large DNA molecules (on average 21 reads per molecule spanning 51 kb and covering 6 phase-informative SNPs), about 0.15% of which were recombinant. We can pin-point over a 1000 mitotic recombination events distributed across the genome (median resolution: 3.9 kb). Crucially, Bloom Syndrome patients show significantly higher mitotic recombination rate than their parents, who are only carriers. We will discuss the underlying genomic features associated with increased mitotic recombination, including mitotic recombination hotspots and their clinical implications.
Haplotagging is a simple, one-step protocol that enables linked-read sequencing using standard Illumina reagents with no instrumentation or extra costs. We show here how we use haplotagging to reveal the recombination landscape in single human individuals. We also anticipate applications in genome assembly and geno- or haplotyping of hundreds to thousands of samples.

# PENETRANCE OF HIRSCHSPRUNG DISEASE DEPENDS ON CELL FATE DECISIONS IN THE ENTERIC NERVOUS SYSTEM (ENS)

Sumantra Chatterjee[1], Elizabeth Vincent[2], Dallas Auer[2], Loyal Goff[2], Aravinda Chakravarti[1]

[1]NYU School of Medicine, Center for Human Genetics and Genomics, New York, NY, [2]Johns Hopkins University School of Medicine, Institute of Genetic Medicine, Baltimore, MD

Despite considerable genetic heterogeneity, ~50% of the genetic risk of Hirschsprung disease (HSCR, congenital aganglionosis) in European ancestry subjects is from common enhancer and rare coding variants at the receptor tyrosine kinase gene *RET*. HSCR has a profound male bias (4:1) but the underlying cause for this sex difference in penetrance has been elusive.

Mouse models with homozygous loss of *Ret* are a close model for HSCR. We have used bulk (E10.5-E14.5) and single cell RNA-seq (E12.5, E14.5) on developing whole gut tissue in wildtype and *Ret* null mice to track the overall transcriptional changes and elucidate how changes in and variation across the ENS. The bulk analyses identified significant transcriptional changes in early transcription factors and genes affecting gut epithelial and mesenchymal development upon loss of *Ret* signaling. However, these studies cannot distinguish cell autonomous from non-autonomous changes or identify which enteric cells contribute to these changes.

Our results show that two distinct *Ret*-expressing cell populations, early progenitor cells expressing glial makers (*S100b, Gfap, Sox10*) and a more fate-committed enteric neuronal population (marked by genes like *Nefh, Nefl, Snap25*), are present early in the developing gut. There is great diversity among the neuronal population which includes Noradrenergic (*Dh, Ddc*), Acetylcholinergic (*Ache, Slc18a3*) and Nitrergic (*Nos, Gal*) neurons, some of which haven't been previously reported in the ENS. Interestingly, loss of Ret signaling affects only a subset of them. Further, a subset of the Ret null neuronal cells in the female gut change fate to a dopaminergic neuronal cell type and express markers (*Th, Ddc, Isl1*) absent or weakly expressed in other cells. These female-specific dopaminergic cells also exclusively express *Gfra3* (a family member of the canonical *Ret* co-receptor *Gfra1*), highlighting the use of alternate co-receptors and alternate signaling cascades in the absence of *Ret*.

Thus, distinct subsets of *Ret*-expressing cells exist and the effect of loss of *Ret* signaling is cell-type and sex-specific in HSCR. Surprisingly, some neuronal cells in females escape cell death and change fate in the absence of Ret signaling, potentially using a non-canonical Ret signaling pathway. We hypothesize that this feature may be a primary cause of the lower female HSCR penetrance.

# EPIML: A WEB SERVER FOR EPISTASIS ANALYSIS USING MACHINE LEARNING METHODS

Junjie Chen, Jia Wen, Xinghua Shi

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

Epistasis describes a phenomenon that different genetic variants affect phenotypes jointly instead of independently. Epistasis among different variants contributes to an appreciable proportion of the hidden heritability of complex traits, and plays an important role in gene regulation. However, genomic data is typically high-dimensional, where the number of features (genetic variants) is usually much larger than sample size. The large number of genetic variants that present in high dimensional genomic data makes it extremely challenging for epistasis analysis since the searching space for epistasis grows exponentially with the number of genetic variants. Penalized regression based machine learning methods work well with analyzing high-dimensional and large-scale genomic data in that they use penalized terms to impose sparsity to simultaneously enable scalable feature selection and dimension reduction to avoid overfitting. Such methods including classical Least Absolute Shrinkage and Selection Operator (LASSO), newly proposed spike-and-slab LASSO generalized linear models (ssLASSO GLMs) and Empirical Bayesian Elastic Net (EBEN), are suitable and scalable for analyzing genomic data. Nonetheless, these methods usually include multiple hyperparameters that need to be carefully trained for optimization on a particular genomic dataset. The intrinsic complexity of optimizing these models limits their wide adoption and makes it hard to reproduce results. To support reproducible epistasis analysis on high dimensional genomic datasets, we have developed an integrated and user friendly web server for epistasis analysis using classical and emerging machine learning methods (EpiML: http://epiml.uncc.edu ). With this web server, a user can upload the data and submit jobs. Our server handles all the data processing and analysis in the background. Once the submitted job is finished, the user can download the identified main and epistatic effects where one or multiple genetic variants affect a particular phenotype. Interactive figures, graphs, tables, and links to meta-data and external databases will be provided to help retrieve, explore and interpret the identified effects. Users can also download the epistasis analysis code encapsulated in Jupyter notebooks, and run it on pre-configured Docker containers.

# IDENTIFICATION OF STRUCTURAL VARIATION ASSOCIATED WITH CARDIOMETABOLIC TRAITS IN THE FINNISH POPULATION

Lei Chen[1], Liron Ganel[1], Haley J Abel[1], David E Larson[1], Allison Regier[1], Indraniel Das[1], Susan K Service[2], FinMetSeq Consortium[9], Aarno Palotie[4,5,6,7], Samuli Ripatti[5], Markku Laakso[8], Nelson B Freimer[2], Michael Boehnke[3], Adam E Locke[1], Nathan O Stitziel[1], Ira M Hall[1]

[1]Washington University School of Medicine, St Louis, MO, [2]University of California, Los Angeles, CA, [3]University of Michigan, Ann Arbor, MI, [4]Broad Institute of MIT and Harvard, Cambridge, MA, [5]University of Helsinki, Helsinki, Finland, [6]Massachusetts General Hospital, Boston, MA, [7]Harvard Medical School, Boston, MA, [8]University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland, [9]FinMetSeq Consortium, Finland

Cardiovascular disease (CVD) is a leading cause of death worldwide and has significant heritability. Most previous genome-wide studies have studied single nucleotide variants and short indels, and the contribution of structural variation (SV) to CVD is largely unknown. We are midway through a large whole genome sequencing (WGS) study of coronary artery disease and CVD risk factors, undertaken by our NHGRI Center for Common Disease Genomics. Here, we describe results from an interim association analysis of SV and cardiometabolic traits in 4,108 individuals from Finland. Relative to prior work, our study benefits from deep WGS (≥20X), extensive trait measurements, and the unique population history of Finland, which has resulted in an excess of low-frequency, high-impact variants that are rare or absent in other populations. We used a combination of read-depth and breakpoint mapping approaches to detect 63,739 high-confidence SVs including deletions (47%), duplications (25%), multi-allelic CNVs (16%), inversions (0.44%), mobile element insertions (3.6%), and other difficult-to-classify SVs (7.3%). We tested 13,217 SVs with MAF>0.1% for association with 137 cardiometabolic traits. We attempted to replicate SVs reaching a suggestive level of statistical significance ($P \leq 10^{-5}$) in an independent exome sequencing dataset of 16,135 Finnish samples. We discovered 11 genome-wide significant associations, including several novel low frequency SVs with large phenotypic effects. For example, deletion of the ALB promoter is associated with decreased serum albumin ($p=2.39 \times 10^{-21}$) and increased cholesterol ($p=1.81 \times 10^{-8}$), a complex multiallelic CNV at PDPR is associated with pyruvate ($p=1.53 \times 10^{-21}$) and alanine ($p=7.50 \times 10^{-11}$), and recurrent gene conversion between the CYP21A2/CYP21A1P gene/pseudogene pair is associated with VLDL ($p=1.3 \times 10^{-7}$). Remarkably, roughly half of trait-associated SVs are complex multi-allelic variants that are not well tagged by SNPs. Our results suggest that SVs underlie some trait associations not detectible from GWAS and have the potential to greatly expand our knowledge of genetic factors underlying CVD risk.

# RE-EXAMINING THE EVIDENCE OF POLYGENIC ADAPTATION AT HEIGHT-ASSOCIATED LOCI IN MAINLAND EUROPE AND SARDINIA

Minhui Chen[1], Masato Akiyama[2], Carlo Sidore[3], Kazuyoshi Ishigaki[2], Yukinori Okada[2], Yoichiro Kamatani[2], David Schlessinger[4], Charleston Chiang[1]

[1]University of Southern California, Center for Genetic Epidemiology, Los Angeles, CA, [2]RIKEN Center for Integrative Medical Sciences, Laboratory for Statistical Analysis, Yokohama, Japan, [3]Consiglio Nazionale delle Ricerche, Istituto di Ricerca Genetica e Biomedica, Cagliari, Italy, [4]NIA, Laboratory of Genetics, Baltimore, MD

Adult height was one of the earliest putative examples of polygenic adaptation in humans. By constructing polygenic height scores using effect sizes and frequencies from hundreds of genomic loci robustly associated with height, it was demonstrated that Northern Europeans were genetically taller than Southern European beyond neutral expectation. However, this conclusion was recently challenged. Sohail et al. and Berg et al. showed the polygenic signature disappeared if GWAS summary statistics from UK Biobank (UKB) were used in the analysis, suggesting that residual uncorrected stratification from large-scale consortium studies like GIANT were responsible for the previously observed difference in genetic. While effect sizes in consortium GWAS were likely over-estimated, it remains an open question whether height loci exhibit signals of polygenic adaptation in any human population. In this study, we re-examine this question focusing on one of the shortest European populations, the Sardinians. We show that previously reported signature of polygenic adaptation in Sardinians remains when using height loci and effect sizes ascertained from UKB (~ 0.37 s.d. shorter than CEU, P = 2.18e-5). To further alleviate concerns of biased ascertainment of GWAS loci, we ascertained height loci from the Biobank of Japan (BBJ), and showed that the results are unchanged (~ 0.36 s.d. shorter than CEU, P = 8.85e-7). We also observed that the exaggerated signature of polygenic adaptation using summary statistics from GIANT was partly due to ascertainment of the most significantly associated SNP from each of 1700 blocks in approximate linkage equilibrium, a practice that enriched for uncorrected stratification. When focusing on mainland European populations, there is a nominally significant signal of shorter height in Southern Europeans (i.e. IBS, ~ 0.12 s.d. shorter than CEU, P = 0.03) if we used only genome-wide significant loci from UKB data, although we did not replicate this signal when using loci ascertained from BBJ data. In summary, by examining height loci ascertained in a distant East Asian population, we further supported evidence of polygenic adaptation in at least some European populations. Whether polygenic adaptation occurred at height loci in mainland Europe remains an open question, and more powerful, within-population haplotype-based approach may be more informative.

# ROBUST AND WINDOW-INSENSITIVE MIXTURE MODEL APPROACHES FOR LOCALIZING BALANCING SELECTION

Xiaoheng Cheng[1,2], Michael DeGiorgio[2,3,4]

[1]Penn State University, Huck Institutes of Life Sciences, University Park, PA, [2]Penn State University, Department of Biology, University Park, PA, [3]Penn State University, Institute for CyberScience, University Park, PA, [4]Penn State University, Department of Statistics, University Park, PA

Long-term balancing selection typically leaves narrow footprints of increased genetic diversity, and therefore most detection approaches only achieve optimal performances when sufficiently small genomic regions (i.e., windows) are examined. Such methods are sensitive to window sizes and suffer substantial losses in power when windows are large. This issue creates a tradeoff between noise and power in empirical applications. Here, we employ mixture models to construct a set of five composite likelihood ratio test statistics, which we collectively term *B* statistics. These statistics are agnostic to window sizes and can operate on diverse forms of input data. Through simulations, we showed that they exhibit comparable power to the best-performing current methods, and retain substantially high power regardless of window sizes. Moreover, these methods display considerable robustness to high mutation rates and uneven recombination landscapes, as well as an array of other common confounding scenarios. Further, we applied these statistics on a bonobo population-genomic dataset. In addition to the MHC locus, we uncovered several novel candidate genes, such as *CSMD1* and *BPIFB3*, which are involved in innate immunity, as well as genes potentially related to sexual dimorphisms, such as *DMRT1*. Finally, we integrated the set of statistics into an open-source software named BalLeRMix, for future applications by the scientific community.

# THE IMPACT OF GLOBAL AND LOCAL POLYNESIAN GENETIC ANCESTRY ON COMPLEX TRAITS IN NATIVE HAWAIIANS

Hanxiao Sun[1], Meng Lin[1], Lynne Wilkens[2], Loic Le Marchand[2], Christopher A Haiman[1], Charleston W Chiang[1]

[1]University of Southern California, Center for Genetic Epidemiology, Department of Preventive Medicine, Los Angeles, CA, [2]University of Hawaii Cancer Center, Epidemiology Program, Honolulu, HI

Epidemiologic studies have shown that Native Hawaiians exhibit an excess risk for obesity, Type-2 diabetes, cardiovascular diseases and several common cancers compared to European or Asian Americans living in the Hawaiian islands today. However, like many non-European or indigenous populations, Native Hawaiians are understudied, particularly from a genetic standpoint. To fill this gap, we took advantage of the availability of genome-wide array data on nearly 4,000 Native Hawaiians from the Multiethnic Cohort to study the risk related to their genetic ancestry for these complex diseases and the associated quantitative traits. Specifically, we first identified a subset of Native Hawaiians with extremely high proportion of Polynesian ancestry to construct an internal reference panel to help us estimate the genomic proportion of Polynesian ancestry ("global ancestry"), as well as Polynesian ancestry along each chromosome ("local ancestry"), in the remaining 3,428 unrelated individuals. We then tested the associations between proportion of global Polynesian ancestry and both binary (obesity, type-2 diabetes, prostate cancer, breast cancer) and quantitative (BMI, glucose, insulin, HDL, LDL, triglyceride, total cholesterol) traits. After adjusting for available covariates on socioeconomic status, we found that per 10% increase in Polynesian genetic ancestry, there is a 17% increase in the odds of being obese ($P < 2x10^{-16}$), a 0.076 s.d. increase in BMI ($P < 2x10^{-16}$) and a 0.046 s.d. increase in HDL ($P = 8.2x10^{-6}$). We also found weaker nominal associations between proportion of Polynesian genomic ancestry and insulin and total cholesterol. To examine regions along the genome that might harbor Polynesian-specific variants associated with risk of disease or biomarkers, we conducted admixture mapping on these traits and identified genomic regions at approximately 57.4Mb-70.3Mb on chromosome 6 and 22.4Mb-29.0Mb on chromosome 22 to be putatively associated with Type-2 diabetes and triglycerides, respectively; neither of the admixture signals are explained by known GWAS variants reported in these regions, suggesting potentially new regions of interest specific to Native Hawaiians.

# THE ADAPTIVE HISTORY OF AFRICAN RAINFOREST HUNTER-GATHERERS USING WHOLE-EXOME SEQUENCING

Jeremy Choin*[1], Marie Lopez*[1], Martin Silvert[1], Martin Sikora[2], Helene Quach[1], Christine Harmant[1], Patrick Mouguiama-Daouda[3], Alain Froment[4], George H Perry[5], Luis B Barreiro[6], Paul Verdu[4], Etienne Patin[1], Lluis Quintana-Murci[1]
*These authors contributed equally to this work

[1]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France, [2]Center for GeoGenetics, University of Copenhagen, Copenhagen, Denmark, [3]Laboratoire de Langue, Culture et Cognition, University Omar Bongo, Libreville, Gabon, [4]Musee de l'Homme, Université Paris Diderot, Paris, France, [5]Department of Biology, Pennsylvania State University, Pennsylvania, PA, [6]Genetics Section, University of Chicago, Chicago, IL

Central Africa, a forested region that supports an exceptionally high biodiversity, hosts the world's largest group of active hunter-gatherers. To dig into the various ecological, nutritional and pathogenic challenges imposed by rainforest environments, we searched for the occurrence and mechanisms of genetic adaptation likely underlying the adaptive history, and distinctive phenotypes, of African “pygmy” groups. To do so, we evaluated the relative contribution of classic sweeps, polygenic adaptation and adaptive admixture characterizing the history of positive selection of these populations. We generated high-coverage whole exome sequences from 566 African rainforest hunter-gatherers and neighboring farmers from 14 different populations across central Africa, which were supplemented with 40 whole genome sequences. We found evidence, for the first time, of a unique selective sweep that is shared among all groups of rainforest hunter-gatherers — from the west to the easternmost part of the Central African belt. This sweep encompasses the regulatory region of TRPS1, a transcription factor associated to short stature and craniofacial traits, notably through the repression of STAT3, and also to immunosurveillance. Additional signals of selective sweeps that are shared across hunter-gatherer groups involve both antiviral and antibacterial functions. Our analyses of gene subnetworks detected signals of polygenic adaptation targeting primarily the glutathione metabolism involved in the fine-tuning of innate immunity toward an antiviral response, a pathway that was also found to be enriched in hunter-gatherer ancestry in a highly admixed group. Other pathways presenting signals of polygenic adaptation include FoxO and Jak-STAT, which modulate a multitude of metabolic processes including innate immunity homeostasis, adiposity, energy expenditure and insulin sensitivity. Interestingly, we detected an enrichment in positive selection signatures across all hunter-gatherer groups in viral interacting proteins, principally for proteins interacting with double-stranded DNA viruses. Collectively, our analyses indicate that immune functions, particularly relating to antiviral functions, have participated to the genetic adaptation of rainforest hunter-gatherers to their unique ecologies, via several mechanisms involving classic sweeps and polygenic adaptation.

# A REFERENCE DATABASE AND DETECTION METHOD FOR ACTIVELY TRANSLATING ORFs FROM RIBO-SEQ DATA

Saket Choudhary*, Wenzheng Li*, Andrew D Smith

University Of Southern California, Quantitative and Computational Biology, Los Angeles, CA

(*equal contribution)

Ribo-seq, the technique for sequencing the ribosome protected fragments of mRNA, has enabled transcriptome-wide monitoring of translation, and has provided avenues for discovering tissue- or condition-specific open reading frames (ORFs). It has expanded the spectrum of translation beyond the annotated coding regions with the discovery of thousands of non-canonical ORFs. However, the detection of actively translating ORFs remains challenging due to high heterogeneity and noise in Ribo-seq data. The presence of amplification bias, non-ribosomal RNA-protein complexes or other contamination can often result in fragments that do not represent active translation. This can be misleading in the interpretation of Ribo-seq data, particularly for detecting short unannotated ORFs that are being actively translated.

The existence of multiple databases of processed public Ribo-seq datasets has provided avenues for testing novel hypotheses on publicly available data. However, there are currently two issues that remain unaddressed. First, the processed data made available can benefit from separating the actively translating fragments from the non-active ones. Second, it should be possible to search for Ribo-seq studies with similar profiles given a query dataset.

We introduce two innovations to the Ribo-seq community:

- RiboCop, a method to detect actively translating ORFs from Ribo-seq data
- ribopod, a database of uniformly processed and quality controlled Ribo-seq datasets.

RiboCop directly leverages the characteristic three-nucleotide periodicity using modified Fourier transform to detect actively translating ORFs and demonstrates unprecedented accuracy, robustness, and speed compared to other methods. The modified Fourier transform includes a normalization step to ensure that RiboCop is not biased towards coverage or length of the ORF as demonstrated using multiple published Ribo-seq datasets in human and mouse, thus making it particularly advantageous for detecting short ORFs.

Ribopod presents a key resource of public Ribo-seq datasets that have been uniformly processed through RiboCop to detect actively translating regions. It provides a multi-dimensional view to the Ribo-seq data at both metagene and transcript level with its dynamic visualization interface. Moreover, it can be used for discovery of non-canonical translating ORFs such as uORFs and stalling sites from multiple species across different conditions and tissues. The search interface of ribopod allows querying Ribo-seq datasets with similar profiles, thus providing a unique resource for discovering datasets exhibiting similar biological phenomena.

# CHARACTERIZING GENETIC AND TRANSCRIPT VARIATION USING LONG SINGLE MOLECULE SEQUENCING IN MAIZE

Kapeel M Chougule[1], Sharon Wei[1], Yinping Jiao[1], Bo Wang[1], Josh Stein[1], Jianing Liu[3], Brett Hannigan[4], Chai Fungtammasan[4], Victor Llaca[5], Kevin Fengler[5], Candice Hirsch[6], Matthew Hufford[7], Kelly Dawe[3], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Laboratory, Cold Spring Harbor, NY, [2]USDA-ARS-NEA, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, [3]University of Georgia, Department of Plant Biology, Athens, GA, [4]DNAnexus, Inc, Mountain View, CA, [5]Corteva Agriscience, Agriculture Division of DowDuPont, Johnston, IA, [6]University of Minnesota, Department of Agronomy and Plant Genetics, Saint Paul, MN, [7]Iowa State University, EEOB Department, Ames, IA

Maize is both an important crop and genetic model system, with high levels of genetic and functional diversity. Gene content can vary by more than 5% across lines, while as much as half of the functional genetic information lies outside of coding regions in highly variable and repetitive intergenic space. Characterization of this diversity has been confounded by the reliance of a single reference genome. Complete and accurate reference genomes from multiple individuals will be needed to characterize the maize pan-genome, and to advance the genetic and functional study of this crop. We describe progress on a new initiative to generate reference genomes for a core set of 26 highly researched inbreds known as the NAM founder lines, which represent a broad cross-section of modern maize diversity. Building off of previous success with generating the B73 reference genome, we describe a strategy for complete and accurate assembly of maize genomes using PacBio SMRT sequencing along with high-resolution Bionano optical. Here we characterize the large-scale structural diversity among 25 maize inbreds with the standalone use of Bionano optical maps. The pipeline involves pseudomolecule construction with de novo-assembled DLE maps and comparative genomic analyses. On the other hand, the long-range contiguity of optical maps enables high-confidence assemblies over the complex repeat-rich area. These high-quality assemblies will be annotated using a standard annotation protocol and gene trees built using the Ensembl compara pipeline, supporting the characterization of gene space structural variation. While these new technologies have greatly lowered the cost and improved quality of reference genome resources, challenges remain in managing and representing data to users.

We gratefully acknowledge funding from NSF award #1744001 and #1127112.

# PROPAGATING MASSIVE LOCAL RELATEDNESS MATRICES LOCUS-BY-LOCUS: PROBABILISTIC HAPLOTYPE SCREENING IN LARGE DATASETS VIA DOUBLY DECOUPLED HMMs

Ryan Christ[1,2], Louis Aslett[2,3], David Steinsaltz[4], Chris Holmes[2,4], Chris Spencer[5], Ira Hall[1]

[1]Washington University School of Medicine in St. Louis, McDonnell Genome Institute, St. Louis, MO, [2]Alan Turing Institute, London, United Kingdom, [3]Durham University, Mathematical Sciences, Durham, United Kingdom, [4]Oxford University, Statistics, Oxford, United Kingdom, [5]Oxford University, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

Approximating the recent phylogeny of N phased haplotypes at sequential loci along the genome is a core problem in modern population genomics and central to performing genome-wide screens for association, selection, introgression, and other signals. Standard window-based methods approximate the recent phylogeny at a given locus by calculating pairwise genetic distances between haplotypes. These distances form a locus-specific, N x N matrix that can then be screened for the signal of interest and removed from memory. This sequential, "test it and forget it" approach avoids having to store many large matrices simultaneously. While these methods can scale to datasets where N is in the hundreds of thousands, they present the trouble of determining window-size and overlap at adjacent loci. These approaches also fail to capture crucial aspects of haplotype block structure, such as staggered endpoints and the non-exchangeability of SNPs. Coalescent-inspired Hidden Markov Models (HMMs) have shown promise in producing locus-specific matrices that overcome these shortcomings; however, due to the inherent bidirectional dependency structure of HMMs, standard forward-backward decoding cannot evaluate loci sequentially, requiring the storage of an intractable number of large matrices. We present a novel HMM checkpointing approach and transformation of the Li & Stephens (LS) Model which together allow us to break this bidirectional dependency structure and obtain genetic distance matrices sequentially at every locus along the genome. This enables scalable "test it and forget it" screening while providing a more accurate representation of local haplotype structure than window-based methods. We show how our approach can be adapted to both sequence and microarray data using a novel parameter-tuning approach that aims to maximize the resolution with which the inferred distance matrices render the underlying phylogenies. These methods are all made easily accessible through our new R package, kalis, featuring a novel reformulation of the LS model which we accelerate with modern CPU vector intrinsics and an optimized cache structure. Finally, we discuss applications of this technology for imputation-free haplotype screening and our ongoing efforts to identify haplotypic complex trait associations in large genotype and genome sequence datasets.

# OUTLYING LEVELS OF DNA METHYLATION ARE ASSOCIATED WITH RARE GENETIC VARIATION AND ALTERED GENE-EXPRESSION

V. Kartik Chundru[1], James G Prendergast[2], Riccardo E Marioni[3,4], Allan J Beveridge[5], Grant W Montgomery[1], Ian J Deary[3], David A Hume[6], Naomi R Wray[1,7], Peter M Visscher[1,7], Allan F McRae[1]

[1]University of Queensland, Institute for Molecular Bioscience, Brisbane, Australia, [2]University of Edinburgh, The Roslin Institute, Edinburgh, United Kingdom, [3]University of Edinburgh, Department of Psychology, Edinburgh, United Kingdom, [4]University of Edinburgh, Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom, [5]University of Glasgow, Wolfson Wohl Cancer Research Centre, Glasgow, United Kingdom, [6]University of Queensland, Mater Research Institute, Brisbane, Australia, [7]University of Queensland, Queensland Brain Institute, Brisbane, Australia

While rare genetic variants have been demonstrated to play a role in an individual's risk to disease, detecting the effects from these variants requires very large sample sizes with whole genome sequencing. As it is not practical to sequence such large samples, an alternative approach is to investigate the role of rare genetic variation on low-level genomic traits where the effects of rare variants are expected to be substantially larger, assuming the variants act on trait variation through effects on gene-expression. In addition, rare epigenetic aberrations have been demonstrated to play a role in susceptibility of disease. We investigate DNA methylation (DNAm) in healthy ageing populations - the Lothian Birth cohorts of 1921 and 1936 - identifying both transient and stable outlying DNAm levels. We find an enrichment of rare genetic variants within 1kb of the DNAm sites in individuals with stable outlying DNAm, indicating underlying genetic control. Using a family based cohort, the Brisbane Systems Genetics Study, we observed increased sharing of DNAm outliers among more closely related individuals, consistent with these outliers being driven by rare genetic variants. Outlying DNAm levels are demonstrated to have a functional consequence on gene-expression, with extreme levels of DNAm being associated with gene-expression levels towards the tails of the population distribution. Overall, this study demonstrates the role of rare variants in the phenotypic variation of low-level genomic traits. By focusing on rare variants associated with outlying levels of DNAm in the regulatory regions of trait associated genes, we may be able to detect new disease associated variants not detectable by genome-wide association studies.

# PREDICTING THE STRUCTURE AND BUTYRATE PRODUCTION OF SYNTHETIC HUMAN GUT MICROBIOME COMMUNITIES

Ryan L Clark[1], David M Stevenson[2], Daniel Amador-Noguez[2], Ophelia S Venturelli[1]

[1]University of Wisconsin - Madison, Biochemistry, Madison, WI,
[2]University of Wisconsin - Madison, Bacteriology, Madison, WI

Metabolic activities arising from microbial community interactions in the human gut microbiome are major drivers of human health and disease. The ability to control metabolic outputs of microbial communities would enable the development of treatments for many microbiome-related diseases. Synthetic microbial communities can be used to connect organismal phenotype and genotype with behavior across various community contexts. In this work, we developed a predictive modeling approach using synthetic human gut microbiome communities to understand how production of the important bacterial metabolite butyrate is determined by microbial community structure. We first characterized growth and metabolite secretion of 25 genotyped human gut bacterial isolates, 5 of whose genomes encode metabolic pathways required for butyrate production. We then characterized the impact of pairwise interactions from these 25 species on growth and butyrate production of each of the 5 butyrate-producing bacteria. Using this initial dataset, we built a minimally-parameterized predictive model using tools from dynamic ecological modeling and machine learning to predict butyrate production by higher-order communities. This model identifies the pairwise interactions that have the largest impact on butyrate production in these synthetic communities. Moving forward, we plan to use active learning techniques with this model to design the most informative sets of experiments to improve the model's predictive capabilities to design highly productive higher-order butyrate-producing communities. The knowledge and methodologies generated in this work will be useful for designing interventions to treat patients with butyrate-deficient gut microbiomes as well as for designing microbial communities with specified functional outputs in general.

# mtDNA eQTLs AND THE $m^1A$ 16S rRNA MODIFICATION EXPLAIN mtDNA TISSUE-SPECIFIC GENE EXPRESSION PATTERN IN HUMANS

Tal Cohen, Chen Mordechai, Alal Eran, Dan Mishmar

Ben-Gurion University of the Negev, Life Sciences, Be'er Sheva, Israel

Expression quantitative trait loci (eQTLs) are instrumental in genome-wide identification of regulatory elements, yet were overlooked in the mitochondrial DNA (mtDNA). By analyzing 5079 RNA-seq samples from 23 tissues we identified association of ancient mtDNA SNPs (haplogroups T2, L2, J2 and V) and recurrent SNPs (mtDNA positions 263, 750, 1438 and 10398) with tissue-dependent mtDNA gene-expression. Since the recurrent SNPs independently occurred in different mtDNA genetic backgrounds, they constitute the best candidates to be causal eQTLs. Secondly, the discovery of mtDNA eQTLs in both coding and non-coding mtDNA regions, propose the identification of novel mtDNA regulatory elements. Third, we identified association between low $m^1A$ 947 MT-RNR2 (16S) rRNA modification levels and altered mtDNA gene-expression in twelve tissues. Such association disappeared in skin which was exposed to sun, as compared to sun-unexposed skin from the same individuals, thus supporting the impact of UV on mtDNA gene expression. Taken together, our findings reveal that both mtDNA SNPs and mt-rRNA modification affect mtDNA gene expression in a tissue-dependent manner.

*RECOUNT-BRAIN*: A CURATED REPOSITORY OF HUMAN BRAIN RNA-SEQ DATASETS METADATA

Ashkaun Razmara[1], Shannon E Ellis[2], Dustin J Sokolowski[3], Sean Davis[4], Michael D Wilson[3], Jeffrey T Leek[2], Andrew E Jaffe[2,5], Leonardo Collado-Torres[5]

[1]Quinnipiac University, Frank H. Netter MD School of Medicine, North Haven, CT, [2]Johns Hopkins Bloomberg School of Public Health, Biostatistics, Baltimore, MD, [3]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [4]National Cancer Institute, Center for Cancer Research, Bethesda, MD, [5]Lieber Institute for Brain Development, Data Science, Baltimore, MD

The usability of publicly-available gene expression data is often limited by the availability of high-quality, standardized biological phenotype and experimental condition information ("metadata"). We released the *recount2* project (DOI: 10.1038/nbt.3838), which involved re-processing ~70,000 samples in the Sequencing Read Archive (SRA), GTEx, and TCGA projects – while samples from the latter two projects are well-characterized, with extensive metadata, the ~50,000 RNA-seq samples in SRA have inconsistent metadata. While tissue type and sex can be estimated (DOI: 10.1093/nar/gky102), dataset and project-specific metadata, like age and diagnosis, must ideally be provided by labs that deposit the data.

In *recount-brain* we took a complementary approach to our prediction work by uniformly curating the metadata for human brain samples across about sixty studies whose expression data is present in *recount2*. The *recount-brain* resource is designed to facilitate re-analyses of human brain expression data and could serve to improve phenotype prediction algorithms. Our key results include:

* We uniformly curated and partially mapped to biomedical ontology terms 48 biological phenotype (sex, age, disease status and type) and experimental condition (library selection, brain bank) variables from 62 human brain SRA projects present in *recount2*.
* We merged GTEx and TCGA brain samples metadata into *recount-brain* to facilitate cross study comparisons. In total, *recount-brain* contains metadata for 6,547 samples out of which 5,330 (81.4%) have expression data in *recount2*.
* We extensively documented our curation process for reproducibility purposes (http://LieberInstitute.github.io/recount-brain/) and integrated *recount-brain* into the recount R/Bioconductor package (DOI: 10.18129/B9.bioc.recount) to facilitate its use and promote reproducibility of downstream analyses.

We envision that *recount-brain* will promote the re-use of smaller human brain RNA-seq studies, thus letting less “useful data go to waste” (DOI: 10.1038/543007a). Furthermore, this resource benefits from other recount2-related projects. Finally, *recount-brain* can serve as an exemplar for other curation-based projects.

# MUTATIONAL AND SELECTIVE PRINCIPLES OF STRUCTURAL VARIATION ACROSS HUMAN POPULATIONS

Ryan L Collins[1,2,3], Harrison Brand[1,2,3], Konrad J Karczewski[1,2], Xuefang Zhao[1,2], Jessica Alföldi[1,2], Amit V Khera[1,2,3], Laurent C Francioli[1,2], Laura D Gauthier[1,2], Harold Wang[1,2], Matthew Solomonson[1,2], Matthew R Stone[1,2], The gnomAD Consortium[1], Benjamin M Neale[1,2,3], Sekar Kathiresan[1,2,3], Mark J Daly[1,2,3], Eric Banks[1], Daniel G MacArthur[1,2,3], Michael E Talkowski[1,2,3]

[1]Program in Medical and Population Genetics, Data Science Platform, and Stanley Center, Broad Institute of MIT and Harvard, Cambridge, MA, [2]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, [3]Division of Medical Sciences, Harvard Medical School, Boston, MA

Structural variants (SVs) rearrange the linear and three-dimensional organization of the genome. Despite their significance in human evolution and disease, the mutational spectrum of SVs has not been routinely evaluated in medical and population genetic studies. Thus, the properties of SVs are incompletely understood. Here, we constructed a reference atlas of SVs from 14,891 deeply sequenced genomes spanning diverse populations (54% non-European) as a component of the Genome Aggregation Database (gnomAD). We discovered 498,327 distinct SVs, including 5,883 multi-breakpoint complex SVs across 13 mutational subclasses, and examples of localized chromosome shattering, like chromothripsis. SVs were non-uniformly distributed across the chromosomes: CNVs were enriched near telomeres and centromeres, whereas balanced and complex SVs were depleted in these same regions. Likewise, mutation rate estimates varied substantially by SV class, with CNVs appearing more mutable than other SVs. Signatures of selection were strongest against inversions and complex SVs, which appeared to be attributable to both coding and noncoding effects. We found strong correlations between constraint against damaging point mutations and rare SVs that both disrupt and duplicate protein-coding sequence, suggesting selection against dosage alterations for a subset of tightly regulated genes. The average genome harbored 8,202 SVs, including eight rare, gene-altering SVs, and we predicted that SVs constitute at least 25% of all rare loss-of-function events per genome. We further observed that 3.7% of genomes harbor rare SVs larger than 1 Mb, half of which are balanced or complex. Finally, we performed analyses to demonstrate the uses of a high-quality SV resource to augment human disease association studies, to redefine odds ratios of genomic disorders, and to estimate the yield of clinically reportable findings from SVs via genome sequencing. The full gnomAD SV map has been directly integrated into the existing gnomAD browser (https://gnomad.broadinstitute.org), which will allow broad, hands-on access to these results as a resource for medical and population genetics.

# THE REGULATORY LANDSCAPE UNDERLYING EPITHELIAL REGENERATION IN *DROSOPHILA*

Elena Vizcaya-Molina[1], Cecilia C Klein[1,2], Carlos Camilleri-Robles[1], Raziel Amador-Rios[1,2], Florenci Serras[1], Roderic Guigo[1,2,3], Montserrat Corominas[1]

[1]Universitat de Barcelona, Genètica, Microbiologia i Estadística, Barcelona, Spain, [2]Centre for Genomic Regulation, Barcelona, Spain, [3]Universitat Pompeu Fabra, Barcelona, Spain

During the process of regeneration a switch in the transcription program occurs in the cells that contribute to reconstruct the missing tissue. Early signals released upon damage are integrated into the chromatin of responding cells to change its activity and function. Changes in chromatin dynamics result in transcriptional reprogramming, this is the coordinated regulation of expression of a specific subset of genes required for the regeneration process. Using genome-wide approaches to interrogate genome activity and chromatin function, we identify the regulatory elements governing tissue recovery in *Drosophila* imaginal discs, which show a high regenerative capacity after genetically induced cell death. Our findings show a burst of transcription and a global co-regulation of gene expression; genes induced early on tend to be located close to one another in the linear genome. Upon injury, a subset of damage-responsive long non-coding RNAs (lncRNAs) is not only activated but also required for proper regeneration. In addition, novel enhancers acting exclusively in the damaged tissue cooperate with enhancers co-opted from other tissues and developmental stages, and with endogenous enhancers that show increased activity after injury. These enhancers host binding sites for regulatory proteins, including a core set of conserved transcription factors that control regeneration across metazoans. The coordinated action of lncRNAs, transcription factors and regeneration enhancers results in the regulatory landscape underlying the transcriptional reprogramming of regeneration.

# GWAS ON LONGITUDINAL GROWTH TRAITS REVEALS DIFFERENT GENETIC FACTORS INFLUENCING INFANT, CHILD AND ADULT BMI

Alexessander Couto Alves[1,2], Vincent Jaddoe[3], Struan Grant[4], Sylvain Sebert[1,5], Mark I McCarthy[6], Marjo-Riitta Järvelin[1,5]

[1]Imperial College London, School of Public Health, London, United Kingdom, [2]Surrey University, School of Biosciences and Medicine, Guildford, United Kingdom, [3]Erasmus MC, University Medical Center, The Generation R Study Group, Rotterdam, Netherlands, [4]The Children's Hospital of Philadelphia, Division of Human Genetics, Philadelphia, PA, [5]Oulu University, Center for Life Course Health Research, Oulu, Finland, [6]Oxford University, Wellcome Centre for Human Genetics, Oxford, United Kingdom

Early childhood growth patterns are associated with adult health, yet the genetic factors and the developmental stages involved are not fully understood. We performed meta-analyses of genome-wide association studies (GWAS) in up to 22,769 children of European ancestry covering six early growth traits, modelled using longitudinal data for height and weight collected from birth to 13 years. We considered the well-established traits: peak height and weight velocity, age and body mass index (BMI) at adiposity peak (Age-AP, BMI-AP) and at adiposity rebound (Age-AR, BMI-AR). GWAS results were followed up for functional, pathway, genetic correlation and genetic risk score analyses to determine how developmental timings, molecular pathways and genetic determinants of these traits overlap with those of adult health.

We identified three variants significantly associated ($P<5x10-8$) with AR traits (rs1421085 (FTO), rs2817419 (TFAP2B) and rs10938397 (GNPDA2)), all of which had previously been associated with adult BMI, obesity and waist circumference. Both Age-AR and BMI-AR had robust genetic correlations with adult BMI, waist circumference, body fat percentage and multiple cardio-metabolic traits. However, AP traits were distinct: neither Age-AP nor BMI-AP are associated with adult BMI traits, and adult BMI-associated variants (and the corresponding genetic risk score) showed no association with Age-AP and BMI-AP. We identified an entirely novel association between variants (led by rs9436303) near the leptin receptor gene and BMI-AP, and showed that this variant colocalized with eQTLs for both LEPR and LEPROT, and also influences circulating soluble leptin receptor levels in adults. Loci associated with AP and AR traits colocalized with variants regulating gene expression in different adult tissues.

These findings suggest that different genetic factors control infant and child BMI. In light of the obesity epidemic, these data are important to inform the timing and targets of prevention strategies.

# *IN-SILICO* EDGETIC PROFILING OF POPULATION-SPECIFIC VARIATION IN HUMAN GENOME

Hongzhu Cui[1], Anastasia Leshchyk[1], Suhas Srinivasan[2], Dmitry Korkin[1,2,3]

[1]Worcester Polytechnic Institute, Bioinformatics and Computational Biology Program, Worcester, MA, [2]Worcester Polytechnic Institute, Data Science Program, Worcester, MA, [3]Worcester Polytechnic Institute, Computer Science Department, Worcester, MA

In spite of the recent efforts in population genetics, medical genomics has focused so far primarily on the people of European descent. Yet, global projects such as the 1000 Genomes Project have revealed intrinsic differences in genetic variation across different ethnic groups. How much of these differences attribute to population-specific traits? To answer this question, one needs to translate the mutation data into functional implications. Recently, the "edgotype" concept was proposed, reflecting the ability of genes and their protein products to interact with each other, forming intricate networks. Unlike the traditional gene-centric paradigm, edgetic network models emphasize the disruption of specific edges (protein-protein interactions, PPIs) caused by non-synonymous variants. Unfortunately, high-throughput experimental methods for edgetic profiling remain costly and time-consuming.

In this work, we create a comprehensive catalog of population-specific edgetic effects at the whole-interactome level. Our work leverages a recently developed novel machine learning approach, SNP-IN tool, that determines interaction-rewiring effects of non-synonymous single nucleotide variants (nsSNVs). The method was applied to more than 11,000 variants associated with diseases, showing feasibility of a large-scale *in-silico* edgetic studiy and revealing insights into the orchestrated play of mutations associated with a complex disease. To compare the functional impact of population-specific variations, we have applied our approach to edgetically annotate an unprecedented set of 97,145 nsSNVs collected from the 1,000 Genomes Project by leveraging the structural information on PPI complexes. The functional impact of a variant on a PPI is annotated as neutral (no change to PPI), disruptive (loss of PPI), or beneficial (increased binding affinity of PPI). We determined that 38,872 SNVs (40.0%) were predicted as disruptive to at least one PPI, indicating that a significant number of nsSNVs from the normal populations can alter protein function. We showed that genes enriched with disruptive mutations obtained from the healthy populations are in fact associated with diverse functions and are implicated in various diseases. Our preliminary analysis shows that edgetic profiles of genes with a high number of disruptive mutations are distinct from each other among 5 major populations. Finally, the difference of the accumulated damage in disease-associated modules suggests population-specific disease susceptibility.

We expect that our approach will provide a more accurate and in-depth characterization of the functional consequences of ethnic-specific alleles, leading to a better understanding of the clinical and phenotypic outcome.

# SINGLE-CELL RNA-SEQUENCING OF DIFFERENTIATING IPS CELLS REVEALS DYNAMIC GENETIC EFFECTS ON GENE EXPRESSION

Anna Cuomo[1], Daniel Seaton[1], Davis J McCarthy[1,4], Iker Martinez[2], Marc Jan Bonder[1,3], Jose Garcia-Bernardo[2], Shradha Amatya[2], Ludovic Vallier[2,7,8], John C Marioni[1,2,5], Oliver Stegle[1,3], Mariya Chhatriwala[2,8]

[1]EMBL-EBI, European Bioinformatics Institute, Cambridge, United Kingdom, [2]Wellcome Trust Sanger Institute, Cellular Genetics, Cambridge, United Kingdom, [3]EMBL, Genome Biology, Heidelberg, Germany, [4]St Vincent Institute, Victoria, Melbourne, Australia, [5]Cancer Research UK, Cambridge Institute, Cambridge, United Kingdom, [6]German Cancer Research Center (DKFZ), Computational Genomics, Heidelberg, Germany, [7]Cambridge Institute, Stem Cell, Cambridge, United Kingdom, [8]University of Cambridge, Department of Surgery, Cambridge, United Kingdom

Recent developments in stem cell biology have enabled the study of cell fate decisions in early human development that are impossible to study in vivo. However, the impact of common genetic variants upon development has not been well characterised. Here, we exploit human induced pluripotent stem (iPS) cell lines from 125 donors, and establish a co-culture system to differentiate 4-6 lines towards endoderm. To understand the interplay between common genetic variants and cell differentiation, we characterize the transcriptional dynamics across the differentiation trajectory using single-cell RNA-sequencing, assaying a total of 36,044 cells using comprehensive full-length transcriptome profiling.

Our data reveal three major stages of differentiation, iPSC, mesendoderm and definitive endoderm, in each of which we map thousands of expression quantitative trait loci (eQTL), including 465 novel eQTL for key developmental genes that have not previously been identified. We rigorously compare single cell QTL mapping with results obtained from bulk RNA-sequencing applied to the same samples, highlighting technical advantages and challenges in using single-cell approaches and co-culture assays.

Beyond the analysis of eQTL in rigidly defined cell types, our study provides an opportunity to analyse the impact of common genetic variants on the differentiation dynamics. By combining a computationally reconstructed pseudo time with allele-specific gene expression quantification we characterise the specific stages of differentiation in which individual eQTL are active, identifying dynamic and compensatory patterns of allele-specific expression changes throughout development. We identify and experimentally validate the functional relevance of these dynamic eQTL, including novel genetic factors that are predictive of variability in differentiation capacity between lines.

# GENOME-WIDE MAPPING OF DNA METHYLATION VARIANTS AFFECTING GENE EXPRESSION LEVELS IN GLIOMAS WITH RESPECT TO THEIR GRADE AND IDH GENE MUTATION STATUS

Michal J Dabrowski[1], Agata Dziedzic[1], Rafał Guzik[2], Michal Draminski[1], Bartosz Wojtas[2], Karolina Stępniak[2], Bartłomiej Gielniewski[2], Jacek Koronacki[1], Bozena Kaminska[2]

[1]Institute of Computer Science, Computational Biology Lab, Warsaw, Poland, [2]Nencki Institute of Experimental Biology, Laboratory of Molecular Neurobiology, Warsaw, Poland

Here we report on differentially methylated promoters in gliomas of different histopathological WHO grades and *IDH* gene mutation status. To obtain methylomes for 20 samples we used SeqCap Epi CpGiant Methylation panel and performed bisulphite conversion followed by Illumina NGS sequencing. For these samples RNA-seq was also performed. In order to analyze methylomes we developed CytoMeth tool based on the Roche pipeline. Our tool compiles a set of open source software enabling fast and transparent rough data processing. The sum of unique CpG sites across all samples was above 5 mln, while the common part constituted ~ 300,000. We analysed genomic regions +/- 2kB from TSS that did not differ in the number of cytosines between samples subgroups. For these cytosines in each of over 34,000 promoters we computed: max, mean, SD, median of B-values, as well as number of CpG sites, % of highly methylated, medium and low methylated cytosines. To select genes whose promoter differs in any characteristic between pilocytic astrocytoma (PA) and other glioma subtypes (diffuse astrocytoma and glioblastoma, hereafter HGG), we used two approaches: (i) Wilcoxon test with false discovery rate correction and (ii) MCFS-ID algorithm which performs non-linear range specific feature selection and discovers interdependencies between features. As biologically relevant and not age-specific, we selected genes that differed in expression between PA and HGG and at the same time did not differ between PA and adult normal brain samples (NB) and did differ between HGG and NB. That pipeline was applied for cytosines intersected with promoters with- and without-respect to the strand, which returned two sets: StSp, NotStSp, respectively. For these sets we performed unsupervised text clustering based on gene descriptions from KEGG and REACTOME and verified whether genes were up- or down-regulated.

There were 70% of genes common for StSp and NotStSp, in consequence clusters corresponding to each other were characterized by similar key-words but only in the StSp set there was gene cluster associated with ABC transporters. Homogenous clusters, where all genes were up-regulated in HGG, were described as follows: [1] homeobox, hoxc, morphogenesis; [2] development, embryonic, differentiation; [3] mitotic, cell cycle, repair, G protein, replication; [4] transcription. Moreover, clusters with majority of up-regulated genes in PA were described by: [5] synaptic, adhesion, membrane, receptor, neurotransmitter, protein tyrosine phosphatases; [6] positive/negative regulation, apoptotic process. It is worth to mention that we also discovered a group of antisense genes i.e. *PLCE1-AS1* whose promoter was highly methylated in PA indicating a multistage regulatory mechanism of its protein coding gene expression.

Supported by the NCN grant 2015/16/W/NZ2/00314.

FUNCTIONAL ANNOTATION OF lncRNAs IN THE HUMAN BRAIN

Anupriya Dalmia[1,2], Natalia Savytska[1], Salvador Rodriguez-Nieto[1], Tenzin Nyima[1], Noemia-Rita Alves-Fernandes[1], Stefan Bonn[1,3], Melissa G Castillo Lizardo[1], Peter Heutink[1,2]

[1]German Centre for Neurodegnerative Diseases, Deparment of Genome Biology, Tuebingen, Germany, [2]RIKEN, Division of Genomic Medicine, Yokohama, Japan, [3]University Medical Center Hamburg-Eppendorf, Department of Medical Systems Biology, Hamburg, Germany

The functional relevance of long non-coding RNAs (*lncRNAs*) is still largely uncharacterized. As part of the FANTOM6 (Functional Annotation of the Mammalian Genome, *http://fantom.gsc.riken.jp/*) consortium, this study aims to characterize the role of lncRNAs in neurodegenerative diseases and aging. The lncRNAs to be investigated have been selected based on several criteria: (i) significant differential expression between cases and controls for Frontal Temporal Dementia (FTD), (ii) antisense transcripts to genes associated to neurodegenerative diseases (Illumina TruSeq Panel), (iii) presence of eQTLs overlapping GWAS loci for neurodegenerative traits and (iv) lncRNAs showing significant differential expression with increasing cognitive decline i.e., from no cognitive impairment (NCI) to mild congnitive impariment (MCI) to Alzheimer's disease (AD) in the ROSMAP (PMID: 22471860) cohorts. Using LNA-modified antisense oligonucleotides (ASOs), we systematically knock down lncRNAs of biological interest in iPS cells that are differentiated into cortical neurons. We measure the effect of the knockdowns on (i) cell growth phenotype and (ii) molecular phenotype using Cap Analysis of Gene Expression (CAGE) sequencing followed by pathway analyses. We capture true biological effect by correcting for cell-type changes using deconvolution networks. We have identified two sets of 41 and 22 lncRNAs each that show a steady decrease and increase, respectively, from NCI to MCI to AD. These gene sets are enriched for biological processes such as synaptic vesicle cycle, innervation, synaptic vesicle transport, neuronal precursor cell proliferation, regulation of signal transduction, among other relevant pathways. Overall, 122 lncRNAs were identified with evidence for a regulatory role in neurodegenerative diseases. The study shows how CAGE trancriptome profiling is useful in uncovering molecular phenotypes induced by lncRNA depletion and is an efficient aid in validating biological explanations of lncRNA function.

# IDENTIFYING AND FUNCTIONALLY CHARACTERIZING THE PROXIMATE EFFECTS OF FETAL-SPECIFIC CARDIAC REGULATORY VARIANTS ON MOLECULAR PHENOTYPES

Matteo D'Antonio, Margaret K Donovan, William W Young Greenwald, David Jakubosky, Paola Benaglio, Erin N Smith, Agnieszka D'Antonio-Chronowska, Kelly A Frazer

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

To fully understand the contributions of genetic regulatory variants to human health and disease, it is essential to characterize their proximate effects in both adult and fetal developmental stages. Here, we used iPSCORE, a collection of iPSC lines generated from over two hundred individuals, to characterize molecular phenotypes (RNA-seq, ATAC-seq, ChIP-seq for H3K27ac) in 180 iPSC-derived cardiovascular progenitor cell (iPSC-CVPC) samples. Comparing the iPSC-CVPC transcriptomes and epigenomes to GTEx and Roadmap showed that they have expression profiles similar to fetal cardiac tissue. We used single cell RNA-seq to examine eight iPSC-CVPCs and found they were comprised of two distinct cell types, which showed both transcriptional and chromatin profiles consistent with being cardiomyocytes (CMs) and epicardium derived cells (EPDCs). We established and used CM-specific and EPDC-specific expression signatures to deconvolute the bulk RNA-seq data for all 180 iPSC-CVPCs, determining the relative ratios of both cell types in all samples. We have leveraged these deconvoluted expression data in combination with whole-genome sequences to perform expression quantitative trait loci (eQTLs), resulting in the discovery of cell-type specific and developmental stage (fetal) specific gene expression regulatory variants that are associated with cardiac disease.

We normalized the expression of 16,156 expressed genes, adjusted them for multiple covariates (age, sex, ancestry and cell population), performed an eQTL analysis on single nucleotide variants, small indels and structural variants detected using a combination of different software tools (GATK, LUMPY, CNVnator, Genome STRiP, MELT and HipSTR), and detected 3,163 genes with significant eQTLs (eGenes). Using an interaction test on the cell population distribution, we observed that 210 eGenes were associated with the cell-type, indicating that their eQTLs have different functions between CMs and EPDCs. We next investigated if fetal-CM and fetal-EPDC eQTLs provide the same associations between genetic variation and gene expression as adult cardiac tissue eQTLs. We performed a colocalization analysis on all eGenes with eQTL data from GTEx cardiac tissues (atrium, ventricle, aorta and coronary arteries), and found that 1,570 (49.6%) eGenes were not shared with adult tissues. To assess whether fetal-specific eQTLs are associated with complex cardiac traits measured in adult individuals, we colocalized eQTLs with summary statistics from cardiac GWAS (pulse rate and coronary artery disease) and found 25 fetal-specific eGenes associated with cardiac disease traits. These results show that analysis of the eQTLs in iPSC-CVPCs identifies cardiac disease GWAS regulatory variants that are active in the fetal heart, but not in adult heart tissues, indicating that they play a role in development.

# HOW SPECIFIC IS TUMOR-SPECIFIC SPLICING?

Julianne K David[1,2], Sean K Maden[1,2], Abhinav Nellore[1,2,3], Reid F Thompson[12,45,67,8]

[1]Oregon Health & Science University, Computational Biology Program, Portland, OR, [2]Oregon Health & Science University, Department of Biomedical Engineering, Portland, OR, [3]Oregon Health & Science University, Department of Surgery, Portland, OR, [4]Oregon Health & Science University, Department of Radiation Medicine, Portland, OR, [5]Portland VA Research Foundation, Portland, OR, [6]VA Portland Healthcare System, Division of Hospital and Specialty Medicine, Portland, OR, [7]Oregon Health & Science University, Department of Medical Informatics & Clinical Epidemiology, Portland, OR, [8]Oregon Health & Science University, Cancer Early Detection Advanced Research Center, Portland, OR

Tumor-specific neoepitopes arising from aberrant RNA splicing are potentially valuable cancer immunotherapy targets, as these peptide sequences differ significantly from those normally generated and presented by cells and are likely to be recognized as non-self by the immune system---more likely to induce an immunotherapeutic response, and less likely to trigger dangerous and unwanted side effects. To improve the robustness of "normal" splicing comparisons, we have pursued an extensive analysis that assesses tumor splicing from The Cancer Genome Atlas RNA-seq samples against all normal splicing found in Genotype Tissue Expression project (GTEx) samples and incorporates junctions from across over 50,000 RNA-seq samples from the Sequence Read Archive (SRA). We found that some normal cell types are largely unrepresented in GTEx but may be important cells of origin for some cancers, such as melanocytes and thymus tissue. We also extracted junctions from SRA developmental tissues, such as placenta, oocyte, and embryo, and from stem cells. We found that many aberrant splice sites thought to lead to splicing neoepitopes are actually found in other normal tissues, and that the total population of non-normal neoepitopes is greatly reduced under the strict filter that junctions not be found in any GTEx samples at all. Additionally, different subtypes of cancer are differentiated by sets of subtype-specific aberrant splice sites, and for some cancer types, this is based on tumor-specific splicing patterns having strong similarities to those of their cell types of origin. For instance, highly shared skin and uveal melanoma junctions are widely present in melanocyte samples on the SRA. Finally, we find that many tumor-specific junctions appear to be reactivation of embryological and developmental pathways. Nonetheless, there remains a small, refined collection of cancer-specific splice sites which may yield real tumor-specific neoepitopes.

# ANALYZING GENETIC DIVERSITY IN MARMOSET COLONIES

Ricardo C del Rosario[1,2], Qiangge Zhang[1,3], Curtis J Melo[1,2], Xian Gao[3], Monika A Burns[4], Fenna Krienen[1,2], Guoping Feng[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]Massachusetts Institute of Technology, McGovern Institute for Brain Research, Cambridge, MA, [4]Massachusetts Institute of Technology, The Division for Comparative Medicine, Cambridge, MA

The common marmoset is a small anthropoid primate (300-400g), reaches sexual maturity in 12-15 months and thus is increasingly being used as a model organism to study psychiatric disorders. Maintaining the genetic diversity of marmosets in various colonies around the world is therefore an important goal. We analyzed the genetic diversity of the marmoset colonies at the Broad Institute and at MIT by performing whole genome sequencing on 81 animals. The animals originated from three colonies: 36 from the New England Primate Research Center (NEPRC), 18 from a company (Company A), and 28 from CLEA Japan, Inc. (CLEA). We initially discovered 17 million SNPs and after applying quality filters and pruning for linkage disequilibrium, we obtained 147,000 independent SNPs. A principal component analysis of the SNP genotypes showed that the colonies are genetically distinct. The average heterozygosity of all marmosets is 1.03 per Kbp which is slightly higher than the average heterozygosity of humans in Africa. However, the heterozygosity of the three colonies vary, with NEPRC the highest at 1.3 per Kbp, followed by Company A (1.03 per Kbp) and CLEA (0.79 per Kbp). We also computed the Fst, a measure of structure in populations, and obtained a value of 0.12. This indicates that 88% of the total genetic variation is shared among colonies. Our results show that genotyping marmosets from other colonies will allow a better understanding of the genetic landscape of captive marmosets and could aid in their breeding strategies.

# A PAN-CANCER TRANSCRIPTOME ANALYSIS REVEALS PERVASIVE REGULATION THROUGH TUMOR-ASSOCIATED ALTERNATIVE PROMOTERS

Deniz Demircioğlu[1], Martin Kindermans[1], Engin Cukuroglu[1], Tannistha Nandi[1], Claudia Calabrese[3], Nuno A Fonseca[3], Andre Kahles[4,5], Kjong Lehmann[4,5], Oliver Stegle[3], Alvis Brazma[3], Angela N Brooks[6], Gunnar Rätsch[4,5], Patrick Tan[1,2], Jonathan Göke[1,7]

[1]Genome Institute of Singapore, Computational and Systems Biology, Singapore, Singapore, [2]Duke NUS, Medical School, Singapore, Singapore, [3]EMBL-EBI, Hinxton, United Kingdom, [4]ETH Zurich, Computer Science, Zurich, Switzerland, [5]Memorial Sloan Kettering Cancer Center, New York, NY, [6]University of California, Santa Cruz, CA, [7]National Cancer Centre, Singapore, Singapore

Cancer is a disease of the genome where alterations in the DNA lead to uncontrollable cell proliferation and division. These modifications in a cell's behavior are reflected at the transcriptome level. Transcriptional regulation, whose central element is the promoter, is responsible for controlling these changes in the expression. International consortia efforts such as TCGA and the ICGC produced vast amounts of publicly available RNA-Seq data to map changes in the cancer transcriptome. However due to lack of ChIP-Seq and CAGE data the role of the promoters in controlling transcriptional changes in cancer is still mostly unexplored.
Here, we developed a framework for estimating promoter activity from RNA-Seq data and used this framework to study the transcriptional regulatory changes that are associated with cancer. We have analyzed 18,468 samples in total covering 42 cancer types including 1188 patient samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG), 11,251 samples from the PanCanAtlas and 6674 samples from the GTEx projects. We demonstrated that our approach accurately identifies active promoters by comparing our promoter activity estimations with H3K4me3 data from the ENCODE project. We found hundreds of tissue specific alternative promoters that are not observable at the gene expression level. Furthermore, we identified promoters with significant activity changes in cancer compared to normal samples for individual cancer types. We examined the associations between highly dynamic active promoter landscape and patterns of somatic mutations as well as patient survival.
In summary, we showed that promoter activity can be estimated using RNA-Seq data and used this approach to identify cancer associated alternative promoters for 42 cancer types. We anticipate that the promoter activity estimation using RNA-seq data will broaden our understanding of the promoters' role in cancer by enabling the use of widely available RNA-Seq data. Furthermore, the catalogue of cancer associated promoters identified here will be a useful resource to uncover the regulatory and transcriptional changes in cancer.

# THE INTERPLAY BETWEEN 3D GENOME ORGANIZATION AND GENETIC VARIATION

Olivier Delaneau[1,2], Mariana Zazhytska[2], Christelle Borel[2], Guillaume Rey[1], Giuliana Giannuzzi[2], Stylianos Antonarakis[1], Alexandre Reymond[2], Emmanouil Dermitzakis[1]

[1]University of Geneva, Genetics and Development, Geneva, Switzerland, [2]University of Lausanne, CIG, 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 here we aim at bridging both by assembling a dataset combining the activity of regulatory elements (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 Fibroblasts).
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). In addition, we also find 25,315 significant associations between CRDs located on distinct chromosomes that form 30 Trans Regulatory Hubs (TRHs) and reflect the global chromatin organization into A/B nuclear compartments. Second, we show that CRDs and TRHs delimit the sets of active regulatory elements controlling the expression for most genes (=82.4%), vary substantially across cell types and are key contributors of cis and trans gene co-expression.
We show that CRDs are under strong genetic control. We discovered 58,968 chromatin peaks affected by nearby genetic variants (cQTLs). At the CRD level, this converts into genetic control of the activity of 6,157 CRDs and the structure of 110 CRDs. In practice, we estimated that 75% of these are eQTLs and showed that CRDs and TRHs capture complex regulatory networks into which the effects of eQTLs are propagated and combined to affect gene expression. Finally, we expand the analysis to three additional blood cell types where not only we replicate the signals observed above but we estimate attributes of cell-type specificity. We then use this level of organization to develop a hierarchical genomics approach to reveal sets of rare and common genetic variants that jointly affect common phenotypes and diseases.
Our analysis reveals population attributes of the 3D genome organization, that is under strong genetic control, exhibits substantial variability and can reveal novel insights to the interplay with genetic variants to determine phenotypic variability.

# STARSOLO: SINGLE-CELL RNA-SEQ ANALYSIS BEYOND GENE EXPRESSION.

Daniel Blibaum, Alexander Dobin

Cold Spring Harbor Laboratory, Genome Center, Cold Spring Harbor, NY

STARsolo is a comprehensive ultra-fast turnkey solution for analyzing droplet single cell sequencing data (e.g. Drop-seq, 10X Genomics Chromium System), built upon the popular RNA-seq aligner STAR. STARsolo takes raw FASTQ reads files as input, and performs the following operations: (i) demultiplexing of reads from different cells using well-established cell barcode whitelisting schemes; (ii) sequencing error correction and collapsing of Unique Molecular Identifiers to de-duplicate reads amplified from the same RNA molecules; (iii) mapping the reads to the reference genome using the standard STAR spliced read alignment algorithm; (iv) quantification of per-cell gene expression by counting the number of reads per gene.

The STARsolo output is designed to be a drop-in replacement for 10X CellRanger gene quantification output and produces nearly identical gene counts in the same format.
Unlike other existing scRNA-seq analysis pipelines, the cell barcode demultiplexing, UMI collapsing, mapping and quantification are integrated into a single code and are performed simultaneously. This avoids input/output bottlenecks and results in massive (5~10 fold) boost in processing speed and a significant reduction in analysis complexity.

In addition to transcript/gene expression evaluation, STARsolo is capable of detecting alternative and novel splicing, novel transcription termination sites, unannotated transcript/genes, and chimeric (fusion) junctions. Detection of these novel transcriptomic features in rare cell types may enable discovery of new biological phenomena unobservable in bulk RNA-seq data. We have used STARsolo to re-analyze the data from two massive single-cell sequencing datasets: Tabula Muris and the 10X Genomics Megacell mouse brain dataset. We have identified thousands of cell-type specific alternative splicing and polyadenylation events, as well as hundreds of previously unannotated genes and isoforms.

# THE CHROMATIN ACCESSIBILITY INFORMATION LANDSCAPE OF HUMAN TRANSCRIPTION FACTOR-DNA INTERACTIONS

Ricardo D'Oliveira Albanus[1], Yasuhiro Kyono[1,2], John Hensley[1], Jacob Kitzman[1,2], Stephen C.J. Parker[1,2]

[1]University of Michigan, Computational Medicine & Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Human Genetics, Ann Arbor, MI

Spatial and temporal gene expression dynamics that lead to biological diversity are orchestrated by transcription factors (TFs), and the DNA binding site motifs they recognize are enriched to overlap disease- and trait-associated genetic variations. Understanding transcriptional regulatory mechanisms is, therefore, a central problem in biomedical research. Chromatin accessibility assays, like ATAC-seq, reveal diverse information about *in vivo* regulatory landscapes, including TF binding patterns. Local properties of chromatin architecture are captured in ATAC-seq fragment lengths in the form of short TF-flanking or longer nucleosomal-sized units. We developed a novel information-theory approach to quantify such positional fragment length patterns in ATAC-seq data, allowing us to measure the impact of TFs on local chromatin organization. This new measure of chromatin accessibility information around TF motifs correlates with nucleosome positioning measured in previous MNase-seq experiments. Application of our method to a collection of 540 non-redundant TF motifs revealed at least two general configurations of chromatin accessibility information: (i) binding at the nucleosome dyad or (ii) binding at nucleosome-depleted regions. We show that the TFs with the highest levels of chromatin accessibility information bind between two well-positioned nucleosomes, suggesting that they induce changes in the local chromatin architecture. We extend our computational approach to quantify asymmetry in chromatin accessibility information around TF binding sites and observe significant asymmetry (Bonferroni adjusted $p < 0.05$, permutation test) for 150 of 540 tested TF motifs, including ELF1 (Z-score = 8.6, permutation test), which agrees with a recent study. Using previous FRAP data, we found that our measure of chromatin accessibility information is a proxy for TF-DNA binding dynamics. As one example, we found that binding sites for CTCF co-bound with cohesin have higher chromatin accessibility information compared to CTCF alone, which agrees with a recent study that reported 10-fold slower DNA binding rates for cohesin compared to CTCF. We observed that ATAC-seq allelic imbalance is significantly shifted towards alleles that are predicted to increase local chromatin accessibility information in GM12878 ($p = 2e\text{-}6$, Wilcoxon test) and human pancreatic islets ($p = 2e\text{-}7$, Wilcoxon test), suggesting that overall chromatin accessibility is driven by TFs with a high capacity to organize local chromatin architecture. We estimated that TFs with high chromatin organizational capacity account for approximately 20% of all TFs, and include lineage-determining TFs. Collectively, our work provides novel insights about the information content of chromatin organizational patterns and the impact these have on gene regulation dynamics.

# CHARACTERIZING THE GENOMIC DETERMINANTS OF CIS-REGULATORY ELEMENT EVOLUTION WITH PHYLOGENETIC MODELS

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

[1]CSHL, SCQB, Cold Spring Harbor, NY, [2]Penn State, Biology Department, State College, PA

With the proliferation of high-throughput functional genomics assays (ie. ChIP-seq, ATAC-seq, PRO-seq, etc.) there has been an explosion in the amount of available epigenetic data within and between multiple 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 (i.e. phastCons, GERP++, and dLess) have been developed to interrogate the forces that shape the evolution of genomic sequences and have been useful in identifying loci associated with disease, development, and molecular phenotypes. However, no similar methods exist for analyzing comparative epigenomic data. Here we propose a new phylogenetic hidden Markov model, epi-Phylo, for reconstructing the evolutionary histories of enhancers and promoters from comparative ChIP-seq data.

Epi-phylo combines a phylo-HMM with a negative binomial error model to jointly infer the location of cis-regulatory elements (CREs) and their evolutionary trajectories from noisy data. We apply epi-Phylo to existing datasets and show that sequence, epigenetic conservation, and stability of gene expression across tissues are positively correlated. To test for groups of CREs under differential selection we develop and apply a second method which uses the probabilistic allele labels from epi-Phylo without the HMM to derive element class specific estimates. We investigate the effects of various genetic covariates to on turnover rate and find that distance and find significant effects based on distance of CRE from target gene, expression of target gene, and variance of gene expression across tissues. We also find a relationship between proxies for dosage sensitivity and turnover rate and propose that dosage sensitivity of target genes can explain the discrepancy between sequence and histone mark turnover rates of associated CREs. In conclusion, epi-Phylo, and its post-hoc method, phylo-GLM, are rigorous statistical phylogenetic models of epigenomic evolution, enabling the inference of explicit evolutionary histories for a variety of regulatory elements.

# USING CITIZEN SCIENCE TO IDENTIFY CATTLE FOR A CHANGING CLIMATE.

Harly J Durbin[1], Troy N Rowan[1], Helen Yampara-Iquise[1], Robert D Schnabel[1,2], Jared E Decker[1,2]

[1]University of Missouri, Division of Animal Sciences, Columbia, MO, [2]University of Missouri, Informatics Institute, Columbia, MO

Beef cattle are subject to a diverse range of environmental stressors compared to other livestock production systems, creating greater opportunity for economic loss. For example, losses associated with heat stress total over $87 million a year and will likely rise in the face of future climate change. Fescue toxicosis, a major biotic stressor, is caused by the consumption of fescue forage infected with the endophytic fungus *Neotyphodium coenophialum* and costs the U.S. beef industry an estimated $1 billion a year through losses in reproductive success, growth, and milk production.

Though no cattle are completely resistant to fescue toxicosis, some are more susceptible than others. Fortunately, susceptibility can be quantified since fescue toxicosis is associated with the animal's ability to adaptively shed hair. At the beginning of the summer when well-adapted cattle are shedding their thick winter coats in preparation for warmer months, severely susceptible cattle tend not to shed their coat. Adaptive hair shedding is also an effective measure of heat tolerance and is closely related to cow performance, particularly 205-day weaning weight of her progeny.

In cooperation with 74 producers across the United States, over 12,000 cattle were scored on a scale of 1-5 for the early-summer hair shedding phenotype in 2016, 2017, and 2018. Participating cattle were genotyped using the GGP-F250 SNP panel developed by the University of Missouri, which contains ~170,000 candidate functional variants and ~25,000 variants in common with beef cattle industry standard genotyping assays.

Genomic breeding values were generated with a repeated records GBLUP model using citizen science-sourced phenotypes and publicly available environmental data. Using genotypes imputed to ~850,000 SNPs against a multi-breed reference panel, we fit several genome-wide association models in order to identify variants associated with the hair shedding phenotype while also taking breed and environmental variation into account. We also fit genotype-by-environment GWAS models to discover variants that interact significantly with temperature, heat index, and latitude/photoperiod. We also demonstrate the relationship between cow productivity and early-summer adaptive hair shedding in heat-stressed and fescue toxicosis-stressed environments. The outcomes of this project will improve selection tools for use in beef cattle production, especially related to improving heat tolerance.

# GENERATING A LIST OF PRIORITIZED GENES ASSOCIATED WITH A PATIENT'S DISEASE AND PHENOTYPE USING THE GENEPANEL.IOBIO WEB TOOL

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

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

The first step in the diagnostic variant analysis process usually involves the identification of variants in genes related to the patient's condition or phenotypes. Compiling a comprehensive list of genes using currently available tools is not straightforward. Additionally, these tools currently have little utility for the physicians or diagnostic clinicians who currently have to rely on the experts to compile such a gene list.

To address these challenges, we have developed ***genepanel.iobio*** (https://genepanel.iobio.io). This web application allows the user to collate lists of genes in commercial panel tests for the patient's condition, using submitter-driven information from h the NCBI's genetic testing registry (GTR); genes associated with the patient's phenotypes according to the Phenolyzer tool; and the clinician/user's own expert gene list. The genepanel.iobio tool integrates gene lists from these sources into a single prioritized gene list.

To start the workflow, the clinician can enter this condition in the search bar of the "Genetic Testing Registry" tool of the ***genepanel.iobio***. The app searches NCBI's genetic testing registry (GTR) for all commercial genetic tests testing the entered condition in real-time and quickly provides a list of associated genes. These genes are ordered on the number of GTR panels a gene occurs in. The "Phenolyzer" tool can be used in a similar way to generate genes associated with one or multiple clinical phenotypes. The clinician now enters a phenotype term to generate a rank-ordered list of genes associated with the phenotype. Additionally, the clinician can add any gene which (s)he suspects that might be relevant for the disease or phenotype but were not included by the previous searches. In all of these steps, genes can be added or removed from the generated list. The final summary compiles all genes into a single list. Lastly, the list can be exported by copying it to the clipboard or by saving it to a file. ***genepanel.iobio*** is available both as a standalone app, and as part of our integrated clinical diagnostic analysis tool clin.iobio (http://clin.iobio.io) presented elsewhere.

# A GLOBAL SURVEY OF TRANSCRIPTION IN 18,000 SAMPLES IDENTIFIES PUTATIVE FUNCTIONS OF HIGHLY REPETITIVE, CANCER-ASSOCIATED RNAS ORIGINATING FROM ENDOGENOUS RETROVIRAL ELEMENTS

Christine Eng*, Shumei Chia*, Engin Cukuroglu*, Lam-Ha Ly, Iwona Szczerbinska, Winston Chan, Ramanuj DasGupta, Jonathan Goeke

Genome Institute of Singapore, Computational & Systems Biology, Singapore

*Authors contributed equally.

Deregulation of transcription is one of the defining molecular characteristics of cancer. In contrast to the protein coding parts of the genome that are often unique, the noncoding genome is interspersed with highly repetitive elements. The largest fraction of such repeats resembles fragments from retrotransposon that have been found to be actively transcribed. However, the repetitive nature makes it a major challenge to estimate their expression, and their contribution to the transcriptome in cancer remains largely unknown.

Here, we analysed more than 18,000 RNA-Seq samples covering all major adult tissues and tumour types to comprehensively investigate expression of retrotransposon-derived RNAs in cancer. Overall, we find that about 4% of the transcriptome resembles fragments from retrotransposons, contributing to transcription start and end sites, splice sites and exons of protein coding and noncoding genes. The most highly expressed RNAs in cancer originate from full length copies of endogenous retroviruses H (HERVH), more specifically, the LTR7Y subfamily of HERVH that are largely inactive in normal tissues. While more commonly known to be essential for pluripotency in embryonic stem cell, we find that HERVH expression is also associated with dedifferentiation in cancer based on estimated stemness indices. Furthermore, we show that LTR7Y is bound by KLF5, a key transcription factor regulating cell proliferation and differentiation. Silencing the expression of these cancer-associated RNAs in primary patient-derived colorectal cancer cell lines results in reduced proliferation and apoptosis. Finally, we observe high variability in expression of LR7Y-HERVH in colorectal cancer patients, which seems to predict clinical outcome.

Together our study demonstrates that the repetitive elements genome contributes significantly to transcription and highlights a new class of highly abundant RNAs that may appear to functionally contribute to the molecular characteristics of cancer.

# DEVELOPING A COMPREHENSIVE VIEW OF THE *PRDM1* NEGATIVE AND POSITIVE REGULATORY LANDSCAPE

Melissa Englund[1], Torrin McDonald[1], Sierra Nishizaki[1], Jessica Switzenberg[2], Alan P Boyle[1,2]

[1]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [2]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI

As data about the structure and content of the human genome accumulates, non-coding DNA regions have been revealed to play a central role in regulating gene expression. Through coordinated interactions, these elements control the specific spatiotemporal gene expression required for complex cell developmental programs. As a result, mutations in regulatory elements are associated with multiple human diseases. While positive regulatory elements (PREs) such as enhancers have been abundantly characterized, negative regulatory elements (NREs) such as silencers and enhancer blockers remain under-studied. Here we utilize a novel transient transfection-based reporter assay panel to rapidly characterize the regulatory capacity of all putative PREs and NREs in the proximal regulatory space of the *PRDM1* (*BLIMP1*) gene.

*PRDM1* encodes a zinc-finger containing protein that can act as either a negative or positive transcription factor for a number of genes involved in innate and adaptive immune response and has a direct role in NK-, T- and B-cell maturation and proliferation. In order to better understand the contribution of changes in the regulatory landscape of *PRDM1* throughout B-cell development to changes in gene expression patterns, we will leverage our assay to determine the regulatory capacity of all non-coding open chromatin regions surrounding *PRDM1* in both K562 lymphoblastoid myelogenous leukemia cells and U266 B lymphocyte myeloma cells. Once regulatory element behavior has been measured, transcription factor binding patterns, overlapping chromatin modifications, and impact on native gene expression of these elements will be explored. Additionally, the combinatorial effects of negative and positive regulatory elements and their cumulative effect on native gene expression will be tested both episomally, and in the native chromatin context of the gene. Not only will this work contribute to a more thorough understanding of the regulatory landscape of an important transcription factor with functions ranging from innate and adaptive immune response to early heart development, proper epidermal formation and cell growth, it will also generate well-characterized, comprehensive examples of novel NREs and their roles in determining specific spatiotemporal patterns of gene expression in combination with positive elements.

# EXTENDING THE RESOURCES CREATED BY THE 1000 GENOMES PROJECT IN THE INTERNATIONAL GENOME SAMPLE RESOURCE (IGSR)

Susan Fairley, Ernesto Lowy, Gabriel Aldam, Paul Flicek

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

The 1000 Genomes Project created valuable data sets, including the largest fully public catalogue of human genetic variation. These data sets continue to be widely used and, via the International Genome Sample Resource (IGSR), we work to maintain access to the data and expand those resources through updating them to GRCh38, adding new data types and adding new samples.

In 2018, we released variation calls generated directly on GRCh38 from alignments of the phase three low coverage and exome data. Our multi-caller approach used BCFtools, FreeBayes and GATK UG to identify sites. Following integration of the call sets, GATK was used in genotyping, prior to using Beagle and SHAPEIT2 to generate a phased biallelic SNV call set, which is available here:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/

We have extended the data sets made available in IGSR through collaboration with data producers, assisting in data coordination and access in line with FAIR principles. This includes working with the New York Genome Center to share newly produced, high coverage sequencing data for the 2,504 phase three panel samples. In addition, we are part of the Human Genome Structural Variation Consortium, who continue to produce data sets for samples from 1000 Genomes.
We are also working to expand the range of data sets which have been aligned to GRCh38, including 90x coverage data for 90 Han Chinese samples sequenced at BGI.
IGSR has also significantly extended the number of samples represented in our collections since the conclusion of the 1000 Genomes Project. Data from 400 samples from the Gambian Genome Variation Project, over 820 samples from the Human Genome Diversity Project and over 280 samples from the Simons Genome Diversity Project have been added to IGSR.
To facilitate navigating the resources and locating data, we provide a data portal as part of our website. This is being enhanced to better integrate new data sets where the population metadata uses a different structure than the 1000 Genomes Project to enable searches across data collections from different projects while clearly identifying samples that are common to multiple groups.
IGSR works closely with the Ensembl genome browser to ensure that available data can be explored in genomic context. We also maintain an email helpdesk, info@1000genomes.org, which answers queries about the data sets and methods of access.
IGSR is open to establishing new collaborations that continue to expand and build on the resources created by the 1000 Genomes Project.

# AN ATLAS OF HUMAN BENIGN VARIATION INFERRED FROM PRIMATE POPULATION SEQUENCING

Kyle Kai-How Farh[1], Hong Gao[1], Laksshman Sundaram[1], Serafim Batzoglou[1], Lukas F Kuderna[2], Marta Gut[3], Ivo Gut[3], Felipe Ennes[4], Takashi Hayakawa[5], Amanda D Melin[6], Katerina Guschanski[7], Jean P Boubli[4,8], Christian Roos[9], Jeffrey Rogers[10], Tomas Marques-Bonet[2,3,11,12]

[1]Artificial Intelligence Laboratory, Illumina, San Diego, CA, [2]Institut de Biologia Evolutiva, UPF-CSIC, Barcelona, Spain, [3]CNAG-CRG, Centre for Genomic Regulation, Barcelona, Spain, [4]University of Salford, Salford, United Kingdom, [5]Japan Monkey Center, Inuyama, Japan, [6]University of Calgary, Department of Anthropology and Archaeology, Calgary, Canada, [7]Uppsala University, Department of Ecology and Genetics, Uppsala, Sweden, [8]Instituto Nacional de Pesquisas da Amazônia, Manaus, Brazil, [9]Leibniz Institute for Primate Research, German Primate CenterGöttingen, Germany, [10]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [11]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain, [12]Institut Català de Paleontologia Miquel Crusafont, Barcelona, Spain

Deciphering the functional consequence of all variants in the human genome is a long-standing goal for human genetics, which holds profound implications for the application of sequencing in routine clinical practice. We approach the problem by employing population sequencing to catalog over 500 million common polymorphisms from one hundred non-human primate species. Analysis of the 30X whole genome sequences of over four hundred individuals from diverse genera reveals for the first time the global landscape of tolerated polymorphisms in the primate lineage. We quantify the fitness consequences of identical-by-state protein-coding mutations within the primate lineage, and find that their selection coefficients are highly congruent across different species, including our own. Therefore, the resulting atlas of common primate polymorphisms can be leveraged to inform the functional consequence of human clinical mutations. We apply this resource to re-classify hundreds of thousands of variants of unknown significance in human disease genes as likely benign, based on their presence at high allele frequency in other primate species.

# THE UNBIASED LENGTH SPECTRUM OF HUMAN DE NOVO MUTATIONS IN 4,330 CHILDREN

Andrew Farrell[1], William Richards[1], Anna Docherty[2], Hilary Coon[3,4], Gabor T Marth[1]

[1]University of Utah, Human Genetics, Salt Lake Cityh, UT, [2]University of Utah, Department of Psychiatry and Human Genetics, Salt Lake City, UT, [3]University of Utah, Department of Psychiatry, Salt Lake City, UT, [4]University of Utah, Department of Biomedical Informatics, Salt Lake City, UT

De novo genetic variation introduces new variations into the population and can cause genetic diseases. Despite its importance, the true spectrum of de novo human mutations has so far been elusive primarily for technical reasons: accurately detecting the roughly 70 de novo mutations in the 6 billion base pairs of a human genome poses a formidable challenge; compounded by the biases of mapping-based variant detection methods. Moreover, distinct classes of tools are currently used to reliably detect single nucleotide variations (SNVs) and short insertions-deletions (INDELs) of up to ~50bp; and 500bp or larger structural variation events (SVs); leaving a "detection gap" for medium-sized (50-500bp) de novo mutation events.

We developed a k-mer based method, RUFUS, for reference-free detection of all types and sizes of de novo mutation events. With RUFUS, mutations are identified and assembled before reads are compared to the reference, thus removing all reference bias and variant size limitation, resulting in completely even sensitivity and specificity across variations of all sizes. Experimental validations of RUFUS mutation calls in multiple datasets indicate extremely high accuracy. We applied this method to detect de novo mutations in 4,330 children in the Simons Foundation SSC (Simons Simplex Collection) dataset, i.e. 2,385 autistic children and 1,945 unaffected siblings, the largest dataset to date in which to study de novo genetic variation. Here we present the first comprehensive de novo somatic mutation dataset where variants of all sizes were detected by a single algorithm. Across both autistic probands and siblings, we saw an average of 72.6 denovo mutations per sample; 86.14% of these are SNVs, 13.58% are 1-50bp INDELs, 0.20% are medium-sized (50-500bp) INDELs, and 0.08% are >500bp SV events (i.e. on average 62.54 SNVs, 9.86 short INDELs, 0.145 medium INDELs, and 0.058 SVs per individual). When accounting for the size of the events, de novo SVs alter the genome of the child by far the most, on average by 3,804bp, and are present in approximately 1 of 17 births. Medium-sized INDELs, the rate of which until this study could not be fully ascertained, alter on average 20.69bp per birth, an effect comparable to that of SNVs (62.54bp per birth), and occur in 1 of 7 births. The overall distribution of de novo event size is indistinguishable between probands and siblings. However, de novo variation within autism-associated genes in the probands is markedly higher higher with RUFUS (i.e. more than double) as compared to the the siblings, a phenomenon part of ongoing investigation.

# GENETIC AND ENVIRONMENTAL DETERMINANTS OF CARDIOMYOCYTE GENE EXPRESSION

A S Findley[1], A L Richards[1], K Rhodes[2], M C Ward[2], C A Kalita[1], A Alazizi[1], X Wen[3], D E Lanfear[4], R Pique-Regi[1], Y Gilad[2], F Luca[1]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Chicago, Department of Human Genetics, Chicago, IL, [3]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [4]Henry Ford Health System, Department of Medicine, Detroit, MI

Cardiomyopathy is a disease of the heart muscle that leads to arrhythmias and heart failure. As the number of people diagnosed with heart failure increases, it is imperative to gain a better understanding of the genetic and environmental factors leading to cardiomyopathy. We have established a cell culture model to explore gene-environment interactions (GxE) in cardiomyocytes (CMs). We have reprogrammed lymphoblastoid cell lines (LCLs) from 6 individuals into induced pluripotent stem cells (iPSCs), which have been further differentiated into CMs. A key feature of this system is the ability to expose these cell types to the same panel of treatments to investigate cell type specific response and treatment-by-cell type interactions. To this end we have exposed all cell lines from the 3 cell types to a panel of 28 treatments, including hormones, vitamins, environmental contaminants, and common drugs. Twelve treatments induced a significant response for hundreds of genes in at least one cell type. We identified 4835 genes expressed in all cell types and that show evidence of treatment-cell type interactions as measured by likelihood ratio test in DESeq2 (FDR < 10%). In particular, 96% of differentially expressed genes in CMs in response to treatment with insulin are not differentially expressed in LCLs or IPSCs. These unique insulin-CM genes are enriched for cholesterol and ADP metabolic processes. To dissect the sources of gene expression variation across individuals, we used a random effects model accounting for interactions with cell-type and treatment, while taking advantage of experimental replication and accounting for other technical variables. We estimate that 64% of interindividual variation is explained by interactions with cell-type, while 34% of variation is due to interactions with treatments (potential GxE). To identify SNPs which could be responsible for these interactions in the CMs, we performed allele-specific expression (ASE) analysis. We identified 615 genetic variants in CMs displaying ASE, corresponding to 242 genes. For example, caveolin-2 is a gene downregulated and with ASE in CMs treated with nicotine, that has been implicated in atrial fibrillation by GWAS. This and other examples show that genetic interactions with environmental exposures (treatments) have a strong contribution to inter-individual variation in health and disease.

# DECIPHERING THE HUMAN GENETIC PANORAMA OF BASQUES

André Flores-Bello[1], Frédéric Bauduer[2], Jasone Salaberria[3], Bernard Oyharçabak[3] Lluis Quintana-Murci[4], David Comas[1]

[1]Institut de Biologia Evolutiva (CSIC-UPF),Universitat Pompeu Fabra, Departament de Ciències de la Salut i de la Vida, Barcelona, Spain, [2]University of Bordeaux I, Laboratoire Maladies Rares: Génétique et Métabolisme (MRGM), EA 4576, Bordeaux, France, [3]Centre National de la Recherche Scientifique, CNRS UMR5478 IKER, Centre de Recherche sur la Langue et les Textes Basques (IKER), Bayonne, France, [4]Institut Pasteur, Human Evolutionary Genetics Unit, Centre National de la Recherche Scientifique URA31012, Paris, France

Basques have been historically distributed along the Western Pyrenees, in the Franco-Cantabrian region, between the Spanish and the French territories currently organised in seven provinces: Gipuzkoa, Bizkaia and Araba in the southern side of the Pyrenees; and Zuberoa, Lapurdi and Nafarroa Beherea located in the northern side. Basque people have been in the limelight of many researches due to several singular cultural and biological traits that, with high controversy, seemed to place them as an heterogeneous, isolated and unique population within Europe, highlighting their own non-Indo-European language, Euskara, and its five main dialects. However, due to the limitations of previous studies about Basque genetic structure, there is still a lively debate about their history and their presumed singularity.

In order to overcome the previous limitations, we have analyzed, at a microgeographical level, genome-wide data (~629K SNPs) of all Basque groups (according to geographic and linguistic criteria) and surrounding non Basque-speaking populations from the Franco-Cantabrian region. Overall, a total of ~200 individuals from 18 different geographical areas were included. We have also considered other external modern and ancient West Eurasian samples from publicly available datasets in order to have external references.

Our results, mainly focused on haplotype-based methods, show a clear differentiation of Basques from other European populations, as previously suggested. Even at a local scale, we detect a genetic differentiation of Basques compared to surrounding Spanish and French groups, with a genetic cline through the non-Basque-speaking Franco-Cantabrians. Moreover, an internal genetic heterogeneity is detected within the Franco-Cantabrian region, with a significant correlation with geography. Finally, our analyses including ancient samples show that the ancestry of Basques is mostly defined by a higher Neolithic farmer component, which challenges the previous ideas of the Paleolithic composition of Basques.

# HUMAN ENHANCERS WITH COMPLEX EVOLUTIONARY ARCHITECTURES ARE MORE ACTIVE THAN ENHANCERS WITH SIMPLE ARCHITECTURES

Sarah L Fong[1], John A Capra[1,2]

[1]Vanderbilt University, Human Genetics, Nashville, TN, [2]Vanderbilt University, Biological Sciences, Nashville, TN

Enhancers are distal genomic regions that bind transcription factors (TFs) and regulate gene expression. Mutations and structural variation in enhancers give rise to new gene regulatory functions, but also associate with disease. While evolutionary conservation in protein coding genes is a powerful predictor of function, low conservation across enhancer landscapes makes enhancer variant interpretation a challenge. For example, only 1% of active human liver enhancers are conserved among mammals, indicating activity turns over quickly though these sequences are often shared across species.

Given the high frequency of enhancer mutations in complex disease and human biology, methods for interpreting these genetic variants are needed. We hypothesize the age of an enhancer sequence may hold information about its function. Enhancers are often composites of old "core" sequences and younger "derived" sequences, but how age architecture relates to function is not understood. We propose young, derived regions of enhancers increase enhancer function by introducing transcription factor binding sites (TFBSs) that potentiate activity.

We explored the relationship between enhancer evolutionary architecture and function by integrating evolutionary analysis of FANTOM enhancers across diverse cells and tissues, TFBS from ChIP-seq, and massively parallel reporter assays (MPRAs). Across tissues, enhancers are enriched in older genomic regions. We find over half of human enhancers arose in the ancestor of placental mammals, validating previous observations in liver tissue. We find 63% of enhancer sequences have "simple" architectures consisting of one age, while the rest are "complex," consisting of sequences of multiple ages. These evolutionarily "simple" enhancers have lower activity than "complex" enhancers. Within complex enhancers, derived regions have higher MPRA activity than older core regions. This supports a model in which the presence of derived regions promotes strong enhancer activity. We hypothesized TFBS density might explain these functional differences. However, derived regions have lower TFBS density than simple enhancers and complex cores, suggesting that specific TFs in derived segments promote activity. We identified 90 TFs enriched for binding in simple enhancers, complex enhancer cores, or derived regions.

We propose human enhancers have two types of age architectures, those composed of composites of multiple sequence ages, and those in sequences with a single age. These two enhancer architectures differ in activity, TFBS density, and the TFs they bind. Complex and simple architectures likely harbor uneven distributions of disease variants. Thus, considering enhancer age architecture may aid the interpretation of human regulatory regions, functional targets, determinants of selection, and disease severity associated with enhancer variants.

# REFINING REFERENCE PROTEIN-CODING GENE ANNOTATION FOR HUMAN AND MOUSE IN GENCODE

Adam Frankish, Paul Flicek, The GENCODE Consortium

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

GENCODE produces foundational genome annotation for the human and mouse reference genomes. We aim to identify and classify, with high accuracy, all protein-coding genes, pseudogenes, long non-coding RNAs and small RNAs based on defined biological evidence and make this annotation freely available for the benefit of biomedical research and genome interpretation.

GENCODE have completed first pass manual annotation of the mouse reference genome. This annotation was released as part of Ensembl 96, GENCODE M21 in April 2019. For the first time there will be essentially equivalent Ensembl/GENCODE manual annotation for the human and mouse genomes, permitting more accurate comparative analysis of their gene complement and initially revealing that (excluding readthrough transcripts that span multiple loci) human currently has 19348 annotated protein coding genes while mouse has 21704. This annotation supports ongoing work to identify pseudogenised and unannotated protein-coding genes in both species and protein-coding genes whose annotation is not supported by current public data. We have previously been successful in using evolutionary conservation of protein-coding potential identified by PhyloCSF to find ~150 unannotated protein-coding genes in human and continue to explore the data to this end in both human and mouse. In addition, combining PhyloCSF with long and short transcriptomic data has permitted the identification of unannotated conserved exons in protein-coding genes. We have also defined a novel class of non-functional exons 'pseudo exons', either ancestral coding exons that have gained fixed LoF mutations in the human lineage (unitary pseudo exons, flagged by PhyloCSF) or exons created by duplications within a gene (duplicated pseudo exons). Identifying and classifying these features for the first time in a reference gene annotation resource will support further study and determination of their biological relevance.

GENCODE have expanded our efforts to work collaboratively with other reference annotation databases. Running for more than a decade, the CCDS project aims to agree one CDS per protein-coding gene. Following on from CCDS, the recently established Matched Annotation from NCBI and EMBL-EBI (MANE) project aims to 1) identify a representative transcript that captures most information for every protein-coding gene and 2) revise the annotation of that selected transcript in the RefSeq and Ensembl/GENCODE sets to completely match in overall exon structure, precise 5' and 3' transcript boundaries, and at the nucleotide sequence level.

# TRANSCRIPTOMICS DATA GOLDMINE ON THE ENCODE PORTAL

Idan Gabdank, Jason Hilton, Yunhai Luo, Jennifer Zamanian, Khine Z Lin, Jennifer P Jou, Zachary A Myers, Emma O'Neill, Paul Sud, Forrest Tanaka, Jin W Lee, Casey R Litton, Keenan Graham, Ulugbek K Baymuradov, Philip Adenekan, Benjamin C Hitz, J Seth Strattan, J Michael Cherry

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

The Encyclopedia of DNA Elements (ENCODE) project is an international collaboration of research groups funded by the NHGRI. The main goal of the project is to identify the functional elements of the human genome. The ENCODE portal (*https://www.encodeproject.org/*), established and maintained by the ENCODE Data Coordination Center (DCC), provides unrestricted access to the data generated by the members of the ENCODE consortium and other related projects such as Model organism ENCODE (modENCODE), Model organism Encyclopedia of Regulatory Networks (modERN), Genomics of Gene Regulation (GGR), and Roadmap Epigenomics Project (Roadmap). The DCC develops and updates the data model and uniform processing pipelines that allow transparency and reproducibility of the experimentally collected and computationally generated data. The metadata is categorized and implemented in a structured, machine-readable data model, ensuring users of the ENCODE portal can search and identify related experiments, understand how assays were performed, and track the provenance of the data. Currently the ENCODE portal hosts data from more than 14 thousand datasets, representing experimental results of more than 40 diverse assay types, performed in multiple cell lines and tissues under various conditions. The portal provides free access to over 640 terabytes of raw experimental, processed, and analysis data files with the corresponding metadata. Richness, diversity, and structure of the data on the ENCODE portal make it an invaluable resource for the scientific community. The presence of curated and standardized experimental and analysis results from RNA microarrays, RNA-seq, TF perturbation followed by RNA-seq, small RNA-seq, microRNA-seq, RAMPAGE, and eCLIP assays turns the portal into a goldmine for scientists seeking their transcriptomics fortune. Here we show how to mine the ENCODE portal for transcriptomics data to enrich and complement your experimental results and facilitate scientific discovery.

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

Thiago Miller[1,2], Gabriela Guardia[1], Helena Conceicao[1,3], Felipe Santos[1,3], Rodrigo Barreiro[1,2], Jose L Buzzo[1,2], Fernanda O Rego[1], Pedro A Galante[1]

[1]Hospital Sirio-Libanes, Molecular Oncology, Sao Paulo, Brazil, [2]Universidade de São Paulo, Bioquimica, Sao Paulo, Brazil, [3]Universidade de São Paulo, Interunidades em Bioinformatica, Sao Paulo, Brazil

Establishing pipelines for accurately extracting the plentiful of next-generation sequencing (NGS) data is achievable only when well-known controls are available. Here, we present SANDY, a general-purpose and easy-to-use tool for simulating NGS data. In brief, SANDY only requires a reference sequence (FASTA file containing genomic or transcriptomic data) and generates single/paired-end reads as in a real NGS experiment. SANDY supports a variety of parameters, which mimic characteristics ranging from the sequencing experiments (DNA or RNA sequencing, the sequencing coverage, the read length, and sequencing error rate, for example) to biological characteristics of the sequenced material (e.g., occurrence of single nucleotide variations, gene fusions, and gene expression profile of specific tissues and organs). SANDY simulates data for the most popular NGS technologies currently on the market, which includes sequencers from Illumina, Pacific Bioscience (PacBio), and Oxford Nanopore (Nanopore). SANDY is freely available at https://galantelab.github.io/sandy/, and it is quick and easy to install in the three most common operating system (Linux, Apple's macOS, and Microsoft Windows).

# POPULATION DIFFERENTIATION OF THE *ADCY9* PHARMACOGENE REVEALS A POTENTIAL EPISTATIC INTERACTION WITH *CETP*

Isabel Gamache[1,2], Magdalena Burchert[1,2], Rocio Sanchez[1], Éric Rhéaume[1,2], Jean-Christophe Grenier[1], Jean-Claude Tardif[1,2], Marie-Pierre Dubé[1,2], Julie Hussin[1,2]

[1]Montreal Heart Institute, Research Center, Montréal, Canada, [2]Université de Montréal, Medicine, Montréal, Canada

Dalcetrapib is a modulator of the cholesterol ester transfer protein (CETP) protein that increases the serum levels of HDL-cholesterol. Pharmacogenomic studies have revealed associations between polymorphisms in the *ADCY9* gene and clinical responses to the drug. Recently, it was shown that Adcy9 inactivation protects against atherosclerosis in mice, but only in the absence of CETP activity, pointing to a functional interaction between them. However, the mechanism behind this interaction is still unknown. Using data from the 1000 Genomes Project and population genetics approaches, we characterized the haplotypes responsible for the pharmacogenomic response and their prevalence in human populations. Differentiation between populations at *ADCY9* polymorphisms showed evidence for natural selection at the locus associated with dalcetrapib response in the Peruvian population. We analysed correlation between *ADCY9* and *CETP* haplotypes, and found significant linkage disequilibrium between highly differentiated SNPs in *ADCY9* and *CETP*, separated by 50 Mb on chromosome 16, suggesting a potential co-evolution between haplotypes within these interacting genes. Using the GEUVADIS transcriptomic dataset, we identified an eQTL-like interaction between the differentiated SNPs in *ADCY9* and *CETP*, influencing CETP expression. In GTEx, we found evidence that specific combinations of these two SNPs' genotypes are associated with acute cardiovascular events, providing a line of inquiry for a potential biological mechanism behind the pharmacogenomic association involving *ADCY9*. After confirming these associations in additional biobank datasets, we will pursue functional validation of this novel putative interaction. This work exemplifies how population genetics combined with well-established transcriptomics datasets and deeply-phenotyped cohorts can shed light on the role of genomic variants in clinical treatment responses.

# EFFECT OF LARGE COPY NUMBER VARIANTS ON GENERAL COGNITION IN 18,329 HEALTHY INDIVIDUALS

Eugene J Gardner[1], Mari Niemi[1], George Kirov[2], Adam S Butterworth[3], Hilary C Martin[1], Matthew E Hurles[1]

[1]Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]Cardiff University, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff, United Kingdom, [3]University of Cambridge, Public Health and Primary Care, Cambridge, United Kingdom

Copy Number Variants (CNVs) represent the deletion and duplication of entire segments of the human genome and have been previously associated with a wide range of phenotypes such as severe psychiatric conditions and several intelligence measures. However, a general model of how CNVs modulate both general cognition and polygenic intelligence in the healthy population has yet to be elucidated. To assess the effects of large CNVs (>10Kbps) on such measures, we called CNVs from genotyping array data generated for the INTERVAL study – a collection of 46,856 putatively healthy individuals recruited for a blood donation study sponsored by the England National Health Service. Following genotyping, we filtered sites with a novel random forest-based approach based on intersection to CNVs called with individual-matched Whole Exome Sequencing (WES) data. Our method achieves a high sensitivity and specificity, with area under the curve (AUC) measures of 0.92 and 0.93 for deletions and duplications, respectively. Following variant quality control, 208,325 redundant CNVs (134,534 deletions, 73,791 duplications) remain, with an average of 4.4 sites per individual. We are interrogating and plan to present the relationship between DNA dosage and protein dosage using proteomics data generated for a subset of INTERVAL participants. Next, we queried our cohort for individuals carrying rare (minor allele frequency < 0.001) CNVs which intersect a gene that is strongly loss of function (LoF) intolerant (pLI > 0.9) and expressed in the fetal brain, for which we identified 2,274 carriers (4.9% of individuals; 915 deletions, 1,391 duplications). To test the effects of such CNVs on general cognition we utilized a linear model incorporating a polygenic score for intelligence, controlled for ancestry, age, and sex. With our model, carriers of deletions which impact crucial genes have a -0.17-standard deviation (95% CI: [-0.07, -0.27], p = 6.6e-4) decrease in g, a normalized metric of general cognition. This finding is similar to the effect size of -0.17 (95% CI: [-0.06, -0.29], p = 3.9e-3) seen for rare nonsense single nucleotide variants in a smaller subset of 1,906 individuals. We likewise tested whether duplications which impact the same classes of genes had any effect but did not find a significant correlation, which indicates that decreased gene dosage uniquely impacts general cognition. Beyond our assessment of the healthy population, we have also used our extensive collection of putatively tolerated CNVs to develop a map of the genome intolerant to dosage changes. We will present how this map will be used to aid in the diagnostic interpretation of developmental disorders.

# MOLECULAR PROFILING OF PRIMARY CUTANEOUS MELANOMAS REVEALS PATHWAYS PREDISPOSING TO BRAIN METASTASES

Manik Garg*[1], Roy Rabbie*[2,3], Nuno A Fonseca[1], Dominique L Couturier[4], Alvis Brazma*[1], David J Adams*[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Functional Genomics, Cambridge, United Kingdom, [2]The Wellcome Trust Sanger Institute, Experimental Cancer Genetics, Cambridge, United Kingdom, [3]Cambridge University Hospitals National Health Service Foundation Trust, Department of Oncology, Cambridge, United Kingdom, [4]University of Cambridge, Cancer Research UK - Cambridge Institute, Cambridge, United Kingdom

*These authors contributed equally.

Brain metastases commonly occur in patients with advanced melanoma. Although targeted therapy and immune checkpoint inhibitors can benefit a subset of patients, the management of brain metastases remains a key unmet need. There are currently no validated clinical, pathologic or molecular factors that predict relapse to the brain. We present data from 1343 adults with stage IIB – IIIC (AJCC 7th Edition) primary cutaneous melanomas from a large UK-based multicentre prospective clinical trial. Mutation panel and gene expression data from a subset of cutaneous primaries and regional lymph nodes were correlated with the first anatomical site of distant relapse with a minimum of 6 years follow-up.

We found no significant differences in the clinico-pathologic characteristics between primary tumours that relapsed to the brain versus those that relapsed to distant extracranial sites. There were also no differences between these metastatic groups in terms of mutations in a select driver gene panel, including genes implicated in the MAPK and PI3K signalling pathways. Differential expression analyses revealed that selective genes involved in the axonal-signalling and oxygen transport pathways were differentially expressed in primaries that relapsed to the brain versus extracranial and these genes were not implicated in other site-specific comparisons. Interestingly, some of these genes have also been found common in other independent studies on brain metastases performed in mouse models (unpublished). We hypothesise that particular neural induction and migration signals may have specific correlates with intracranial invasion.

# CROSS-SPECIES ANALYSIS ACROSS 450 MILLION YEARS OF EVOLUTION REVEALS CONSERVATION AND DIVERGENCE OF THE MICROGLIA PROGRAM

Laufey Geirsdottir[1], Eyal David[1], Hadas Keren-Shaul[1], Daniel Erny[2], Marco Prinz[2], Ido Amit[1]

[1]Weizmann Institute of Science, Department of Immunology, Rehovot, Israel, [2]Uniklinikum Freiburg, Institute of Neuropathology, Freiburg, Germany

Understanding the biology of microglia is of significant interest in the prevention and treatment of neurodegenerative diseases. Efforts to translate animal neurodegenerative models to human therapies call for better characterization of microglia pathways across species. Using immunohistochemistry with computational reconstitutions, and a combined single-cell, and bulk analysis strategy, we characterize microglia across seventeen species, spanning more than 450 million years of evolution. We find that while in most species, microglia show a single dominant transcriptional state, humans demonstrate significant intra-species heterogeneity. We demonstrate that microglia express a conserved core gene program of orthologous genes from zebrafish to human. Nevertheless, we observed significant cross-species differences in metabolic, complement, phagocytic and signaling pathways that are enriched for susceptibility genes to brain disorders, including Alzheimer's and Parkinson's. Thus, our study provides an important resource of divergent and conserved microglia pathways across evolution with important implications for development of microglia-based therapies.

# ON THE IMPORTANCE OF LONG-RANGE PHASING FOR NON-INVASIVE PRENATAL TESTING AT LOW FETAL FRACTIONS

Giulio Genovese[1,2,3], Po-Ru Loh[3,4], Robert E Handsaker[1,3], Seva Kashin[1,3], Steven A McCarroll[1,2,3]

[1]Broad Institute, Stanley Center, Cambridge, MA, [2]Harvard Medical School, Genetics, Boston, MA, [3]Broad Institute, Medical and Population Genetics, Cambridge, MA, [4]Brigham and Women's Hospital, Genetics, Boston, MA

Since the discovery of cell-free fetal DNA (cffDNA) in maternal plasma in 1997, many developments have aimed at diagnosing fetal aneuploidy non-invasively by analyzing the relative abundance of cffDNA in maternal blood. Commercial companies offer screening tests to detect fetal trisomies 21, 18, and 13, whose occurrence increases with age, as early as 9 weeks gestation age (GA).

However, when the amount of cffDNA in maternal plasma, i.e. the fetal fraction (FF), is much lower than the 10% average, it becomes difficult to detect chromosomal imbalances. This issue is exacerbated by the fact that aneuploid pregnancies such as trisomy 13 and 18 associate with lower FFs due to smaller placentas. To increase sensitivity at low FFs, we investigated the role that long-range phasing information, that is, haplotypic information about the genome of the mother, plays in the detection of fetal aneuploidy from maternal plasma.

From a pregnant mother carrying a fetus with trisomy 18 we generated targeted sequence data across chromosomes 13, 18, 21, and X at 11 weeks GA through a commercial company performing non-invasive prenatal testing (NIPT) using single nucleotide polymorphisms (SNPs) and whole-genome sequence data at 15 weeks GA using massive parallel sequencing. We further generated genotype array data for the mother and the father and their parents to infer long-range phase information through trio-phasing.

Through an extension of a novel framework we previously developed to detect mosaic chromosomal alterations at low cell fractions (https://github.com/freeseek/mocha), we modeled the haplotypic inheritance pattern of the fetus. The framework identified trisomy 18 as the best model explaining both the data from the targeted approach (~28M sequence reads, FF 3.4%, LOD>15) and the whole-genome approach (~232M sequence reads, FF 6.1%, LOD>40). In both cases the use of long-range phasing was key to integrate the evidence across the chromosomes and in both cases the data was consistent with a trisomy 18 due to an error in maternal meiosis II.

We finally speculate on the practical use of population-based long-range phasing for the purpose of NIPT at low fetal fractions as well as of early detection of cancer. While we obviated the problem of obtaining long-range phasing through trio-phasing, in practical clinical applications the problem might be best solved by having access to large biobanks with genotypes for million of individuals.

# *ONCOGENE.IOBIO*: A WEB APP FOR REAL-TIME, INTEGRATIVE EXAMINATION AND FUNCTIONAL PRIORITIZATION OF TUMOR MUTATIONS.

Stephanie J Georges, Yi Qiao, Xiaomeng Huang, Gabor T Marth

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

Genetic testing and clinical tumor sequencing are rapidly becoming part of the standard of care for a large subset of cancers. At present, a scattered framework of resources, often requiring expensive hardware and heavy computational skills, exists to aid clinicians and researchers in the bioinformatic analysis process. To assist in addressing the need for a comprehensive tool that operates intuitively, and at the pace of critical therapeutic timelines, we have developed *oncogene.iobio*. Utilizing our novel iobio analysis ecosystem (https://iobio.io), *oncogene.iobio* visualizes genomic, transcriptomic, and epigenetic data to aid in the discovery of variants with the highest potential for association with the patient's cancer.

*Oncogene.iobio* consumes primary sequencing data (in bam, vcf, fasta/fastq formats) sourced from local, or cloud-based repositories. In real-time, variants are identified and annotated on a gene-by-gene basis, and prioritized using functional annotations (sourced from SNPEFF, VEP, SIFT, PolyPhen) and cancer databases (COSMIC, cBioPortal). Uniquely, *oncogene.iobio* identifies compound heterozygotes between inherited cancer-implicated variants and somatically-acquired tumor mutations, dynamically filters mutations by allele frequency, and analyzes somatic variants within areas of chromosomal CNVs or LOH. Areas with large transcriptional change or allele-specific expression are highlighted when RNAseq data is present, along with hypo- or hyper-methylated sites given bisulfate-treated DNA sequencing reads. Ranked, candidate gene lists (sourced from ICGC, cBioPortal), as well as cancer-associated variants (sourced from GEMINI), are provided by *oncogene.iobio* to guide users in the analysis process, and to quickly elucidate suspect loci.

Some analyses may benefit from the more advanced features of *oncogene.iobio*, such as the ability to recall variants in real-time, using our Freebayes algorithm attuned with tumor-specific parameters. This workflow assists in discovering false negatives, and large-scale deletions and duplications, normally missed in low-coverage areas analyzed using canonical methods. Additionally, *oncogene.iobio* provides the ability to corroborate true positives and reduce reference mapping biases using our graph-based adjudication tool GRAPHITE ( https://github.com/dillonl/graphite). In conjunction with the rapidly-informative and visually-intuitive primary features, these options allow for critical biomedical findings sourced from high-confidence data.

With extensive integration and immediate feedback, *oncogene.iobio* aims to aid the discovery of impactful variants in oncological research and diagnostic care.

# COMPREHENSIVE FUNCTIONAL GENOMIC RESOURCE AND INTEGRATIVE MODEL FOR THE HUMAN BRAIN

Daifeng Wang*[1,2,3], Shuang Liu*[1,2], Jonathan Warrell*[1,2], Hyejung Won*[4,5], Xu Shi*[1,2], Fabio Navarro*[1,2], Declan Clarke*[1,2], Mengting Gu*[1,2], Prashant Emani*[1,2], Daniel Geschwind[6,7,8], James Knowles[9], Mark Gerstein[1,2,10,11]

[1]Yale University, Program in Computational Biology and Bioinformatics, New Haven, CT, [2]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT, [3]Stony Brook University, Biomedical Informatics, Stony Brook, NY, [4]University of North Carolina, Genetics, Chapel Hill, NC, [5]University of North Carolina, UNC Neuroscience Center, Chapel Hill, NC, [6]University of California–Los Angeles, Human Genetics, Los Angeles, LA, [7]University of California–Los Angeles, Program in Neurobehavioral Genetics, Los Angeles, LA, [8]University of California–Los Angeles, Neurology, Los Angeles, LA, [9]SUNY Downstate Medical Center College of Medicine, Brooklyn, NY, [10]Yale University, Computer Science, New Haven, CT, [11]Yale University, Statistics and Data Science, New Haven, CT

Despite progress in defining genetic risk for psychiatric disorders, their molecular mechanisms remain elusive. Addressing this, the PsychENCODE Consortium has generated a comprehensive online resource for the adult brain across 1866 individuals. The PsychENCODE resource contains ~79,000 brain-active enhancers, sets of Hi-C linkages, and topologically associating domains; single-cell expression profiles for many cell types; expression quantitative-trait loci (QTLs); and further QTLs associated with chromatin, splicing, and cell-type proportions. Integration shows that varying cell-type proportions largely account for the cross-population variation in expression (with >88% reconstruction accuracy). It also allows building of a gene regulatory network, linking genome-wide association study variants to genes (e.g., 321 for schizophrenia). We embed this network into an interpretable deep-learning model, which improves disease prediction by ~6-fold versus polygenic risk scores and identifies key genes and pathways in psychiatric disorders. This work is a capstone project of PsychENCODE Consortium and was published in Science (Vol. 362, Issue 6420, eaat8464, 2018).

# BIOLOGICAL INSIGHTS FROM THE INTEGRATED GENOMIC ANALYSIS OF 143 HUMAN EMBRYONIC STEM CELL LINES

Sulagna Ghosh*[1,4], Florian Merkle*[2], Giulio Genovese[1], Robert Handsaker[1], Seva Kashin[1], Konrad Karczewski[5], Steven McCarroll[1], Kevin Eggan[3]

[1]Broad Institute, Stanley Center for Psychiatric Research, Cambridge, MA, [2]University of Cambridge, Metabolic Research Laboratories and Medical Research Council Metabolic Diseases Unit, Cambridge, United Kingdom, [3]Harvard Medical School, Department of Genetics, Boston, MA, [4]Harvard University, Department of Stem Cell and Regenerative Biology, Cambridge, MA, [5]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA

Human embryonic stem cells (hESCs) hold great promise for disease modeling and regenerative medicine because of their unlimited potential for self renewal and ability to differentiate into multiple somatic lineages. Despite their widespread use and ongoing clinical application, surprisingly little is known about their genetic makeup and variation. Since genetic differences between hESCs can contribute to variability in transcriptional and cellular phenotypes or render cell lines less suitable for transplantation and disease related models, a necessary first step is to systematically characterize the genetic landscape across these cell lines. To this end, we applied high density DNA microarrays, whole exome sequencing (WES) and whole genome sequencing (WGS) to identify potentially deleterious sequence and structural variants in 143 readily-obtained hESC lines. The sensitivity of this integrated strategy revealed a surprising extent of undocumented first-degree relationships among these cell lines, and identified large and potentially deleterious structural variants, many of which were undetectable with traditional technologies. Furthermore, we identified many single nucleotide variants (SNVs) in genes associated with autosomal dominant disease, infertility, and cancer that may alter cellular phenotypes or compromise the safety of hESC-derived cellular products transplanted into humans. Together, the results described here serve as a resource to enable the rational selection of hESC lines for applications ranging from cellular models of disease to transplantation medicine. Since many groups would likely benefit from these genetic data, we developed a user-friendly online data portal with the aim of facilitating informed selection of lines and improving the safety and reproducibility of studies carried out with hESCs.

**Authors contributed equally to this work*

# OPEN TARGETS GENETICS: INTEGRATING GENETICS AND FUNCTIONAL GENOMICS TO IDENTIFY AND PRIORITIZE DRUG TARGETS

Maya Ghoussaini[1,2], on behalf of Open Targets[1]

[1]Open Targets, EMBL-EBI South, Wellcome Genome Campus, Hinxton, United Kingdom, [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom

The majority of drugs that enter clinical trials fail due to the lack of efficacy. This highlights the need to explore additional sources of evidence to inform drug discovery. Drugs that have their targets supported by genome-wide association study (GWAS) evidence are twice as likely to be approved for clinical use. Here we present Open Targets Genetics, an open-access resource that makes robust connections between disease-associated loci and likely causal genes, to enable the identification and prioritization of new drug targets.

We combine GWAS associations curated from literature and the recently released NHGRI-EBI GWAS Catalog summary statistics database (including over 3,000 UK Biobank phenotypes) to locate over 100,000 disease-associated loci. We then use statistical fine-mapping (or linkage-disequilibrium expansion if no summary statistics are available) to identify a set of potentially causal variants at each locus. Lastly, variants are integrated with functional genomics data (e.g. promoter capture Hi-C, DNase hypersensitivity sites), molecular quantitative trait loci (e.g. eQTLs, pQTLs) as well as in silico functional predictions to link each variant to its target gene(s) using a single evidence score. Most recently, we have added systematic disease-molecular trait colocalization analysis using an extensive collection of eQTLs.

The results of Open Targets Genetics are made available through our web portal (genetics.opentargets.org), for bulk download and via a GraphQL API, enabling users to easily prioritize genes at disease-associated loci and assess their potential as pharmaceutical targets.

# SIGNATURES OF TRANSCRIPTOMIC INDIVIDUALITY REVEALED BY THE NINE-BANDED ARMADILLO

Sara Ballouz[1], Megan Crow[1], Maria Pena[2], Jesse Gillis[1]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY, [2]Health Resources and Services Administration, US Dept of Health and Human Services, Baton Rouge, LA

A delicate balance between robustness and stochasticity is present within developmental programs in mammals. The outcome is individuality, or phenotypic uniqueness. One way to assess purely epigenetic variability leading to different phenotypes is to look at the variation of phenotypes across identical genomes, such as twins. But because twinning occurs rarely and randomly in humans, there are still uncertainties associated with the contributions of "nature, nature or noise". Here, we exploit the obligate polyembrony of the wild nine-banded armadillo (Dasypus novemcinctus), which are programmed to produce monozygotic quadruplets, to test for 'epigenetic signatures' of individuality.

First, we sequenced the blood transcriptomes of five litters of armadillo quadruplets to generally assess and quantify epigenetic variation, using gene expression as a proxy for epigenetic marks. Then, we looked longitudinally across a year to see if this variability is itself stable, and unique to genetic background or a common signature of armadillos. Unsurprisingly, we find a 'sibling-effect' - i.e., within-litter transcriptional similarity is always significantly higher than between-litter similarity. But remarkably, we find weak but significant 'individuality' signatures of siblings. Mostly, these signatures are unique to each armadillo litter, implying distinct processes being affected. While expression level of individual genes alone provides a significant signal ($p<0.05$), allele-specific expression patterns exhibit much greater individuality ($p<1E-3$), allowing armadillos to be uniquely identified relative to their siblings over time.

We then performed a very high-throughput meta-analysis of human differential expression data to identify any correlates of this effect, characterizing differential expression across more than 635 datasets totaling more than 27000 samples. Our hypothesis is that gene expression which is not subject to state-dependent changes is a likely candidate for signatures of individuality; even slight changes are likely to be characteristic once fixed within the individual. Consistent with this, we find that genes which are essentially never differentially expressed within humans are very strong predictors of armadillo identity ($p<1E-3$). We close with a targeted assessment of the same effect within human twin data.

Our work reveals the transcriptional stability and variability associated with genetic backgrounds with important implications for the interpretation of gene expression and developmental programs.

*CIS*-REGULATION PROVIDES A LINK BETWEEN THE HUMAN Y CHROMOSOME AND SEX DIFFERENCES IN NON-REPRODUCTIVE TISSUES.

Alexander K Godfrey[1,2], Sahin Naqvi[1,2], Lukas Chmatal[1], Joel M Chick[3], Richard N Mitchell[4], Steven P Gygi[3], David C Page[1,2,5]

[1]Whitehead Institute, Cambridge, MA, [2]Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, [3]Harvard Medical School, Department of Cell Biology, Boston, MA, [4]Harvard Medical School, Brigham and Women's Hospital, Department of Pathology, Boston, MA, [5]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA

Human males and females differ in each diploid cell of the body by their second sex chromosome: females have a second X chromosome, whereas males have a Y chromosome. A growing body of evidence ties this difference in sex-chromosome complement to differences between the sexes in the function of non-reproductive tissues. While evolutionary studies have affirmed the importance of Y-chromosome genes beyond the gonad, little is known about how Y-chromosome genes might directly account for sex differences in non-reproductive contexts. Key players are likely to be a group of widely expressed, dosage-sensitive regulatory genes on the Y chromosome that retain similar but non-identical homologs on the X chromosome. The proteins encoded by Y- and X-linked homologs likely remain functionally redundant, though they might have diverged in expression. Through a quantitative survey of Y-chromosome gene expression across 36 adult human tissues, we find that Y-chromosome genes show unexpected variation in their expression across non-reproductive tissues and compared to their X-linked counterparts. A notable example is Y-linked EIF1AY and its X-linked homolog EIF1AX, which encode functionally redundant isoforms of eukaryotic initiation factor 1A (EIF1A), a factor essential for initiating protein synthesis in eukaryotes. Although EIF1AY and EIF1AX show similar levels of expression in most tissues, two Y-specific nucleotide substitutions disrupted a microRNA target site within EIF1AY's 3′ UTR, leading to tissue-specific upregulation of EIF1AY in the heart. As a result of EIF1AY's over-expression, EIF1A protein is nearly twice as abundant in male heart tissue compared to female heart tissue. Much as non-coding variants are thought to contribute significantly to phenotypic variation within and across species, phenotypic differences between the sexes likely stem in part from the Y chromosome's non-coding differences from the X. These differences enable male-specific regulation, and sometimes higher male expression, of important regulatory genes that the Y and X chromosomes share.

# SINGLE-CELL DNA SEQUENCING REVEALS EXTENSIVE CLONAL HETEROGENEITY AND BIASED T-CELL INFILTRATION IN DEDIFFERENTIATED LIPOSARCOMA

Rodrigo Gularte Mérida[1], Timour Baslan[2], Richard W Cass[1], Yawei Shen[1], Arnaud Da Cruz Paula[1], Christina Antonescu[3], Nicholas Socci[4], Tomoyo Okada[1], Samuel Singer[1]

[1]Memorial Sloan-Kettering Cancer Center, Department of Surgery, New York, NY, [2]MSKCC, Cancer Biology and Genetics Program, New York, NY, [3]MSKCC, Department of Pathology, New York, NY, [4]MSKCC, Marie-Josée & Henry R. Kravis Center for Molecular Oncology, New York, NY

Liposarcoma is the most common type of soft tissue sarcoma. Well-differentiated liposarcoma (WDLS) and dedifferentiated liposarcoma (DDLS) represent the most common biologic group of liposarcoma and are characterized by a 12q amplification, complex rearrangements, and copy number alterations (CNA). DDLS are resistant to radiation and to current chemotherapies; even if treated by aggressive surgery, they are often lethal. To understand the clonal architecture of DDLS we carried out single-cell DNA sequencing from five regions taken along the largest central axis of a 30 cm retroperitoneal DDLS. The five regions were high-grade anaplastic tumor concordant with DDLS. Four regions displayed DDLS with features of a rhabdomyosarcoma divergent differentiation (RMSd), and one classified as spindle cell DDLS. Single-nuclei from each region were isolated based on DNA content by FACS into 96 well plates, amplified and indexed by DOP-PCR, pooled, and sequenced. After alignment to the human reference genome GRCh37, and CNA analysis we obtained 797 cells with high-quality CNA profiles (169 DDLS, and 628 DDLS w/RMSd). Tumor/normal cells were distinguished by marked amplifications across CDK4 (n=285), MDM2 (n=281), and HGMA2 (n=280) on 12q13.3-23.1; TRIO (n=281) on 5p15, and TCF21 (n=278) on6q23.2. In addition, cancer cells had a lower ploidy than normal cells ($T=1.8\pm0.11$; $N=2.0\pm0.01$; $p=10^{-16}$). Broad CNA stratified our single-cell dataset into eight cell populations with five sharing a clear cancer lineage and sub-clonal structure. The remaining were partitioned into two normal cell populations and one undetermined population characterized by a loss of one copy of chromosome X. Cancer clones from the DDLS w/RMS regions had 10 extra sub-clonal chromosome and arm level gains or losses than spindle cell DDLS indicating a distinct divergent lineage. An amplification on 19p13.2 was enriched in tumor cells in the spindle cell DDLS. Integer CNA states of MDM2, HMGA2, and CDK4 focal amplifications showed a continuous distribution, and moderate pairwise correlations ($0.57< r <0.71$; $p\leq10^{-16}$) suggestive of double-minute chromosome instability and independent evolution. Lastly, 172 cells had a polymorphic TCRα V(D)J recombination signature in an otherwise normal 2N genome indicative of heterogeneously infiltrating T-cell population. Interestingly, the inferred T-cells were enriched in the spindle cell DDLS, 44% compared to only 15% in the DDLS w/RMSd ($p=10^{-14}$). This suggests a biased T-cells infiltration to the spindle cell DDLS component as compared to the DDLS w/RMSd. In conclusion, our massively parallel, multiregional single-cell DNA sequencing results reveal the extensive heterogeneity and complexity of both spindle cell DDLS and DDLS with RMS differentiation and show the level of T-cell infiltration in each tumor subtype.

# NEW IDEAS FOR GWAS: EPIGENETIC "FINGERPRINTS" TO SYSTEMATICALLY IDENTIFY DISEASE CAUSAL CELL TYPES FOR THOUSANDS OF TRAITS IN THE GWAS CATALOG AND UK BIOBANK

Cong Guo[1], Karsten B Sieber[1], Aidan MacNamara[2], Matthew R Nelson[1], Yu Zhang[3]

[1]GlaxoSmithKline, Human Genetics, Collegeville, PA, [2]GlaxoSmithKline, Functional Genomics, Stevenage, United Kingdom, [3]Penn State University, Biostatistics, University Park, PA

Identifying disease causal cell types and tissues underlying a genetic association is a crucial step towards executing downstream target validation experiments. Target validation can be challenging given that thousands of loci associate with hundreds of complex diseases. Here, we predicted causal cell types for ~1,100 traits in the NHGRI-EBI GWAS Catalog and ~2,000 traits in the UK Biobank by matching genetic association signals with cell-type specific gene regulatory elements. The genome segmentation was conducted on 150 cell types using an Integrative and Discriminative Epigenome Annotation System (IDEAS), a 2D genome segmentation method that accurately annotates regions of the genome with gene regulatory potential developed by Zhang and colleagues. We then developed a method called IDEASpredict that integrates GWAS posterior probabilities from statistical finemapping with the posterior probabilities from regulatory elements defined by IDEAS to estimate the proportion of genetic signal that can be attributed to each cell type for a given trait. Unlike existing enrichment analysis methods, we account for the correlation of regulatory events across cell types and estimate the cell type in which causal variants are most likely to be functional.

The results from IDEASpredict provided new insights for thousands of complex traits. To illustrate, we observed that ~300 of the traits we investigated had > 50% of their signal attributed to immune cells, with the most significant traits being allergies, celiac disease, lupus and Crohn's disease. We were able to further distill the immune component to specific immune cells and subtypes. For example, for rheumatoid arthritis, we found that 60% of the immune signal mapped to T-cells followed by smaller contributions from B-cells and monocytes. We were also able to disentangle mechanisms for related traits. A comparison between ulcerative colitis and Crohn's disease revealed that while Crohn's disease had 2x more immune derived signal, ulcerative colitis had 4x greater epithelium derived signal. In summary, we present a new integrative approach to identify cell type specific enrichments, which will increase our understanding of the various pathways and tissues involved in multifactorial diseases. We expect this approach to substantially improve our ability to take advantage of GWAS to identify gaps in knowledge and new prospective therapeutic targets.

# TARGETED GENOTYPING OF OXFORD NANOPORE READS EMPOWERS CLINICAL APPLICATIONS THROUGH HIGH SINGLE MOLECULE REFERENCE-ANCHORED ACCURACY

Jeffrey A Rosenfeld[1], Marcus Stoiber[2], Melissa delaBastide[3], Tim Massingham[2], Sara Goodwin[3], Shridar Ganesan[1]

[1]Rutgers Cancer Institute of New Jersey, Patology, New Brunswick, NJ, [2]Oxford Nanopore Technologies, Oxford, United Kingdom, [3]Cold Spring Harbor laboratory, Genome Center, Cold Spring Harbor, NY

Sequencing of humans, and other well-characterized organisms, is often directed towards the highly accurate detection of a small number of known or clinically relevant SNPs or small variants. Novel adaptations of the "flip-flop" neural network base caller to specifically propose alternative sequences (SNPs) enables accurate single molecule detection of variants and rapid genotyping of samples. Rather than base calling each read, losing information about alternatives calls and their confidence, we condition the information-rich output of the neural network on a reference followed by specific proposal of modifications greatly increasing the power to correctly identify the correct sequence.

Applying this method to artificial SNPs placed in the genome of the NA12878 shows 99+% single molecule accuracy. For the NA19700, HapMap s samples we were able to match previously identified SNPs from MinIon coverage of less than 5 reads. Using the SKBR3 tumor cell line sample, we were able to confidently identify clinically important SNPs from the COSMIC cancer mutation database.

While Oxford Nanopore sequencing remains vital to de novo discovery via traditional base calling, the ability to extract high confidence, single molecule SNV calls and genotype in real time opens the door to many more clinically relevant applications.

# FROM GENETIC RISK VARIANTS TO CONVERGENT PROTEIN NETWORKS: AN INTEGRATIVE APPROACH TO ELUCIDATE THE CAUSAL MOLECULAR MECHANISMS OF SCHIZOPHRENIA.

Kasper Lage[1,2], Eugene Nacu*[1,3], April Kim*[1,2], Edyta Malolepsza[1,2], Natalie Petrossian[1], William Crotty[1], Tarjinder Singh[1,2], Ruize Liu[1,2], Greta Pintacuda[1,3], Matthew Johnson[1,4], Kamal Sharma[5], SCHEMA Consortium[1], Richard Huganir[5], Beth Stevens[1,4], Mark Daly[1,2], Hailiang Huang[1,2], Jake Jaffe[6], Kevin Eggan[1,3]

[1]Broad Institute, Stanley Center for Psychiatric Research, Cambridge, MA, [2]Massachusetts General Hospital, Department of Surgery, Boston, MA, [3]Harvard University, Department of Stem Cell and Regenerative Biology, Cambridge, MA, [4]Boston Children's Hospital, Department of Neurology, Boston, MA, [5]Johns Hopkins School of Medicine, Department of Neuroscience, Baltimore, MD, [6]Broad Institute, Proteomics Platform, Cambridge, MA

Recent studies have revealed the highly polygenic nature of schizophrenia, but the specific underlying gene networks (pathways) remain obscure in many cases. This is a key bottleneck towards biological understanding and therapeutic intervention.
Human upper layer cortical excitatory neurons are particularly relevant to schizophrenia. We developed a production framework of homogeneous cell populations at scale (>10 billion neurons per year) and analyzed neuron-specific protein-protein interactions of schizophrenia risk genes using tandem mass spectrometry. In parallel, we developed a computational platform (Genoppi) to QC the data, to integrate proteomic and genetic datasets, and to prioritize emerging pathway models for functional validation experiments.
Many of the high-quality protein interactions we identified are unique to human neurons (i.e., not reported in the literature nor seen in non-brain tissues). By integrating data from the SCHEMA exome sequencing consortium and the East Asia Cohort of the Psychiatric Genomics Consortium, we found that the interaction networks of several schizophrenia risk proteins (e.g., CACNA1C, HCN1, and SYNGAP1) are significantly enriched for common and/or rare risk variants, illustrating how common and rare genetic risk can converge onto the same cellular networks in human neurons. Functional studies of interaction partners identified by our approach show that knockdowns at the gene level result in synaptic phenotypes in primary mouse neurons.
Overall, we describe an integrated experimental and computational framework to 1) map interactomes of schizophrenia proteins in human neurons, 2) model pathway relationships through integration of proteomic and genetic data, and 3) functionally validate key schizophrenia network modules in human neurons, human brain tissue, and animal models. Our approach bridges the evolutionary gap between current model organisms and humans to maximize functional insights from recent genetic data.

# PUTATIVE NON-CODING CIS-REGULATORY DRIVERS IN CHRONIC LYMPHOCYTIC LEUKAEMIA AND SKIN CANCER

Halit Ongen, Olivier Delaneau, Nikolaos M Lykoskoufis, Aline Real, Christelle Borel, Michael W Stevens, Cedric Howald, Emmanouil T Dermitzakis

University of Geneva, Department of Genetic Medicine and Develpoment, Geneva, Switzerland

Perturbations of the coding genome and their role in cancer development have been studied extensively. However, the non-coding genome's contribution in cancer is poorly understood, since it is difficult to define the non-coding regulatory regions and the genes they regulate, and there is limited power due to the small size of regulatory regions. We try to resolve this by defining modules of coordinated non-coding regulatory regions of genes (Cis Regulatory Domains or CRDs) by using the correlation between histone modifications, assayed by ChIP-seq, in population samples of immortalized B-cells and skin fibroblasts. We screen for CRDs that accumulate an excess of somatic mutations in chronic lymphocytic leukaemia (CLL) and skin cancer, after accounting for somatic mutation patterns and biases. At 5% FDR, we find 16 CRDs with significant excess of somatic mutations in CLL, 15 of which regulate 44 genes, and in skin cancer 235 significant CRDs, 85 of which regulate 109 genes. We show that these results are unlikely to be false positives through a simulation analysis of mutation rates and patterns. The genes identified include ones already implicated in tumorigenesis, and are enriched in pathways involved in the respective cancers, like the BCR and MAPK signalling pathways in CLL and skin cancer, respectively. Since functional somatic mutations will be in heterozygosity, they will affect allelic expression, so, under selection, we expect significant CRD genes to show allelic imbalance. In an CLL RNA-seq dataset, we tested 32 out of the 44 genes identified for extreme allelic imbalance compared to genes with no significant CRD effects. We find 9 genes have significantly more extreme allelic imbalance ($p = 0.018$), corroborating their driver potential. We discover that the somatic mutations in the significant CRDs of CLL are hitting bases more likely to be functional than the mutations in non-significant CRDs. Moreover, in both cancers, mutational signatures observed in the regulatory regions of significant CRDs deviate significantly from the ones observed in non-significant CRDs. These results indicate selection acting on CRDs during tumorigenesis. We find that the transcription factor biding sites that are disturbed by the somatic mutations in significant CRDs are enriched for factors known to be involved in cancer development. Finally, we are conducting cell culture experiments to show the driver potential of the genes identified. This study describes a new powerful approach to discover non-coding regions involved in tumorigenesis in CLL and skin cancer that could be generalized to other cancers.

# IDENTIFYING DISORDER-SPECIFIC METHYLATION SIGNATURES IN PATIENTS WITH SEVERE DEVELOPMENTAL DISORDERS

Juliet E Handsaker[1], Eugene J Gardner[1], Patrick J Short[1], Guiseppe Gallone[1], Leopold Parts[1], Jeffrey C Barrett[1,2], Matthew E Hurles[1]

[1]Wellcome Trust Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]Genomics PLC, Cambridge, United Kingdom

Chromatin-modifying genes (CMGs) are a diverse set of genes whose protein products bind to chromatin and regulate cellular processes. CMGs typically act via the modification of histones, with effects such as promoting or limiting cell proliferation and regulating gene expression. Unsurprisingly given the importance of CMGs in several higher-order cellular functions, deleterious mutations in over 50 CMGs have been robustly associated with severe developmental disorders (DDs).

Previous studies have shown that perturbations of the histone code impact genome-wide DNA methylation patterns. To determine if individuals with deleterious mutations in CMGs have distinct methylation signatures, we characterised methylation profiles in blood-derived DNA from 128 DD probands with de novo mutations in one of 22 disease-associated CMGs from the Deciphering Developmental Disorders (DDD) study, a UK-wide cohort of 13,600 children with DD. A median of 5 patient samples were selected per CMG and, along with age and sex matched negative controls, were methylation profiled using the Illumina EPIC array.

Using our methylation data as input to a linear correlation model and an elastic net, we sought to identify a CpG methylation signature discriminating between patients and control individuals for each CMG disorder. We were able to establish a signature based on differential methylation for 10 CMGs (*NSD1, DNMT3A, BCL11A, ARID1A, KMT2A, CHD4, CREBBP, EP300, ADNP, KMT2B*), which indicates that a perturbed methylation phenotype is a general characteristic of CMG disorders and that these signatures could be diagnostically predictive. These methylation signatures were validated with additional *de novo* cases and replicated utilising independent data from other studies. Additionally, we applied the methylation signatures to potential cases where the *de novo* status of the variant was unavailable or the severity of the variant was uncertain. These results were concordant with established metrics of determining variant severity (CADD, MPC, LOFTEE), exemplifying the diagnostic potential of our signatures. Overall, our study highlights the utility of DNA methylation profiles as biomarkers supporting the diagnosis of CMG disorders.

# BUILDING BETTER PERSONAL CANCER VACCINES

Catherine J Wu

Dana-Farber Cancer Institute and Harvard Medical School, Department of Medical Oncology, Boston, MA

With the recent availability of novel immunologic agents, priority has shifted to understanding the mechanisms of and predicting responses to each treatment. At the heart of cancer and host immune cell interactions is the tumor antigen and host antigen-specific T cell interaction, with the cytotoxic T cell-cognate antigen interaction forming the mechanistic basis for immune-mediated recognition and the killing of malignant cells. While the search for immunogenic tumor antigens has been the subject of decades-long studies, multiple lines of evidence have convincingly demonstrated tumor neoantigens as an important class of immunogenic tumor antigens. Neoantigens arise from amino acid changes encoded by somatic mutations in the tumor cell and have the potential to bind to and be presented by personal HLA molecules. Using next-generation sequencing approaches, we can now systematically identify mutations leading to amino acid changes that can be potentially recognized immunologically through the implementation of neoantigen discovery pipelines. In recent studies, we have demonstrated that neoantigens can be safely and feasibly targeted to generate customized cancer vaccines. We have been undertaking pilot clinical trials to develop personal cancer vaccines in melanoma and glioblastoma that utilize synthetic long peptides as delivery approach for this therapy. Recent results and new directions will be discussed.

# A NOVEL METHOD FOR JOINT ANALYSIS AND COLOCALIZATION APPLIED TO 10,000 GWAS DATASETS IDENTIFIES MOLECULAR MECHANISMS UNDERPINNING GWAS HITS, REVEALS SURPRISING SHARED BIOLOGY BETWEEN DISEASES, AND INCREASES THE POWER OF POLYGENIC RISK SCORES ACROSS POPULATIONS

Peter Donnelly, Gerton Lunter, Eva Krapohl, Andreas Heger, Reka Nagy, Peter Sorensen, Mike Sivley, Mike Weale, Jeffrey Barrett, Gilean McVean, Vincent Plagnol, Chris Spencer

Genomics plc, Oxford, United Kingdom

We have assembled, harmonized, and imputed 10,000 GWAS summary statistic datasets. This combines genome-wide genetic data on human traits ranging from molecular and cellular phenotypes, through biomarkers, to physiological and cognitive quantitative traits and disease outcomes. Typically, these would be analysed individually or in pairs, as current approaches to joint analysis and co-localisation simply do not scale to large numbers of phenotypes.

Here we describe a novel, scalable, statistical method and the associated computational infrastructure to jointly analyse tens of thousands of GWAS datasets. The core of the approach utilizes model-based clustering in a non-parametric Bayesian framework to identify groups of phenotypes which share the same causal variant locally.

There are major advantages to joint analysis of all 10,000 traits, which we illustrate. (1) Co-localisation of GWAS hits for a disease outcome with molecular/cellular traits can identify the underlying molecular mechanism, pinpoint the relevant gene, and reveal key aspects of the path through human physiology to the disease. (2) The sharing of information across traits can provide compelling evidence of a genetic effect when the marginal evidence is at most equivocal, for example because it lies well below genome-wide significance in the relevant disease GWAS. This identifies often surprising shared biology between different diseases and traits (eg *PDE5A*, the target of Viagra is also associated with diverticulitis, *C1orf106*, a known locus for IBD is associated with multiple other inflammatory diseases and schizophrenia). (3) Identification of multiple phenotypes driven by the same causal variant (genome-wide, our clusters contain a median of 30 and a mean of 57 studies) massively improves power to do finemapping, and thus to identify causal variants.

The approach yields important and novel insights into human biology. One natural application is in drug discovery, particularly to the identification and validation of novel drug targets. A second significant application is to polygenic risk scores (PRS). Our data resource allows us to derive PRSs simultaneously for over 100 diseases. Further, greater precision in identifying causal variants can not only improve the power of PRS but it also improves behaviour of PRSs in other ancestry groups.

# SURVEYING STRUCTURAL VARIATION LANDSCAPES IN SINGLE CELLS BY STRAND-SPECIFIC SEQUENCING REVEALS COMPLEX GENOMIC REARRANGEMENT PROCESSES.

Ashley D Sanders*[1], Sascha Meiers*[1], Maryam Ghareghani*[2,3], David Porubsky*[2,3], Tobias Marschall**[2,3], Jan O Korbel**[1,4]

[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, [2]Saarland University, Center for Bioinformatics, Saarbrücken, Germany, [3]Max Planck Institute, Max Planck Institute for Informatics, Saarbrücken, Germany, [4]University of Heidelberg, Molecular Medicine Partnership Unit (MMPU), Heidelberg, Germany

*authors contributed equally **senior authors contributed equally

Somatic structural variation (SV) is a major source of human genetic variation. It can arise in rapid bursts mediating large-scale genomic rearrangements and disrupt cancer-related pathways. Mapping the landscape of SVs in single cells is essential to understand how these variants arise in healthy tissues and when they lead to disease. Until now, comprehensive detection of SVs in heterogeneous cell samples has remained challenging because subclonal copy-neutral and complex rearrangements are largely intractable. To address this, we developed scTRIP, a novel methodology that integrates three complementary data layers: read depth, template strand and haplotype phase, in strand-specific sequence data to systematically discover SVs in a cell-by-cell basis. We applied scTRIP to survey the SV landscapes of 565 single-cell genomes, including transformed epithelial cells and patient-derived leukemic samples. In these single cells, we identified a wide variety of SV types that previously escaped detection, including translocations, inversions, repeat-embedded and copy-neutral rearrangements leading to oncogene dysregulation. We compiled these data to build a 'molecular karyotype' of the patient samples and simultaneously examined the clonal structure of the tumour. Finally, our analysis enabled us directly measure complex rearrangement processes, such as breakage-fusion-bridge cycles and subclonal chromothripsis events. The ability to now map SV landscapes with single cell resolution positions us to study dynamic clonal expansions, genetic mosaicism and structural rearrangement processes with a new lens, which can provide a clearer picture of the processes behind development, aging, and disease.

# TREX1 AND APOBEC3B INDUCE CHROMOTHRIPSIS & KATAEGIS

Aikaterini Chatzipli[1], John Maciejowski[2], Peter J Campbell[1], Titia de Lange[3]

[1]Memorial Sloan Kettering Cancer Center, New York, NY, [2]Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [3]Rockefeller University, New York, NY

Telomere crisis occurs during tumorigenesis when depletion of the telomere reserve leads to frequent telomere fusions. The resulting dicentric chromosomes drive genome instability and give rise to aneuploidy, chromosome rearrangement, and the clustered mutational phenomenon chromothripsis and kataegis. Kataegis mutations clusters are characterized by a proximity to rearrangement breakpoints, and exhibit a C>T or C>G mutational signature at TpC dinucleotides. This signature has proposed links with the APOBEC family of cytosine deaminases, but mechanistic certainty and the identity of the primary driver of mutagenesis have remained elusive. Here, we show that APOBEC3B is solely responsible for the induction of kataegis in a model system of telomere crisis. Mutagenesis occurs independently of APOBEC3B upregulation, but depends on resection and the exposure of single stranded DNA by the cytoplasmic 3' nuclease TREX1. We also show that APOBEC3B editing promotes the induction of chromosome rearrangement breakpoints on chromothripsis chromosomes. We propose that uracil excision and consequent apurinic/apyrimidinic cleavage enable APOBEC3B to fragment chromosomal DNA.

# QUANTIFYING THE CONTRIBUTION OF X-LINKED CODING VARIANTS TO DEVELOPMENTAL DISORDERS

Hilary C Martin[1], Ana Lisa Taylor Tavares[1,5], Kaitlin E Samocha[1], Joanna Kaplanis[1], Nadia Akawi[2], Alejandro Sifrim[3], David R FitzPatrick[4], Helen V Firth[5], Matthew E Hurles[1]

[1]Wellcome Sanger Institute, Wellcome Sanger Institute, Wellcome Genome Campus, United Kingdom, [2]University of Oxford, Radcliffe Department of Medicine, Oxford, United Kingdom, [3]KU Leuven, Laboratory of Reproductive Genomics, Leuven, Belgium, [4]MRC Institute of Genetics & Molecular Medicine, MRC Human Genetics Unit, Edinburgh, United Kingdom, [5]Cambridge University Hospitals NHS Foundation Trust, East Anglian Medical Genetics Service, Cambridge, United Kingdom

The Deciphering Developmental Disorders study is a large cohort of >13,000 exome-sequenced patients with severe undiagnosed developmental disorders (DDs). Within the cohort, there are 1.4-times more males than females, and we hypothesized that this male bias might be due to variants on the X chromosome. In fact, we found that ~7% of patients of both sexes can be attributed to rare and de novo X-linked coding variants, implying that the male bias in the cohort is not driven by such variants. About 80% of this burden is in known DD genes, and 35% is due to loss-of-function (LoFs) variants versus 65% to missense variants or inframe indels ("functional variants"). In contrast, in ClinVar, 55% of the reported pathogenic variants in X-linked DD genes are LoFs, implying an ascertainment bias in clinical practice. We found that, in known X-linked DD genes, >80% of de novo LoF and functional mutations and of ultra-rare inherited LoFs are likely to be pathogenic in males, versus only ~18% of maternally-inherited ultra-rare missense variants, highlighting the care that needs to be taken with clinical interpretation of such variants. In known X-linked recessive DD genes, 23% of the pathogenic variants in males were de novo, which is consistent (chi-squared p=0.2) with a theory developed by Haldane in 1935.

To maximise power for X-linked gene discovery, we analysed de novo mutations and inherited variants in males alone to identify X-linked recessive genes, de novo mutations in males and females together to identify X-linked dominant genes, and de novo mutations in females alone to identify X-linked semi-dominant (male-lethal) genes. Of the 25 genes that passed multiple-testing correction, all were already known, but our results nonetheless allow us to start to delineate the different classes of X-linked disorder. For example, MED12 is annotated in the literature as X-linked recessive, but we found a significant enrichment of de novos in females, implying that it also has an X-linked dominant mode. For DDX3X and MECP2, we saw significant enrichment of de novos in females compared to males, confirming that these genes are X-linked semi-dominant and that LoFs are male-lethal.

These results shed important light on the genetic architecture of developmental disorders and indicate that more needs to be done to understand the male bias. Our results should also inform diagnostic testing by better delineating the inheritance modes of X-linked disorders and providing prior probabilities that particular classes of variation are pathogenic.

# MAPPING THE FUNCTIONS OF HUMAN DISEASE RISK ALLELES DURING NEURONAL DEVELOPMENT USING MULTIPLEXED DIFFERENTIATION OF HUMAN INDUCED PLURIPOTENT STEM CELLS

Julie Jerber[1,2], Daniel Seaton[3], Daniel Pearce[2], Oliver Stegle[3], Florian Merkle[4], Daniel Gaffney[2]

[1]Open Targets, Wellcome Genome Campus, Hinxton, United Kingdom, [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [3] EMBL-EBI, Wellcome Genome Campus, Hinxton, United Kingdom, [4]Metabolic Research Laboratories and Medical Research Council Metabolic Diseases Unit, Wellcome Trust-Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge, United Kingdom

Human Induced pluripotent stem cells (hiPSCs) offer a controlled experimental system to investigate the impact of genetic variation in otherwise inaccessible cell lineages and states. When coupled with single cell RNA-seq, hIPSCs can also be used in highly efficient "multiplexed" experimental designs, where cell lines from large numbers of donors can be cultured and differentiated in pools, then deconvoluted, enabling large sample sizes to be achieved even in complex differentiation systems. Here, we used this approach to differentiate 145 hIPSC lines to functional midbrain dopaminergic neurons, a key pathogenic cell type in Parkinson's disease (PD). Our design enabled us to collect samples across 4 differentiation time points, including a stimulation with the PD-associated compound, rotenone, while controlling for between-differentiation batch effects.

By comparing with in vivo human data, we identified 7 primary cell types generated during the course of our differentiation, including quiescent and mitotic floorplate progenitors, floorplate neuroblasts, ependymal cells, astrocytes, dopaminergic and serotonergic neurons. We generated the first maps of expression quantitative trait loci (QTLs) in each of these lineages, both in quiescent cell states, and in response to cellular stress induced by rotenone exposure. We colocalized disease-associated loci for multiple neurological disorders, including PD, Alzheimer's disease and schizophrenia with eQTLs in developing neural cell types, and identified the stages and cell types in which they become active. We identified a subset of PD risk variants with effects that change upon treatment by rotenone suggesting potential mechanisms for genotype-environment interaction effects. Our study demonstrates how scRNA-seq coupled with hiPSC technology can be used to efficiently perform large population studies in inaccessible cell states and lineages, and provides a novel framework for understanding the function of disease risk loci function across many cell types and states in in a single model system.

# GENERATION OF THE FIRST REGULATORY MAP OF HUMAN PRIMARY MICROGLIA FROM 141 NEUROSURGICAL PATIENTS.

Fiona Calvert*[1], Adam Young*[2,3], Natsuhiko Kumasaka*[1], Robin Franklin[3], Daniel Gaffney[1,4]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridgeshire, United Kingdom, [2]Division of Academic Neurosurgery, Department of Clinical Neurosciences, Cambridge University Hospital, Cambridge, United Kingdom, [3]Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom, [4]Open Targets, Wellcome Genome Campus, Cambridgeshire, United Kingdom

Dementia is estimated to affect nearly 50 million people worldwide, with that number expected to triple by 2050. Alzheimer's disease (AD), the most common form of dementia, is complex with many factors appearing to affect disease development and progression. Multiple lines of evidence, including large-effect AD risk variants in myeloid-specific genes such as TREM2 and a striking enrichment of AD loci risk variants near genes with immune function, have suggested that microglia, the tissue resident macrophages of the CNS, as key players in AD. Here, we present the largest collection to date of single cell and bulk RNA-seq data from human primary microglia and macrophages, collected from 141 neurosurgical patients from a range of clinical presentations and ages. We combine bulk RNA-sequencing data with genotype information to generate the first eQTL map in primary microglia and colocalise with genome wide association studies (GWAS) for a range of neurodegenerative disorders . We identified multiple AD risk loci that colocalise with an eQTL in primary microglia. We also identified multiple loci where an AD association signal colocalises with an eQTL in a immune cell type, but not microglia, suggesting these loci may function in infiltrating rather than resident immune cells. Using our single cell data, we illustrate how the population composition of microglia changes between different clinical phenotypes, sex and age and identify clinical pathology as the driver of much of the microglial heterogeneity. Our work provides the largest resource of adult human primary microglia RNA-sequencing to-date and our large sample size allows for robust statistical comparison of transcriptomes across clinical groups as well as the generation of the first eQTL map within this cell type.

# TRACING THE PEOPLING HISTORY OF THE SOUTHWEST PACIFIC THROUGH WHOLE GENOMES

Javier Mendoza-Revilla*[1], Jeremy Choin*[1], Lara R Arauna[1], Christine Harmant[1], Maximilian Larena[2], Olivier Cassar[3], Jean F Deleuze[4], Antoine Gessain[3], Mattias Jakobsson[2], Mark Stoneking[5], Etienne Patin[1], Lluis Quintana-Murci[1]
*These authors contributed equally to this work

[1]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France, [2]Uppsala University, Department of Organismal Biology, Uppsala, Sweden, [3]Unit of Epidemiology and Physiopathology of Oncogenic Viruses, Institut Pasteur, Paris, France, [4]CNRGH, CEA, Evry, France, [5]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany

The Southwest Pacific region represents both the endpoint of the early migration of modern humans out-of-Africa and of the most recent human expansion into empty territories. Further evolutionary interest in the region comes from the fact that populations inhabiting the these Pacific islands present the highest combined archaic ancestry at the worldwide scale. Recent studies indicate that genetic structure in this region is driven by the levels of Papuan- and Austronesian-like ancestry. Yet, the demographic history of the Southwest Pacific in terms of population splits, effective population sizes and introgression from archaic hominins based on whole-genome sequence (WGS) data is missing.
Here, we generated novel high-coverage (~40×) WGS data from more than 300 unrelated individuals from 15 populations covering the whole transect at the basis of the peopling of the Pacific, including Taiwan, the Philippines, and Near and Remote Oceania (Solomon Islands, Santa Cruz and Vanuatu), which were supplemented with available data from Papua New Guinea and the Bismarck archipelago. First, using a simulation-based composite likelihood method, we dated the earliest separation among Near Oceanian populations to the Late Pleistocene (~20,000-30,000 years ago), suggesting an early population structure following the settlement of the region. Marked differences in effective population sizes were also found between Austronesian-speaking Taiwanese aboriginals and Filipinos, consistent with the occurrence of demographic growth among the latter during the Holocene. Focusing on archaic introgression, while the levels of Neanderthal ancestry were relatively homogeneous (~2.1 to 2.7%), those of Denisovan ancestry varied markedly across populations (0 to ~2.9%). Denisovan ancestry was found to be highly correlated with Papuan-like ancestry (r2~0.8), with the exception of the Agta foragers from the Philippines. This suggests two independent introgression events, a model that is further supported by our coalescent-based simulations. Finally, we detected a depletion of Denisovan ancestry centred on promoters, in agreement with recent findings based on Neanderthal ancestry, highlighting the key evolutionary importance of these gene regulatory regions.
Collectively, our analyses provide novel insights into the genetic history of Southwest Pacific populations, including independent interactions with archaic hominins, early splits during the Palaeolithic, and more recent fluctuations in population sizes associated to the range expansion of Austronesian speakers.

# MONOGENIC AND POLYGENIC INHERITED FORMS OF CLONAL SELECTION

Po-Ru Loh[1,2], Giulio Genovese[2,3,4], Steven A McCarroll[2,3,4]

[1]Brigham and Women's Hospital and Harvard Medical School, Division of Genetics, Department of Medicine, Boston, MA, [2]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [3]Broad Institute of MIT and Harvard, Stanley Center for Psychiatric Research, Cambridge, MA, [4]Harvard Medical School, Department of Genetics, Boston, MA

Clonal expansion of hematopoietic cells with acquired mutations (clonal hematopoiesis, CH) is a common, age-acquired condition with many adverse health outcomes. To identify mutations and alleles that lead to clonality, we identified 17,113 cases of CH with 19,635 autosomal mosaic chromosomal alterations (mCAs, including deletions, duplications, and copy-number-neutral loss of heterozygosity [CNN-LOH]) among 482,857 UK Biobank participants. Analysis of these acquired mutations, along with inherited variation in the same genomes revealed seven monogenic-inherited vulnerabilities to the development of expanded clones with CNN-LOH events involving 30 rare, large-effect inherited risk variants (OR>10). At all seven loci, the inherited alleles appeared to be substrates for clonal selection: subsequent, acquired mutations systematically replaced the inherited risk allele (at MPL) or duplicated it to the other chromosome (at *FH, NBN, MRE11, ATM, SH2B3*, and 15q26.3). Three of the seven genes with large-effect inherited alleles (*MRE11A, NBN, ATM*) encode components of the MRN-ATM pathway, which limits cell division after a cell experiences DNA damage.

In addition to these monogenic inherited forms of CH, we found a common, pervasive and surprisingly polygenic substrate of clonal selection: CNN-LOH events across the genome tended to cause chromosomal segments with inherited sets of alleles that more strongly promote proliferation to replace their homologous (allelic) counterparts.

This dynamic reveals a basic challenge for lifelong cytopoiesis in any genetically diverse species: many individuals inherit unequal proliferative potentials on paternally and maternally derived chromosome-pairs, and readily-acquired mutations that replace chromosomal segments with their homologous counterparts give selective advantage to mutant clones.

# NEANDERTHAL FRAGMENTS LENGTHS AND MUTATIONAL SIGNATURES SHOW DIFFERENT HISTORICAL GENERATION TIMES IN NON-AFRICAN POPULATION OVER THE PAST 50,000 YEARS

Moises Coll*[1], Laurits Skov*[1,2], Mikkel H Schierup[1]

[1]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark, [2]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany

Modern humans and Neanderthals interbred around 50,000 years ago, leaving 1-3% Neanderthal DNA in all present day non-African populations. These Neanderthal fragments have since been broken down into small pieces by recombination. The length distribution of these fragments is informative about the number of generations that has passed since the admixture event and on the number of admixture events. We detect Neanderthal fragments in the individuals of Simons Genome Diversity Project and compare the fragment length distribution among the different continental populations. We find that fragment length distributions differ systematically across the world, e.g. with a mean length in West Eurasian population of 73 kb and a mean length of 82 kb in East Asian populations. Since we can show that population specific Neanderthal admixtures are unlikely to affect the length distribution, the recombination clock must have run at different speeds across the world, with the fastest speed in West Eurasia and slowest in East Asia. We also find that fragment lengths in ancient humans samples are longer in ancient European samples than in samples from East Asia and America. These differences can be explained by historical generation times differing by up to 20% between populations.

Mutation accumulation is also informative about generation time since the mutational spectrum of the offspring depends on paternal and maternal age at conception. We therefore studied the accumulation of derived alleles the same data after first removing all regions affected by archaic introgression and SNPs seen in Africa. We estimate recover mutational spectrum differences between populations previously reported, including the C -> T transitions in a TCC three base pair context with high abundance in Europeans. We then correlated for each individual the relative frequency of mutation types with the mean of its archaic fragment lengths. We observe a strong correlation with the specific mutational types that are strongly affected by parental age in the deCODE dataset of 1536 trios. Differences are also consistent with generation time differences among non-African populations of about 20%.

Our two independent measures of historical generation time correlate strongly in that individuals with longer Neanderthal fragments have a higher fraction of the mutation types that are enriched in older parents. We estimate that Europeans have had the shortest historical generation time. They also have slightly more derived mutations in accordance with this, and the European enrichment in C -> T in TCC context can be explained by differences in generation times alone. Our results imply that the human generation time has been dependent on some environmental or cultural differences in ancient societies, perhaps related positively with population densities or grain based versus meat-based life styles.

# TISSUE VARIABILITY OF REGULATORY VARIANT EFFECTS AND THE ROLE OF TRANSCRIPTION FACTORS

Elise D Flynn[1,2], Stephane Castel[1,2], Xiaoting Li[3], Hanrui Zhang[4], GTEx Consortium[5], Raag Agrawal[3], Harmen J Bussemaker[3], Pejman Mohammadi[1,2,5], Tuuli Lappalainen[1,2]

[1]Columbia University, Systems Biology, New York, NY, [2]New York Genome Center, New York, NY, [3]Columbia University, Biological Sciences, New York, NY, [4]Columbia Univerity, Medicine -- Cardiology, New York, NY, [5]GTEx Consortium, Bethesda, MD, [6]The Scripps Research Institute, Integrative Structural and Computational Biology, La Jolla, CA

The GTEx Consortium has identified thousands of genetic regulatory variants (local expression quantitative trait loci [cis-eQTLs]) that affect gene expression in multiple tissues. Over a third of GTEx eQTLs are active in most tissues, often with varying effect sizes, but the molecular mechanisms of cross-tissue eQTL effect size variability are poorly understood. Since eQTLs are enriched in transcription factor binding sites (TFBS), we developed a model of eQTL effect size as a function of transcription factor (TF) levels across tissues.

We investigated mechanisms of tissue variability of eQTL effect size in the GTEx v8 dataset (838 individuals with WGS, 15,201 RNA-seq across 49 tissues). We first examined the correlation of eGene expression with eQTL absolute effect size across tissues and identified 1,864 expression-correlated eQTLs at a 5% FDR, with 51% positive and 49% negative correlations. This shows that expression level and genetic regulatory effect are not independent phenomena. We then proceeded to correlate two measures of TF activity (one based on TF expression, the other as inferred by genome-wide target gene expression) with eQTL effect size across tissues. We find that over a third of all tested eQTLs correlate with the activity of at least one TF, and we found an enrichment of the relevant TFBS overlap among the TF-correlated eQTLs. We highlight an interesting example of PU.1 regulation of a LIPA gene eQTL that colocalizes with a GWAS signal for coronary artery disease.

Our results provide a basis for identifying regulatory mechanisms of eQTL activity and for understanding eQTL effect size variability in the context of TF activity. This has major implications for future work investigating context-specific mechanisms of GWAS loci and their effects on human phenotype and disease.

# DETECTING SIGNATURES OF SELECTION AND LOCAL ADAPTATION IN UNITED STATES *BOS TAURUS* BEEF CATTLE

Troy N Rowan[1], Robert D Schnabel[1,2], Jared E Decker[1,2]

[1]University of Missouri, Division of Animal Science, Columbia, MO,
[2]University of Missouri, Informatics Institute, Columbia, MO

Cattle are under intense selection pressures, both artificial and natural, making them an interesting system in which to study how selection acts on short timescales. This project leverages temporally-stratified genotypes to identify genomic changes that have occurred in cattle populations over the last ~50 years. Additionally, using geographic information and historical climate data, we identify unique signatures of selection associated with regional environments. Using an evolutionary approach to study artificial selection allows us to create deliverable results for use in genetic evaluations while providing important insights into adaptation biology.

Genotypes for 22,000 Red Angus cattle, born from 1970 to 2016, were phased with Eagle v2.4 and imputed with Minimac3 up to ~850,000 SNPs. To identify selection occurring within the Red Angus breed, we performed Generation Proxy Selection Mapping. This method implements a genome-wide linear mixed model that uses an individual's birth year as a proxy for generation number as the phenotype to identify shifts in allele frequency over time more significant than those caused by drift alone. This analysis found that genomic regions near the PRDM9 and TG genes have been under strong selection in Red Angus cattle over the last ~50 years. This model also showed signatures of polygenic selection throughout the genome. We also identified multiple different genomic regions under selection using SmartPCA with a PCA-based test for selection.

We leveraged the diverse geographic distribution of our samples to further identify regions of the genome contributing to local adaptation. We used FLK selection scans to identify variants under strong selection in each of our 9 distinct regional environments. In addition, we used 30-year mean temperature, precipitation, and elevation measurements for each sample as phenotypes in univariate and multivariate genome-wide association analyses. These tests identify loci associated with these environmental gradients, and these variants are presumably under selection. While univariate analyses were underpowered, the multivariate analysis identified 8 genomic regions strongly associated with continuous climate attributes. A GxE Generation Proxy Selection Mapping model tested for significant interactions between historical allele frequency changes and climate variables. Loci associated with regional or environmental differences are being used to calculate region-specific genomic predictions via genomic feature selection.

# POPULATION-SCALE ANALYSIS OF COPY NUMBER VARIATION FROM WHOLE GENOME SEQUENCING

Robert E Handsaker[1,2], Seva Kashin[1,2], Christina Usher[1,2], Christopher Whelan[1,2], Michele T Pato[3], Carlos N Pato[3], Michael Boehnke[4], 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]SUNY Downstate Medical Center, Brooklyn, NY, [4]University of Michigan, School of Public Health, Ann Arbor, MI

Human genomes exhibit abundant structural variation; this class of variation has been shown to impact many important human phenotypes. However, the full spectrum of structural variation in humans has not been as deeply or as thoroughly ascertained as fine-scale variation (e.g. SNPs and small indels), especially in large populations.

We have performed an initial analysis of copy number variation – including deletions, duplications and multi-allelic CNVs – from whole-genome sequencing a cohort of 13,713 individuals from the Whole Genome Sequencing in Psychiatric Disorders (WGSPD) consortium. This cohort is comprised of schizophrenia cases, controls and unaffected family members and includes individuals of diverse ancestry including more than 7,500 African Americans and 250 Latinos. Accurate and efficient methods for ascertaining copy number variation at this scale from whole-genome sequencing data have been lacking. To address this need, we have enhanced our Genome STRiP CNV research software to efficiently process cohorts of this size in a cloud computing environment.

To date, we have ascertained more than 60,000 copy number variants of length greater than one kilobase, including variants in many repetitive regions of the genome. We have performed uniform genotyping of these variants across the cohort to allow accurate characterization of allele frequencies and to place the structural alleles on SNP haplotypes. Most of this variation appears to be segregating in human populations at allele frequencies below 1%. We estimate that 58% of all protein-coding genes are impacted by at least one copy number variant ascertained in this cohort.

This set of broadly genotyped CNVs will make it possible to begin to understand the patterns of rare, segregating copy number variation in large populations and across different ancestry groups. To facilitate access to this resource, we have developed an interactive web interface to allow researchers to search and explore this data set and view the evidence for each variant.

# HARLEE (HADOOP ARCHITECTURE LAKE OF EXOMES) FACILITATES THE DISCOVERY OF CANDIDATE MENDELIAN DISEASE GENES THROUGH A GENOCENTRIC APPROACH

Adam W Hansen[1,2], Mullai Murugan[2], He Li[2], Michael Khayat[1,2], Liwen Wang[2], Jill Rosenfeld[1], Shalini Jhangiani[2], Zeynep Coban Akdemir[2], Fritz Sedlazeck[1], Donna Muzny[1,2], Aniko Sabo[1,2], Jennifer E Posey[1,2], Michael F Wangler[1], Christine Eng[1], Vernon R Sutton[1], James R Lupski[1,2,3,4], Eric Boerwinkle[2,5], Richard A Gibbs[1,2]

[1]Baylor College of Medicine, Molecular and Human Genetics, Houston, TX, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [3]Baylor College of Medicine, Pediatrics, Houston, TX, [4]Texas Children's Hospital, Genetics, Houston, TX, [5]UTHealth, School of Public Health, Houston, TX

The advent of inexpensive, clinical exome sequencing (ES) has led to the accumulation of genetic data from thousands of samples from individuals affected with a wide range of diseases, but for whom the underlying genetic and molecular etiology of their clinical phenotype remains unknown. These data, when studied in aggregate, allow for a genocentric approach to elucidating disease-gene associations, where a molecular diagnosis delineates a clinical entity enabling disease cohorts that are initially defined by pathogenic variants affecting the same gene. By integrating gene-level annotations, variant-level annotations, and known gene/disease-association data in an Apache Hadoop data lake, we developed a tractable genocentric approach to the study of ultra-rare variation in Mendelian disease, identifying 154 candidate Mendelian disease genes. This Hadoop ARchitecture LakE of Exomes (HARLEE) currently includes data from 18,696 ES tests implemented on subjects with clinical phenotypes, that prompted a visit to a physician, and some blood relatives often including parents of probands. Family based genomics studies using these data have identified 154 candidate Mendelian disease genes. In the months following our initial analysis, 13 of the 154 candidates were reported as causative of a Mendelian trait. The HARLEE environment now enables ongoing prospective gene discovery efforts allowing molecular insights into perturbations of biological homeostasis presenting as disease.

# A CRYPTOCURRENCY FUNDED, BLOCKCHAIN NOTARIZED AND REVIEWED, PUBLICLY AVAILABLE AND USABLE REFERENCE GENOME OF *CANNABIS SATIVA L.*

Timothy T Harkins[1], Brittany Sexton[2], Yvonne Helbert[1], Liam T Kane[1], Heather Ebling[1], Lei Zhang[1], Biao Liu[1], Zachary Eaton[1], Luo Sun[2], Eileen Dimalanta[2], Sarah Kingan[3], Primo Baybayan[3], Maximilian Press[4], William Barbazuk[5], Keerthana Krishnan [2], Kevin J McKernan[1]

[1]Medicinal Genomics, R&D, Beverly, MA, [2]New England Biolabs, R&D, Ipswich, MA, [3]Pacific Biosciences, R&D, Menlo Park, CA, [4]Phase Genomics, R&D, Seattle, WA, [5]University of Florida, Department of Biology and the Florida Genetics Institute, Gainsville, FL

As the cannabis industry is on track to become a $50B+ industry, it would benefit from having access to a freely available high quality, reference genome to help standardize strain types, properly identify and classify them, provide a foundation for marker assisted breeding projects, and reduce the risk of IP that threatens to throttles the entire industry. As seen in many other economically important crops, advanced genetic tools can accelerate the production yields, focusing on narrowly defined traits. The basis for these genetic tools is access to a usable, reference genome.

Medicinal Genomics has recently blockchain published the most contiguous and complete reference genome for Cannabis sativa L. Recently, the team has made further improvements in the assembly, boosting the N50 contig length to over 7.6Mb with less than 400 contigs representing the genome and a BUSCO score of >98%. The addition of HiC data has generated scaffolds of N50 83Mb and when small (<50kb contigs are removed) the overall assembly is in 58 scaffolds, representing the 10 chromosomes.

To demonstrate the usability of this assembly, the team has completed both RNA-Seq and whole genome methylation studies to identify unique transcripts along with corresponding regulatory elements. A few examples include the identification of the regulatory regions for the Edestin and Cs2S genes along with two other sites important for seed production. We observe the CBCA, THCA and CBDA synthase gene clusters, important for cannabinoid production, have been phased onto respective contigs demonstrating tandem repeat expansions. Several gene families that contain up to 30 paralogs, including the important THCA Synthase and CBCA Synthase genes have been identified. Interestingly, many of these sequences contain full length genes with most variation occurring at the 3' and 5' termini. By using RNA-Seq, we are actively annotating and identifying which sequences are expressed and biologically active. To ensure the highest quality assembly is public and open access, we recruited 3 independent reviewers to comment, challenge, and improve the assembly published on the DASH blockchain, enshrining those reviews as part of the open publication process.

# ASSOCIATING GERMLINE VARIANTS WITH SOMATIC CHROMOSOME BREAKS IN CANCER

Meng Xiao He[1,3,4], Sahar Shahamatdar[2], Amaro Taylor-Weiner[4,3], Travis Zack[5], Sohini Ramachandran[2], Eliezer Van Allen[3,4]

[1]Harvard University, Biophysics Graduate Program, Boston, MA, [2]Brown University, Center for Computational Molecular Biology, Providence, RI, [3]Dana-Farber Cancer Institute, Medical Oncology, Boston, MA, [4]Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, [5]University of California, San Francisco, San Francisco, CA

Many cancer genomes contain rearrangements and somatic copy number alterations arising from chromosome breaks. Features of the local genomic context, such as presence of certain retrotransposons, influence the locations of such breaks. Additionally, in myeloproliferative neoplasms and in blood from cancer-free individuals, inherited variants are associated with specific chromosomal alterations. Here, we attempted to broadly identify associations between inherited variants and the location of somatic breakpoints across cancers.

We used Affymetrix Genome-Wide Human SNP Array 6.0 data from matched tumor-normal pairs from TCGA (n = 5,923 pairs remaining after quality control and selecting individuals of predominantly European ancestry). Using normal samples, we called germline genotypes with Birdseed v2. Somatic chromosomal segment calls were generated via allele-specific copy number analysis using both normal and tumor samples for each individual using ASCAT v2.5.1. To guard against false positive associations resulting from contaminating segment calls corresponding to germline structural variation, we removed segments overlapping known germline structural variants. Subsequently, intervals between somatic chromosomal segments were called as breakpoints.

For each non-overlapping 5 megabase bin of the genome, we assigned binary phenotypes corresponding to whether a somatic breakpoint was detected in a given individual. We then performed an association analysis for each phenotype, adjusting for sex and ancestry via logistic regression with plink2. Analyses were performed separately for each disease type followed by a meta-analysis of association summary statistics with METAL.

We identified several candidate associations between inherited variants and somatic breakpoints in specific regions of the genome (using a p-value cutoff of 5e-8 for each phenotype). Our results suggest that heritable genetic variation may be associated with risk for localized somatic chromosome breaks. Ongoing work is aimed at validating the putative associations (for example using whole genome sequencing data from the PanCancer Analysis of Whole Genomes study) and identifying potential mechanisms of putative risk alleles.

# UNDERSTANDING REGULATION OF mRNA N6-METHYLADENOSINE (M6A) AND ITS CONTRIBUTIONS TO HUMAN DISEASES BY QTL MAPPING

Zijie Zhang[1], Kaixuan Luo[2], Maguanyun Qiu[1], Matthew Stephens[2], Xin He[2], Chuan He[1]

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

N6-methyladenosine (m6A) plays a critical role in regulating various aspects of mRNA metabolism and translation in eukaryotes. Despite rapid progresses in this field, little is known about how DNA sequence variations may affect the m6A modification and the role of m6A in common diseases. To answer these questions, we mapped ~88K genetic variants associated with m6A levels in 60 Yoruba lymphoblast cell lines. By analyzing these variants, we uncovered several factors affecting m6A installation, including binding by specific RNA binding proteins (RBPs), RNA secondary structure, and transcriptional processes. Our joint analysis of QTL data of m6A and related molecular traits suggests that the downstream effects of m6A are highly context-dependent. We discovered several RBPs that may mediate functions of m6A, most of which are not known m6A reader proteins. Integrated analysis with GWAS data highlights m6A variation as an important mechanism linking genetic variations to complex diseases.

# MECHANISMS OF TISSUE SPECIFIC GENETIC REGULATION REVEALED BY LATENT FACTORS ACROSS eQTLs

Yuan He[1], Kaushik Srinivasan[2], GTEx Consortium[3], Christopher D Brown[4], Alexis Battle[1,5]

[1]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, Department of Applied Mathematics and Statistics, Baltimore, MD, [3]GTEx, Consortium, MA, [4]University of Pennsylvania, Department of Genetics, Philadelphia, PA, [5]Johns Hopkins University, Department of Computer Science, Baltimore, MD

Genetic regulation of gene expression varies across tissues in complex patterns, ranging from universal effects, to effects shared among a subset of tissues, to highly tissue-specific effects. However, it remains unclear how these complex patterns of tissue specific gene regulation arise or what the functional consequences of such variability are. To systematically explore the patterns and mechanisms of genetic regulation of gene expression across tissues, we use gene expression quantitative locus (eQTL) data across 49 human tissues from the Genotype-Tissue Expression (GTEx) project. We learn a lower-dimensional representation of eQTL effect sizes across tissues by sparse semi-nonnegative matrix factorization. The learned factors include a dense factor representing effects shared across all tissues, factors representing effects shared only among brain regions, and factors specific to arteries, skin, and gut, together demonstrating biologically relevant patterns of genetic effects across tissues. By identifying the factors that appear to underlie each eQTL, we are able to more precisely categorize eQTLs according to the combination of tissues each affects, and use the implicated variants to investigate tissue-specific mechanisms. We find that for each factor, eQTL variants are significantly enriched in enhancers and DNAse hypersensitive regions in matched corresponding tissues. We also identify specific transcription factor binding sites that are enriched for tissue-specific eQTL variants. These include known key regulators in relevant tissues, such as *NEUROD1*, *BHLHE41* and *CLOCK* for variants affecting brain regions, and *MYOD1* and *MYOG* for muscle. Together these results provide a framework to interpret tissue specificity of eQTLs and help characterize the mechanisms that influence this tissue specificity.

# APPLICATION OF VARIATION GRAPH FOR THE DETECTION OF ADDITIONAL SOMATIC MUTATIONS

Cecilia Wieder, Garima Khandelwal, Javier Herrero

Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London, United Kingdom

It has now become evident that a using linear structure for the human reference genome is a limiting factor and could bias the alignment of short reads. New graph representations have been developed and used recently to capture and handle the variability in the human population, reduce the reference genome bias, improve the alignment of short reads and finally enable the detection of additional rare variants that were otherwise undetected as the corresponding supporting reads would not align correctly.

Here we present a proof-of-concept on the value of using a graph-based approach for calling somatic variants in a whole-exome tumour sample from a multiple myeloma patient. We use the **variation graph (vg)** framework to build a graph based on the reference genome and the germline mutations. In order to compare the effect of using a graph instead of a linear reference genome, we call variants from the vg alignment and from the standard alignment using the same methodology (VarScan).

The original germline mutations are derived using GATK4. The vg graph is built using the GRCh38 reference genome and the germline mutations. On the one hand, the reads from the tumour sample are aligned to the graph using 'vg map'. This creates a GAM file (a graph-like BAM file) which is further surjected onto the linear reference genome to obtain a BAM file. On the other hand, the same reads are aligned to the GRCh38 linear reference genome with bwa. Variants are called on both alignment and compared.

We observe an increase in the number of called somatic mutations. Visual inspection reveals that, as expected, some of the newly detected somatic mutations lie next to existing germline mutations. This coincides with additional reads being aligned on those loci as the graph-based approach overcomes the reference allele bias. We are in the process of further validating these results using a different variant caller, namely MuTect2.

In summary, we show the usefulness of accounting for the germline mutations within a graph framework for the alignment of reads and the subsequent calling of somatic mutations.

# VLINC273 IS REQUIRED FOR NORMAL DNA REPLICATION TIMING OF CHROMOSOME 6 IN THE GM12878 CELL LINE

Michael B Heskett[1], Leslie Smith[2], Takayo Sasaki[3], David M Gilbert[3], Matthew J Thayer[2], Paul T Spellman[1]

[1]Oregon Health and Science University, Medical and Molecular Genetics, Portland, OR, [2]Oregon Health and Science University, Biochemistry, Portland, OR, [3]Florida State University, Biological Science, Tallahassee, FL

Genome instability is observed in virtually all cancer types and is associated with aggressive tumors and poor patient outcomes. A better understanding of the molecular basis of genome instability may lead to better usage of current therapies and development of new therapeutic approaches. By studying specific chromosomal rearrangements that lead to chromosome instability, the Thayer lab recently discovered two cis-acting long non-coding RNAs, ASAR6 (ASynchronously replicating Autosomal RNA) and ASAR15, on chromosome 6 and 15, that are monoallelically expressed and required to maintain normal DNA replication timing and chromosomal stability. To test the hypothesis that all autosomes contain ASARs we performed a computational search for long, L1-rich monoallelic RNA in the GM12878 cell line and found in total 238 putative ASARs across all autosomes. We validated the expression of one striking result, the previously unannotated vlinc273 on chromosome 6, with RNA-DNA FISH and found that CRISPR mediated deletion of the active allele of vlinc273 led to delayed DNA replication timing and delayed mitotic condensation. In summary, we have found a new member of the ASAR family of functional long non-coding RNA and show evidence supporting their necessary role in maintenance of chromosome stability.

# THE FOUNDATIONAL DATA INITIATIVE FOR PARKINSON DISEASE (FOUNDIN-PD)

Peter Heutink[1], David W Graig[2], Kendall Van Keuren-Jensen[3], Steve Finkbeiner[4], Mark Cookson[5], Mike Nalls[6], Raphael J Gibbs[5], Andrew Singleton[5]

[1]German Center for Neurodegenerative Diseases, Genome Biology of Neurodegenerative Diseases, Tuebingen, Germany, [2]USC, Institute of translational Genomics, Los Angeles, CA, [3]TGEN, Neurogenomics division, Phoenix, AZ, [4]UCSF, Center for Systems and Therapeutics & Taube/Koret Center for Neurodegenerative Disease, Gladstone Institute, San Francisco, CA, [5]NIA IRP NIH, Laboratory of Neurogenetics, Bethesda, MD, [6]Data tecnica International, Glen Echo, MD

**Objectives:** FOUNDIN-PD (https://www.foundinpd.org/wp/) will differentiate 100 iPS cell lines from the Parkinson Progression Marker Initiative (PPMI) to dopaminergic (DA) neurons. With these DA neurons we will complete molecular assays to assess transcriptional activity, epigenetic state, genome organization and cellular phenotypes at three different time points (differentiation day 0, 25, and 65). Datasets will be integrated into networks across molecular assays to compare results across cell lines by genetic risk score (GRS) to analyze how Parkinson's disease (PD) associated genetic variants influence gene regulatory networks in DA neurons.

**Methods:** We will complete DA neuron differentiation on a fully automated cell culture platform to generate DA neurons for all PPMI lines. At each time point we will use RNA-seq to examine expression profiles and Illumina EPIC arrays to assess DNA methylation. Transcriptional regulatory elements and chromatin organization will be described using ATAC-seq and Hi-C. We will also use proteomics in a subset of cell lines to quantitatively asses protein expression and longitudinal single cell imaging, Calcium imaging and electrophysiology (MEA) to examine cellular phenotypes.

**Results:** At differentiation day 65, greater than 60% of cells express markers of DA neurons and display electrical activity. Pilot data from 10 PPMI lines is now available to identify differentially regulated genes, eQTL, methQTL expression networks and enriched pathways in DA neurons.

**Conclusions:** Here we present the FOUNDIN-PD project to assess molecular signatures of DA neurons from 100 PPMI iPS cell lines carrying different PD risk variants. Assay data from all time points in the first 10 cell lines have been analyzed and made publicly available. Data from subsequent batches will be made available during the course of 2019.

# FUNCTIONAL ANNOTATION OF NON-CODING GENOMIC REGIONS IN NON-HUMAN PRIMATES

Naoki Hirose[1], Masatsugu Ema[2], Erika Sasaki[3], Jun Kawai[4], Yasuhiro Murakawa[1], Hideya Kawaji[1,4]

[1]RIKEN, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [2]Shiga University of Medical Science, Research Center for Animal Life Science, Otsu, Japan, [3]Central Institute for Experimental Animals, Marmoset Research Department, Kawasaki, Japan, [4]RIKEN, RIKEN Preventive Medicine and Diagnosis Innovation Program, Yokohama, Japan

Understanding non protein coding regions in our genomes attracts wide attentions from aspects of gene regulation in physiological and disease states. In previous studies in FANTOM5, we produced atlases of non-coding elements in human genome, consisting of promoters, enhancers, and long non-coding RNAs (Forrest et al 2014; Andersson et al 2014; Hon et al 2017), where cis-regulatory elements were identified by transcription initiation profiles in nearly 2,000 human samples including ~180 tissues with CAGE (Cap Analysis of Gene Expression) and novel transcript structures were predicted in combination of CAGE and RNA-seq. While a series of functional genomics data in humans and mice has been produced by such as FANTOM5 and ENCODE, evolutional insights into non-coding regions are still limited because (i) conservation of genome sequences in functional but non-coding regions is not necessarily high between humans and mice, and because (ii) functional genomics data in non-human primates (NHPs) is insufficient.

To overcome the limitation, we aim to produce a functional genomics atlas in NHPs, in particular, crab-eating macaques (*Macaca fascicularis*) and common marmosets (*Callithrix jacchus*). We obtained transcription initiation profiles from 270 and 90 tissue samples respectively, so that majority of them matched the human tissues investigated in FANTOM5, leading to more than 100,000 CAGE peaks identified as promoters. Notably, 95% and 90% of the identified promoters in macaques and marmosets are alignable to the human genome, while 70% of the mouse promoters are. The data set also enabled us to identify more than 10,000 transcribed enhancers, and we found that 68% and 48% of the identified enhancers are alignable to the human genome respectively, while 30% of the mouse enhancers are. We additionally performed RNA-seq in a subset of the samples mentioned above, and produced 5'-end accurate transcript models by incorporating RNA-seq data produced by NHPRTR (the Nonhuman Primate Reference Transcriptome Resource; Pipes et al, 2013). It resulted in ~163,000 transcript models, including more than 13,000 loci matched to human non-coding RNA loci, while only ~3,000 mouse long non-coding RNAs are reported as human orthologs (Breschi et al 2017). Although we are still expanding the data set, our resource provides a crucial baseline of a functional annotation to study evolutional aspects of non-coding regions in the human genome.

# GENOME-WIDE GERMLINE CORRELATES OF THE EPIGENETIC LANDSCAPE OF PROSTATE CANCER

Kathleen E Houlahan[1,2,3], Yu-Jia Shiah[1], Musaddeque Ahmed[4], Michael Fraser[1], Housheng H He[2,4], Robert G Bristow[2,4,5], Paul C Boutros[1,2,6,7]

[1]Ontario Institute for Cancer Research, Adaptive Oncology, Toronto, Canada, [2]University of Toronto, Department of Medical Biophysics, Toronto, Canada, [3]Vector Institute, Toronto, Canada, [4]University Health Network, Princess Margaret Cancer Centre, Toronto, Canada, [5]University of Toronto, Department of Radiation Oncology, Toronto, Canada, [6]University of California, Department of Human Genetics, Los Angeles, CA, [7]University of California, Jonsson Comprehensive Cancer Center, Los Angeles, CA

Although germline polymorphisms can modulate cancer risk, little is known about how they influence the somatic epigenomic profiles that emerge in primary tumours. To fill this gap, we present the first systematic survey of the influence of germline single nucleotide polymorphisms (SNPs) on the somatic methylome of localized prostate tumours. We first show that known risk SNPs influence the prostate epigenetic landscape, uncovering a mechanism for cancer susceptibility. We then identify and validate 1,178 tumour methylation quantitative trait loci (meQTLs) with regulatory impacts propagating through the central dogma to modulate RNA and protein abundances. Amongst these are a haplotype on chromosome 14 associated with tumour-specific methylation and expression of the oncogene AKT1 and a haplotype on chromosome 10q associated with tumour-specific epigenetic regulation of TCERG1L – a transcription elongation factor predictive of rapid biochemical relapse following treatment. Taken together, these data reveal a strong interplay between the germline mutational profile and the epigenomic features of primary tumours.

# MULTI-OMICS DATA ANALYSIS IN CARDIOVASCULAR RESEARCH

Julie Hussin

Université de Montréal, Faculty of Medicine, Montreal, Canada

Building predictive models of complex human disease that are useful in a medical setting has proved extremely challenging. This is particularly true for cardiovascular diseases (CVD), a group of disorders impacting the heart and the circulatory system. Annually, more people die from CVD than from any other disease worldwide, with major acute life-threatening events including heart attack and stroke. Rates and clinical manifestations of subtypes of CVD vary across human populations as well as between sexes, with women being often misdiagnosed and undertreated. Extensive research has focused on the discovery by genome-wide association studies (GWAS) of thousands of inherited disease-associated SNPs, which provided overwhelming evidence for the role of genetics in CVD predisposition. On the other hand, genetic data do not show the full picture, as these disorders are of multi-factorial nature. They occur through a complex web of interactions between molecular phenotypes (transcripts, metabolites, micro-organisms) and environmental factors. This quest for full molecular profiling has now led to an explosion of 'omics' technologies producing datasets at all biological levels, ultimately modulated by individuals' genetic makeup. Therefore, our research activities are focused on combining genomics with multi-omics data to understand cardiovascular health, with applications for unraveling molecular processes behind myocardial infarction, pulmonary hypertension and heart failure, along with co-occurring conditions such as diabetes, auto-immune disorders and reproductive challenges. To reach these goals, my research group collaborates closely with clinicians and fundamental researchers at the Montreal Heart Institute to generate and analyze omics data according to the most pressing research questions in cardiovascular research. Here, I will present the research questions currently tackled by my group, all of which are relevant to precision cardiovascular medicine, along with the data available to us to advance this field. I will also describe the different computational challenges and the novel techniques employed to answer these questions, which includes artificial intelligence methods developed in collaboration with the Montreal Institute for Learning Algorithms. These projects hold the potential to improve cardiovascular disease risk-estimating algorithms and predictive models of human health.

# GENOMICS2020: AN OVERVIEW OF THE ONGOING NHGRI STRATEGIC PLANNING PROCESS

Carolyn M Hutter, on behalf of the NHGRI Strategic Planning Group

National Human Genome Research Institute, Division of Genome Sciences, Bethesda, MD

The National Human Genome Research Institute (NHGRI) has a rich tradition of strategic planning that dates back to the Human Genome Project. The significant advances in genomics since the publication of NHGRI's last strategic plan in 2011 illustrate the need for an updated vision about the future of genomics for 2020 and beyond. In February 2018, NHGRI launched a new round of strategic planning, one that will yield publication of a new strategic plan in Fall 2020. The new plan will detail the most compelling areas of genomics research and their application to human health and disease; it will also inform NHGRI's scientific priorities and shape its research portfolio, ensuring that the Institute continues to be a driving force at the Forefront of Genomics. To formulate relevant and forward-reaching goals, NHGRI is fostering partnerships with research, healthcare, education, and policy communities, as well as the public. Starting in 2018, NHGRI has engaged a wide variety of stakeholders through workshops, town halls, advisory meetings, meeting sessions, and webinars. Over the next year, NHGRI plans to use the received input to develop draft tenets for the emerging plan as well as collect additional input through more focused activities. We continue to encourage the genomics community to participate in the process via genome.gov/genomics2020, genomics2020@nih.gov, and #genomics2020. This poster, presented by NHGRI staff, provides an overview of the ongoing NHGRI strategic planning process to date.

# A HUMAN PROTEIN-PROTEIN INTERACTION NETWORK FOR PD GWAS RISK LOCI TO CATALYZE GENOMIC INTERPRETATION

Anastasia Illarionova, Anamika Giri, Joachim Taeger, Peter Heutink

Deutsches Zentrum fur Neurodegenerative Erkrankungen (DZNE), Genome Biology of Neurodegenerative Diseases, Tuebingen, Germany

Genome wide association studies (GWAS), leverage linkage disequilibrium to successfully uncover disease associated common variants and shed light on the underlying pathological processes. However, GWAS itself lacks the explanatory power to identify the causal genes and variants which has to be compensated by the development of methods for true causative variant identification. This challenge can be resolved by combining different approaches including the construction of biological networks, such as protein–protein interaction (PPI) networks, and subsequent analysis of network properties (PMID: 27892958).
One of the examples of heterogeneous disorders with familial and sporadic forms which is actively studied in large patient-control cohorts is Parkinson`s disease, a complex neurodegenerative disorder characterized by accumulation of a selective loss of dopaminergic neurons in the substantia nigra pars compacta (SNpc) and, α-synuclein aggregation into Lewy Bodies (LB). In recent years GWAS for PD have uncovered more than 90 independently associated common SNPs, including variants in the vicinity of TMEM175, MCCC1, STX4-BCKDK, SIPA1L2 among others (PMID: 25064009, PMID: 28892059)
This study aimed to apply a multistep approach allowing to prioritize candidate genes under PD GWAS hits (Nalls et al., 2018, unpublished manuscript) based on experimentally supported protein-protein interactions (PPIs) and revealed putative PD-associated loci connections. For this purpose, modules of genes with highly correlated RNA expression were identified via weighted gene correlation network analysis (WGCNA) on GTEx transcriptomic data from five brain regions: substantia nigra (SN), prefrontal cortex Broadman area 9 (BA9), caudate, putamen and nucleus accumbens (NA). Protein-coding genes from the modules with neurodegeneration related GO terms, comprising PD mendelian genes, were used to build the PPI networks. The PD disease modules (DM) were identified around a set of known PD causative proteins with the use of a DIseAse MOdule Detection (DIAMOnD) algorithm (PMID: 25853560). The candidate hub genes under PD GWAS hits located within the identified DM were further prioritized by means of a literature research and sorted into several main groups according to gene functions and localization. In conclusion, this study lead to generation of PPI networks that uncovered a set of putative interactions between PD associated risk loci. This can be further used for hypothesis generation and subsequent validation on the established PD models.

# COMPARISON OF THE GENE EXPRESSION PROFILE IN THE DIFFERENT CULTURED CONDITION BY SINGLE CELL TRANSCRIPTOME ANALYSIS

Sadahiro Iwabuchi, Saki Kurokawa, Yu Kamide, Shinichi Hashimoto

Kanazawa University, Department of Integrative Medicine for Longevity, Kanazawa, Japan

The aim of this study is to know how the gene expression pattern in 1-cell level differs from original *in vitro* cultured cell to the *in vivo* transplanted cultured cell. Here, we will show that fatty acid-binding protein 5 (Fabp5) slightly decreased during the tumor growth not in the *in vitro* cultured but *in vivo* transplanted culture of mouse hepatocellular carcinoma (HCC) cells. After HCC cells were cultured in *in vitro* condition, cells were collected and transplanted subcutaneously into C57/BL6N mouse. After 2 months housing, tumor tissues were dissected and isolated into single cell. At the same time, the *in vitro* cultured HCC cells were also isolated. The device of our Nx1-seq (next generation 1-cell sequencing) was microwell based, and isolated single cell was applied to the microwell, and cell bar-code beads could capture mRNA. After clustering analysis by t-SNE, it revealed that the *in vitro* cultured HCC cells had 3 HCC cell populations and transplanted cells had 7 clusters (4 HCC-like cell population, 2 fibroblast or smooth muscle cells, 1 macrophage). In this transplanted tumor tissue, there were a few infiltrations of immune cells by t-SNE analysis followed with immunohistochemisty. *Fabp5* expressed in all the 3 clusters of the *in vitro* cultured HCC cells; however, it only expressed at macrophages in the *in vivo* transplanted cultured cells. *Fabp5* was correlated to overall survival of Kaplan-Meier estimate in liver cancer referred from *The Cancer Genome Atlas*. Our result indicates that the changes of microenvironment could reduce anti-inflammatory gene Fabp5 to promote the tumor growth. We are ongoing study to find how the expression of Fabp5 is down-regulated in HCC cells and how macrophage affects the phenomena. In summary, Nx1-seq clearly shows the gene expression pattern of a cell is affected by the microenvironment of the tumor.

# UNDERSTANDING GENETIC VARIATION USING TARGETED AND SINGLE-CELL LONG-READ SEQUENCING

Shruti Iyer[1,2], Sara Goodwin[1], W. Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Genome Center, Woodbury, NY, [2]Stony Brook University, Genetics, Stony Brook, NY

Understanding the basis of complex diseases relies heavily on the ability to effectively study the genomic loci associated with these conditions. Structural variations (SV), a hallmark of genomic instability in cancer, include insertions, deletions, duplications, inversions, or translocations that can either activate oncogenes or inactivate tumor suppressor genes. These variants contribute to the complexity of the cancer genome but are often difficult to resolve. While short-read sequencing has aided cancer genomics, it has performed poorly in SV detection, with false positive and false negative rates of 50% or more. Long-read sequencing generates reads that are several (>10) kb long. This technology has helped identify thousands of genomic features that are possibly pertinent to cancer that were previously missed by short-read sequencing. Long reads have the potential to span difficult SV or repetitive regions with a single continuous read, giving a clearer idea of the variation and eliminating ambiguity in position or size. However, the throughput and coverage, along with the high cost for whole genome long-read sequencing limits its wide adoption. Additionally, long-read sequencing methods are can be greatly limited during the DNA preparation step due to compromised fragment lengths and yields through standard extraction methods. Moreover, long-read platforms require large amounts (several μg) of DNA for library preparation, which prevent their adoption for applications where there is limited in DNA. We use different extraction, amplification, fragmentation, and enrichment strategies to efficiently target and resolve complex regions of the genome that are often associated with cancer. Currently we are developing methods that employ the CRISPR/Cas system to target and enrich for genes that are susceptible to large genomic rearrangements using the breast cancer cell line, SK-BR-3. Our overarching goal is to develop tools that will leverage the long-read sequencing technology to increase the resolution with which we can look at large numbers of cancer genomes. The approaches we develop can be translated to studying several complex disorders and could help identify variants that are important for diagnostic or therapeutic applications.

# CNET: A MULTI-OMICS APPROACH TO DETECTING CLINICALLY ASSOCIATED, COMBINATORY GENOMIC SIGNATURES

Peilin Jia[1], Guangsheng Pei[1], Zhongming Zhao[1,2]

[1]The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX, [2]Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN

Motivation: Genome-wide multi-omics profiling of complex diseases provides valuable resources and opportunities to discover associations between various measures of genes and diseases. Currently, a pressing challenge is how to effectively detect functional genes associated with or causing phenotypic outcomes. We developed CNet to identify groups of genomic signatures whose combinatory effect is significantly associated with clinical and phenotypical outcomes.
Results: CNet builds on a generalized sequential feedforward method, augmented by a down-sampling bootstrap strategy to reduce random hitchhiking signatures. It further applies a dynamic trimming procedure to remove relatively less informative signatures at every step. CNet can manage heterogeneous genomic signature profiles simultaneously and select the best signature to represent a specific gene. To deal with various forms of clinical and phenotypical measurements, we introduced four models to deal with continuous, categorical, and censored data. We tested CNet using drug-response data, multidimensional cancer genomics data, and genome-wide association study data for multiple traits. Our results demonstrated that in various scenarios, CNet could effectively identify signatures that are associated with the outcomes. In addition, we applied CNet to identify likely disease-causing chains involving somatic mutations, pathway activities, and patient outcomes. With appropriate setting, CNet can be applied in many biological conditions.

# THE ENCODE TRANSCRIPTOME PIPELINES: REPRODUCIBLE, PORTABLE TOOLS FOR RNA-SEQ ANALYSIS

Otto Jolanki[1], Ulugbek K Baymuradov[1], Idan Gabdank[1], Keenan Graham[1], Ben C Hitz[1], Jin W Lee[1], Sorena Rahmanian[2], Vanessa Sochat[1], J Seth Strattan[1], Dana Wyman[2], Ali Mortazavi[2], J Michael Cherry[1], ENCODE Data Coordination Center[1]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]University of California Irvine, Department of Developmental and Cell Biology, Irvine, CA

The ENCODE Consortium seeks to annotate and understand the functional elements of the human genome. The Consortium makes all primary and analysis data public through the ENCODE Portal and also shares the analysis tools used to produce the results. Core transcriptome assays, such as poly-A and total RNA-seq, micro-RNA-seq, and long-read RNA-seq, have been performed in Consortium labs in more than 2,000 experiments in over 300 different cell types. Together, the uniformly processed results from these assays present a broad and coherent sampling of transcriptional states and provide a valuable resource for scientists who seek to compare transcript identities and quantifications across cell types. While all data processing for ENCODE is standardized and centralized to ensure comparable results across labs and time, until now it has been difficult for some researchers to compare data from their own experiments with ENCODE data because the software ENCODE used was not always available on multiple platforms in a reproducible deployment. To address this difficulty, the ENCODE Data Coordinating Center (ENCODE-DCC) have developed a general pipeline architecture that includes containerization technology with Docker and Singularity, a domain-specific workflow description language, and best-practices software deployment to cloud, HPC, and local platforms. Here we describe the ENCODE RNA-seq, micro-RNA-seq, and long-read RNA-seq analysis pipeline deployments in the ENCODE framework and show how anyone can run these pipelines on their own data, ensuring their results will be comparable to those from ENCODE. The ENCODE-DCC pipelines and architecture are free and open-source. The code and instructions to run the ENCODE transcriptome pipelines can be found at the following github repositories: https://github.com/ENCODE-DCC/rna-seq-pipeline https://github.com/ENCODE-DCC/mirna-seq-pipeline https://github.com/ENCODE-DCC/long-read-rna-pipeline. The pipeline architecture is demonstrated at https://github.com/ENCODE-DCC/demo-pipeline. ENCODE data are distributed through the ENCODE Portal at https://www.encodeproject.org.

# SIXTEEN DIVERSE LABORATORY MOUSE REFERENCE GENOMES DEFINE STRAIN-SPECIFIC HAPLOTYPES AND NOVEL FUNCTIONAL LOCI

Thomas M Keane[1], Mouse Genomes Project Consortium

[1]European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge, UK, United Kingdom

We report full-length draft *de novo* genome assemblies for 16 widely used inbred mouse strains and find extensive strain-specific haplotype variation. We identify and characterize 2,567 regions on the current mouse reference genome exhibiting the greatest sequence diversity. These regions are enriched for genes involved in pathogen defence and immunity and exhibit enrichment of transposable elements and signatures of recent retrotransposition events. Combinations of alleles and genes unique to an individual strain are commonly observed at these loci, reflecting distinct strain phenotypes. We used these genomes to improve the mouse reference genome, resulting in the completion of 10 new gene structures. Also, 62 new coding loci were added to the reference genome annotation. These genomes identified a large, previously unannotated, gene (Efcab3-like) encoding 5,874 amino acids. Mutant Efcab3-like mice display anomalies in multiple brain regions, suggesting a possible role for this gene in the regulation of brain development.

Further details: Lilue *et al.* (2018) Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci, *Nature Genetics*, 50, pages1574-1583

# IDENTIFICATION OF DISTINCT MOLECULAR SUBTYPES OF CROHN'S DISEASE THROUGH GENE EXPRESSION AND CHROMATIN ACCESSIBILITY

Benjamin P Keith[1,2], Takahiko Toyonaga[2], Michelle Hoffner O'Connor[2], Praveen Sethupathy[3], Shehzad Z Sheikh[1,2], Terrence S Furey[1,2]

[1]The University of North Carolina at Chapel Hill, Department of Genetics, Chapel Hill, NC, [2]The University of North Carolina at Chapel Hill, Center for Gastrointestinal Biology and Disease, Chapel Hill, NC, [3]Cornell University, Department of Biomedical Sciences, Ithaca, NY

Crohn's disease (CD) is a chronic gastrointestinal inflammatory. Current clinical diagnostic methods are ineffective towards predicting disease progression and response to therapy. Given the success of molecular subtyping in various cancers, there is substantive interest in implementing this methodology to predict disease progression and treatment response in CD. Through analysis of gene expression and chromatin accessibility in colon tissue from a small cohort of adult CD patients and non-CD controls, we identified two molecular subtypes associated with distinct clinical behaviors. We are further validating these CD subtypes using gene expression and chromatin accessibility profiles in adult and pediatric patients and aim to also characterize cellular mechanisms underlying these subtypes to identify novel therapeutic targets.

We collected samples from the non-inflamed regions of colon tissue from 110 CD and 51 non-CD adult patients along with samples from non-inflamed sections of the colon and ileum from 76 CD and 51 non-CD pediatric patients. RNA-seq, small RNA-seq, ATAC-seq, and patient genotyping was performed on adult samples and small RNA-seq was performed on pediatric samples. Deconvolution of colonic tissue expression was performed to estimate specific colonic cell proportions, such as intestinal epithelial cells and colonic immune cells, and used to remove variation in tissue composition. Principal component analysis and hierarchical clustering were performed to jointly group samples with similar gene expression and open chromatin profiles. Differentially expressed genes and regions of differential chromatin accessibility were identified. Finally, annotations from the ENCODE and Roadmap Epigenomics Projects were used for functional annotation of open chromatin data, and motif enrichment analysis was performed.

Initial results confirm that at least two distinct molecular CD subtypes are present and that additional subgroups may be present through further refinement of the data. One of the subtypes is associated with alterations in intestinal epithelial cell differentiation that may be contributing to defects in barrier integrity. MicroRNA-31 was identified as a driver of CD subtypes in adults and was associated with the development of fibrotic CD in our pediatric population. We will further present on our joint analysis of gene expression and chromatin accessibility, which we expect will provide insight into the perturbation of cellular mechanisms that underlie CD subtypes and associated clinical phenotypes.

Our results provide the most comprehensive characterization of the regulatory impact of chromatin accessibility on gene expression in CD to date. Greater understanding of the molecular mechanisms that are perturbed in heterogeneous CD subtypes will be vital to provide avenues for novel treatment strategies.

# SINGLE-CELL DISSECTION OF ALZHEIMER'S PATHOLOGY: GENETICS, EPIGENOMICS, TRANSCRIPTION, SEX, AGE, AND DISEASE PHENOTYPES

Manolis Kellis[1,2], Hansruedi Mathys[3,4], Jose Davila-Velderrain[1,2], Liang He[1,2], Shahin Mohammadi[1,2], Carles Boix[1,2], Maria Kousi[1,2], David A Bennett[5], Li-Huei Tsai[3,4]

[1]MIT, Computer Science and Artificial Intelligence Laboryatory, Cambridge, MA, [2]Broad Institute, of MIT and Harvard, Cambridge, MA, [3]MIT, Picower Institute, Cambridge, MA, [4]MIT, Biology Department, Cambridge, MA, [5]Rush University, Alzheimer's Disease Center, Chicago, IL

Alzheimer's disease (AD) is a pervasive neurodegenerative disorder characterized by cognitive de-cline and neuronal loss, whose complex etiology involves multiple brain cell types, but remains poorly characterized. To address this challenge, we analyze 80,660 single-nucleus transcriptomes from prefrontal cortex of 48 AD vs. non-AD individuals, and integrate these data with providing the first single-cell cortical view of AD pathology. We identified transcriptionally-distinct cell subpopulations capturing six major brain cell-types, and analyzed their association with AD pathology. We uncovered coherent neuronal and glial subpopulations associated with AD, and identified regulators of myelination, inflammation, and neuron survival as top markers characterizing them. We find that the strongest AD-associated changes appear early in disease progression, and are highly cell-type specific. By contrast, genes up-regulated in late-stage disease progression are common across cell types, and primarily involved in global stress response. Surprisingly, we found that cells isolated from female individuals are overrepresented in the AD-associated cell subpopulations, revealing substantially different transcriptional responses between sexes. We also combined single-cell profiles, tissue-level variation, and genetic variation across healthy and diseased individuals to deconvolve bulk profiles into single-cell profiles, to recognize changes in cell type proportion associated with disease and aging, and to partition genetic effects into the individual cell types where they act. These revealed a change in cell type proportion associated with aging and with AD, with decreased excitatory and inhibitory neurons, and increased astrocytes, which can become cytotoxic. They also revealed genetic variants underlying cell-type-proportion, including in the TMEM106B locus, which has been previously implicated in Fronto-temporal lobal degeneration (FTLD), and was associated with decreased fraction of inhibitory neurons, but not associated with AD directly. These results provide a roadmap for translating genetic findings into mechanistic insights and ultimately new therapeutic avenues for complex disease and cancer.

# GENE-ENVIRONMENT INTERACTIONS IN THE UK BIOBANK

Matthew Kerin[1,2], Kevin Sharp[1], David Steinsaltz[1], Jonathan Marchini[2]

[1]University of Oxford, Department of Statistics, Oxford, United Kingdom, [2]University of Oxford, Wellcome Trust Center for Human Genetics, Oxford, United Kingdom

The size and scale of the UK Biobank dataset presents a unique opportunity to investigate the contribution of gene-environment interactions to the etiology of many traits and diseases. We aimed to go beyond the simple paradigm of analyzing environmental variables one at a time. To do this we developed a method that estimates the gene-environment heritability explained by a linear combination of a large number of environmental variables. In addition, the method tests each SNP for non-zero interaction effects with the learned linear combination of environmental variables.

We use a Bayesian whole genome regression approach and Variational Bayes for model fitting with a mixture of exact and gradient-based updates to optimize over the hyper-parameters. The method is implemented in C++ using multi-threading to increase computational efficiency. The computational complexity of the method is linear in sample size and number of SNPs.

On simulated data with 25,000 individuals at 100,000 SNP loci, we performed simulations with 30 environmental variables, of which 6 have a non-zero weight, and true PVE for main effects and SNP-environment interactions of 15% and 3%. When compared to a standard linear model applied to each SNP independently, our method has 2.3 times more power to detect true SNPs that interact with environment at a false positive rate of 5%.

We applied the method to a GxE association study of Body Mass Index in 281,184 white British UK Biobank participants, using 642,254 genotyped SNPs and 41 environmental variables. We learn a linear combination of environments that put high weight on alcohol intake frequency, Townsend Index and physical activity measures and identify 17 SNPs that interact with the linear combination with $p<5 \times 10^{-8}$. Amongst others, we find GxE signals at SNPs proximate to *SEC16B*, *FOXO3* and *DIO2*.

# TELOMERE-TO-TELOMERE ASSEMBLY OF A COMPLETE HUMAN X CHROMOSOME

Karen H Miga*[1], Sergey Koren*[2], Arang Rhie[2], Shelise Brooks[2], Gerard G Bouffard[2], Valerie Maduro[2], Nancy Hansen[2], Edmund Howe[3], Jennifer L Gerton[4], Jo Wood[3], Milinn G Kremitzki[5], Matthew Loose[6], Urvashi Surti[7], Kerstin Howe[4], Tina G Lindsay[5], Alice Young[2], James Mullikin[2], Adam M Phillippy[2]
* authors contributed equally

[1]University of California, UC Santa Cruz Genomics Institute, Santa Cruz, CA, [2]National Institutes of Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [3]Stowers Institute for Medical Research, Kansas City, MO, [4]Wellcome Sanger Institute, Cambridge, United Kingdom, [5]Washington University School of Medicine, Department of Genetics, St Louis, MO, [6]University of Nottingham, School of Life Sciences, Nottingham, United Kingdom, [7]University of Pittsburgh, Department of Pathology, Pittsburgh, PA

Release of the first human genome assembly was a landmark achievement, and after nearly two decades of improvements, the current human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever produced. However, no chromosome has yet been finished end to end, and hundreds of gaps persist across the genome. These unresolved regions include segmental duplications, ribosomal rRNA gene arrays, and satellite arrays that harbor unexplored variation of unknown consequence.

Here we announce a whole-genome de novo assembly that surpasses the continuity of GRCh38, along with the first complete, telomere-to-telomere assembly of a human X chromosome. In total, we collected 45X coverage of ultra-long Oxford Nanopore sequencing for the CHM13hTERT cell line, including 53 Gb of sequence in reads >100 kb. A de novo assembly combining this nanopore data with 70X of existing PacBio data achieved an NG50 contig size of 75 Mb (versus to 56 Mb for GRCh38), with some chromosomes broken only at the centromere. The assembly had the highest de novo segmental duplication resolution and the highest concordance with previously finished CHM13hTERT BACs. It accurately captured gene copy number of large repeat arrays, such as the CT45 and GAGE loci on chrX, as well as intra-repeat structural variation.

Using this assembly as a basis, we chose to manually finish the X chromosome. The few unresolved segmental duplications were assembled using reads spanning the individual copies, and the ~2.7 Mbp X centromere was assembled by identifying unique variants and using these to anchor overlapping ultra-long reads. Orthogonal technologies, such as BioNano, 10X, ddPCR and PGFE, were used to validate the assembly. All raw data and assemblies are available on AWS as an open dataset. Our future work will focus on completing and validating the remainder of the genome.

# µSPLiT: HIGH-THROUGHPUT SINGLE-CELL RNA SEQUENCING PLATFORM FOR MICROBIAL COMMUNITIES

Anna Kuchina*[1], Leandra M Brettner*[2,3], William R DePaolo[3,4], Georg Seelig[1,5]

[1]University of Washington, Electrical and Computer Engineering, Seattle, WA, [2]University of Washington, Bioengineering, Seattle, WA, [3]University of Washington Medical Center, Center for Microbiome Sciences & Therapeutics, Seattle, WA, [4]University of Washington, Gastroenterology, Seattle, WA, [5]University of Washington, Computer Science and Engineering, Seattle, WA

*authors contributed equally

Transcriptomic analysis of individual cells has become an essential tool for characterizing cellular diversity. Single-cell RNA sequencing (scRNA-seq) methods allow profiling of up to hundreds of thousands of cells to create comprehensive atlases of cell types in a wide range of organisms. However, so far, these methods have not been compatible with characterizing gene expression in microbes. While the gene expression profile of the complex microbial communities such as the gut microbiome as a whole may be obtained through metatranscriptomics, we cannot with certainty tell which strains or species are responsible for turning on specific gene expression pathways. This lack of single-cell resolution hinders our understanding of the functional consequences of changes in microbiota composition or its interaction with the host environment.

**I present µSPLiT, a high-throughput scRNA-seq approach for microbial communities.** µSPLiT is based on SPLiT-seq, an affordable, high-throughput approach for characterizing gene expression at the single-cell level in mammalian cells. SPLiT-seq employs a combinatorial split-pool indexing strategy, where in each round groups of cells are randomly divided into 96 containers or 'wells' and have their cDNA labeled with a well-specific oligonucleotide sequence tag, then pooled back together. Even though many cells might get the same tag in each round, repeating the split-pool rounds several times ensures that each cell travels through a unique combination of wells with high likelihood. µSPLiT scales exponentially with the number of barcoding rounds, enabling a massive increase in the number of cells that can be sequenced, which makes it uniquely suitable for profiling a vast number - up to millions - of microbial cells. µSPLiT does not require isolation of individual microbes in droplets or wells which can be challenging due to the variable sizes and shapes of microbes.

We develop and validate µSPLiT with synthetic microbial communities including both gram-negative and -positive bacteria, with the outlook of applying it to a number of *in vivo* model systems. Ultimately, microbial scRNA-seq will be applied to resolve heterogeneous functional cell states and cell-to-cell variation in the gut microbiome.

# THE PRIMATE GENOME SEQUENCING INITIATIVE: ASSESSING PRIMATE GENOMIC DIVERSITY AT A GLOBAL SCALE

Lukas F.K. Kuderna[1], Marta Gut[2], Ivo Gut[2], Felipe Ennes[3], Takashi Hayakawa[4], Amanda D Melin[5], Katerina Guschanski[6], Jean P Boubli[3,7], Christian Roos[8], Jeffrey Rogers[9], Kyle Kai-How Farh[10], Tomas Marques-Bonet[1,2,11,12]

[1]Institut de Biologia Evolutiva, UPF-CSIC, Barcelona, Spain, [2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona, Spain, [3]University of Salford, ., Salford, United Kingdom, [4]Japan Monkey Center, ., Inuyama, Japan, [5]Department of Anthropology and Archaeology, University of Calgary, ., Calgary, Canada, [6]Department of Ecology and Genetics (Animal ecology), Uppsala University, ., Uppsala, Sweden, [7]Instituto Nacional de Pesquisas da Amazônia, ., ., Brazil, [8]German Primate Center, Leibniz Institute for Primate Research, ., Göttingen, Germany, [9]Human Genome Sequencing Center, Baylor College of Medicine, ., Houston, TX, [10]Illumina Artificial Intelligence Laboratory, Illumina Inc, ., San Diego, CA, [11]Institució Catalana de Recerca i Estudis Avançats (ICREA), ., Barcelona, Spain, [12]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, ., ., Spain

Around 60% of the world's primate species are currently threatened with extinction. The majority have declining population sizes, mainly because of anthropogenic factors. While several specific species of primates have been studied in depth mainly to better understand our own evolutionary trajectories, there is still a lack of data to understand primate evolution and genomic diversity at a global scale. We produced whole genome shotgun sequences to an average coverage of at least 30X for several hundred primates, and currently cover over a hundred different species. Our samples span 39 different genera and currently include 36% of all described species of Old-World monkeys and New-World monkeys in Africa, Asia and the Neotropics. More than half of these samples are from species threatened with extinction, and the majority are from wild born individuals. We explore how genetic variation within different primate lineages has been shaped across the globe and find New-World monkeys and Asian Old-World monkeys to exhibit much lower degrees of genetic diversity than African Old-World Monkeys. We furthermore study how the IUCN extinction risk categories correlate with estimates of genetic diversity in several groups. Our dataset provides the most extensive and complete catalogue of primate genetic variation to date, with which we aim to help guide future conservation efforts.

# PREDICTIVE MODELING IN BIOLOGICAL SYSTEMS USING MULTI-OMICS DATASETS IN OPEN-SCIENCE, COMMUNITY PLATFORM OF KBASE.

Vivek Kumar[1], Sunita Kumari[1], Zhenyuan Lu[1], Doreen Ware[1,2], Bob Cottingham[3], Chris Henry[4], Adam Arkin[5]

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

The Systems Biology Knowledgebase (KBase) is an open-science, open-source, software and data platform funded by U.S. Department of Energy in support of biological and environmental research to investigate the complex interactions within biological systems. Supported by scalable computing infrastructure, it enables sharing, integration, and analysis of multi-omic datasets related to prokaryotes and eukaryotes, particularly microbes, plants, fungi and their communities.

KBase offers a suite of scientific applications to enable users to build sophisticated analytical workflows and share their findings. The nearly 200 apps in KBase offer diverse scientific functionality across the realms of comparative genomics, communities analysis, metabolic modeling, and transcriptomics. 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.

KBase is also actively engaged with the external community to improve its suite of available tools and workflows and to accelerate scientific discovery, improve reproducibility, and foster open collaboration. KBase is a freely available and developer extensible platform that enables scientists to analyze their own data within the context of public data and share their findings across the system. Developers can build, test, register, and deploy new or existing software as KBase apps using the Software Development Kit, thereby extending the platform's scientific capabilities.

# EFFICIENT AND HIGH-THROUGHPUT GENOTYPING OF ELUSIVE HUMAN POLYMORPHIC INVERSIONS MEDIATED BY INVERTED REPEATS

Marina Laplana[1], Sergi Villatoro[1], Roser Zaurín[1], José Luis Royo[2], Marta Puig[1], Mario Cáceres[1,3]

[1]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra (Barcelona), Spain, [2]School of Medicine, University of Malaga, Department of Surgery, Biochemistry and Immunology, Málaga, Spain, [3]Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

Structural variants are large genomic rearrangements that affect a big number of bases, with the potential to have a wide phenotypic impact. Among them, inversions are genomic regions that change orientation without gain or loss of DNA and many of them have highly identical inverted repeats (IRs) at the breakpoints. Moreover, they tend to be recurrent and not linked to SNPs. Thus, the complexity of such regions has made inversion characterization difficult. Previously, we developed a combination of inverse PCR and MLPA to simultaneously interrogate 24 common human inversions mediated by IRs in hundreds of individuals. Here, we have optimized this method by taking advantage of the iPLEX technology to detect one single nucleotide change specific of each orientation using mass spectrometry and extending the analysis to additional inversions. Also, we have improved genotype accuracy by designing specific assays that take into account common polymorphisms affecting restriction-enzyme target sites. Currently, we are able to genotype 37 inversions amplified in four different PCR reactions from as little as 400 ng of DNA. In order to test the performance of the method, we have genotyped 95 individuals from three populations of the 1000 Genomes Project with 94.3% genotyping rate and an error rate of ~3% for a single experiment. The new genotype data generated allow us to have a first glimpse on the functional and evolutionary consequences of these inversions. In addition, having a high-throughput genotyping technique paves the way to unravel the role of inversions in complex traits and disease susceptibilities.

# DETECTION OF TISSUE-SPECIFIC RNA IN THE PLASMA OF CANCER PATIENTS: A CIRCULATING CELL-FREE GENOME ATLAS (CCGA) SUBSTUDY

Matthew H Larson[1], Hyunsung J Kim[1], Wenying Pan[1], Sarah Stuart[1], Yiqi Zhou[1], Archana Shenoy[1], Monica Pimentel[1], Per Knudsgaard[2], Vasiliki Demas[3], Earl Hubbell[1], Alexander M Aravanis[3], Arash Jamshidi[1]

[1]GRAIL, Inc., Bioinformatics, Menlo Park, CA, [2]GRAIL, Inc., Software Engineering, Menlo Park, CA, [3]GRAIL, Inc., Research and Development, Menlo Park, CA

**Background**: Cell-free RNA (cfRNA) is a promising analyte for cancer detection, but a comprehensive assessment of cfRNA in persons with and without cancer is lacking. To characterize tumor-derived RNA in plasma, we performed an exploratory CCGA substudy to examine cfRNA expression in participants with and without cancer.

**Methods**: Whole-transcriptome sequencing of cfRNA was performed on 210 participants from the previously-described CCGA training set, of which 98 were diagnosed with stage III cancer at the time of blood draw (stage III selected to maximize signal in the blood and avoid confounding signal from secondary metastases). The cancer types were breast (47), lung (32), colorectal (15), and anorectal (4); 112 participants were frequency-age-matched individuals without cancer. When available, RNA from buffy coat and formalin-fixed paraffin-embedded (FFPE) tumor tissue was analyzed.

**Results**: Of 57,783 annotated transcripts, 39,564 (68.5%) were not detected in the cfRNA of non-cancer individuals, and were defined as “dark channels.” A subset of dark channel transcripts was recurrently detected in individuals with cancer and was defined as “dark channel biomarkers” (DCBs). Twelve DCBs were identified in lung cancer and nine DCBs were identified in breast cancer. No DCBs were identified in anorectal or colorectal cancer. A large percentage of DCBs (50% of lung DCBs, and 44% of breast DCBs) are tissue-specific, and a subset of DCBs were specific to lung and breast cancer subtypes. Detection of DCB transcripts was also correlated with tumor fraction in plasma and expression in matched tissue.

**Conclusions**: cfRNA DCBs may provide a unique opportunity to detect cancer and identify tissue of origin and cancer subtype. DCBs have low background in plasma from non-cancer individuals and are specific to tumor tissue and plasma from cancer participants. In addition, DCB signal amplification due to high tumor tissue expression may enable detection of cancer in patients with low tumor fraction in cfDNA. Taken together, these data suggest that tissue-specific transcripts have potential for use in multi-cancer detection.

Funding: GRAIL, Inc.

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

Ryan M Layer[1], Brent S Pedersen[2], Aaron R Quinlan[2], Fritz J Sedlazeck[3], Eric Boerwinkle[4]

[1]University of Colorado, BioFrontiers Institute, Boulder, CO, [2]University of Utah, Human Genetics, Salt Lake City, UT, [3]Baylor College of Medicine, Genome Sequencing Center, Houston, TX, [4]University of Texas, Health Science Center, Houston, TX

SNV population frequency from resources such as gnomAD 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 that makes it challenging to use SV call sets to determine the prevalence of variants observed in new patients since it is impossible to tell if a variant is absent rare or if it was removed from the call set by a filter.

To fix this, we built STIX to index thousands of genomes (BAMs/CRAMs) to enable rapid searches for any evidence that supports the existence of an SV. From this search, we can better differentiate between common and rare SVs (deep evidence in many samples vs. no evidence). To test STIX, we indexed 2,535 genomes from the 1,000 Genomes Project (1KG) and 253 genomes from the Simons Foundation Genome Diversity Project (SGDP), 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 (Sudmant et al. 2015), STIX identified evidence for 16.2% of the COSMIC SVs (~1 sec per SV) across the two cohorts (7.3% in 1KG and 8.8% in SGDP), implying that many of these SVs do not play a role in cancer.

STIX can also help prioritize SVs found in disease studies. For example, among the 1,640 deletions (>500bp) found in the GiAB Ashkenazi Jewish trio, STIX found evidence for 88.3% of the SVs in 1KG (92% in SGDP) while only 48.7% appeared in the 1KG SV call set. Importantly, of the 8 purported de novo SVs in the child, STIX found evidence for 5 of these in 1KG while the filtered call set only contained 2.

We are actively improving the power of STIX by greatly expanding the number of genomes that we index. In collaboration with the University of Texas Health Science Center and Baylor College of Medicine Human Genome Sequencing Center, we will soon make over 10,000 human genomes available for STIX SV population frequency queries. It is our hope this resource will increase the rate at which SVs are considered in disease gene discovery and diagnosis.

# CONTIGUOUS ASSEMBLY OF BIONANO DLS CHEMISTRY REVEALS STRUCTURAL VARIANTS AT CHALLENGING REGIONS

Joyce Lee, Andy Wing Chun Pang, Thomas Anantharaman, Ernest T Lam, Alex Hastie, Mark Borodkin

Bionano Genomics, Research Development, San Diego, CA

The human reference, hg38, released in 2013 is by far the best annotated human assembly and has been used in many clinical studies. The reference is about 3.2 Gbp; it is a haploid representation of the human genome with about 200 Mbp of gaps. Since a complete and accurate anchor is very important for clinical studies, closing of these gaps and accurate representation of challenging regions, such as tandem repeats and segmental duplication regions, are important. With the advancement of technologies, recent studies have also made an effort in identifying structural variants that are due to the incompleteness or misrepresentation of the hg38 reference.

One of the advanced technologies is Bionano Genomics' Saphyr™ system and its direct labelling and staining (DLS) chemistry; physically intact molecules of this technology have N50 length of about 250 kbp and maximum length of about 5 Mbp. Challenging regions of the human genome, including regions that are incomplete or repeats that are collapsed in the reference, become accessible and easy to assemble with high accuracy using these ultra-long molecules. Moreover, with these contiguous molecules, assemblies with N50 of about 60 Mbp are assembled with the well understood overlap layout consensus algorithms. Such contiguous assembly provides the basis for high sensitivity in structural variation detection, which is very important in clinical studies.

One of the latest efforts to close gaps and resolve challenging regions in the hg38 reference used a combination of multiple platforms and analyzed structural variants called from fifteen genomes (*Audano et.al*). Using a 2.99 Gbp CHM1 assembly with N50 of 60 Mbp constructed from Bionano DLS molecules, about 78% of the 811 deletions and 1989 insertions overlapped the structural variants found in Audano's study; with about 37% and 46% of the Bionano unique insertions and deletions involved in segmental duplication regions respectively, and about 2.5% and 28% of the Bionano unique insertions and deletions involved in gap regions respectively. Most of the Bionano calls >50 kbp do not have concordant overlaps despite strong intact DLS molecule support; this might be caused by the uniqueness in detecting large structural variants using ultra-long molecules as well as a resolution difference between technologies.

In conclusion, Bionano DLS chemistry provides contiguous assemblies with high accuracy; given such assembly, structural variants in challenging regions are revealed and well supported by intact ultra-long molecules. This helps in resolving challenging regions and covering gaps in the current hg38 reference.

# THE INFLUENCE OF HUMAN INVERSIONS ON GENE EXPRESSION, EPIGENETIC CHANGES AND PHENOTYPIC VARIATION

Jon Lerga-Jaso[1], Marta Puig[1], Alejandra Delprat[1], Sergi Villatoro[1], Marina Laplana[1], Alba Vilella-Figuerola[1], Teresa Soos[1], Claudia Ramírez[1], Clara Vizuete[1], Olivier Delaneau[2], Tõnu Esko[3], Emmanouil T Dermitzakis[2], Mario Cáceres[1,4]

[1]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra (Barcelona), Spain, [2]University of Geneva, Department of Genetic Medicine and Development, Geneva, Switzerland, [3]University of Tartu, Estonian Genome Center, Tartu, Estonia, [4]Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

Despite structural variants contributing significantly to genetic diversity, their association to phenotypic variation is still poorly characterized compared to SNPs. This is even more obvious for inversions, many of which appear in complex repetitive regions and constitute a technical challenge. Here we analyzed the potential consequences of more than 100 inversions accurately genotyped in individuals from several human populations with diverse approaches, which represents the most complete resource for this type of variants so far. First, we measured their effect on gene expression using transcriptome data from different datasets of blood-derived cells. We detected that inversions maintain divergent haplotypes associated to gene-expression differences, but they can also act directly as eQTLs by disrupting gene structures or creating novel fusion transcripts. Moreover, we looked for additional inversion effects through linked SNPs already reported as eQTLs by the GTEx or alternative eQTL collections. This analysis confirmed our previous results and identified more candidates that may affect gene regulation in other tissues or specific conditions, such as infection. Second, for those inversions in non-coding regions acting as eQTLs, we checked epigenetic patterns, finding that inversion orientation can modulate chromatin accessibility, DNA methylation and the presence of certain histone marks. Finally, we investigated the role of inversions in human phenotypes directly and by linkage disequilibrium with other variants, finding associations with neurological disorders, atrial fibrillation or immune response. In addition, inversion regions are enriched in GWAS signals, supporting their possible implication in disease. Interesting examples include HsInv0031, which is associated to epigenetic changes and *FAM92B* expression in cerebellum, and is linked to a SNP conferring risk for Alzheimer's disease. Importantly, recurrent inversions are poorly imputed with the most used bioinformatic methods, meaning that their effects have been lost in GWAS analysis. Thus, these results highlight the functional role of inversions and reveal previously missing variants responsible for human traits and diseases.

# EVOLUTIONARY DIVERGENCE OF MAMMALIAN SPERMATOGENESIS: SINGLE-CELL COMPARATIVE ANALYSES ACROSS MOUSE, MONKEY, AND HUMAN TESTES

Xianing Zheng#, Adrienne Shami#, Qianyi Ma, Hailey Larose, Gabriel Manske, Sue S Hammoud*, Jun Z Li*

University of Michigan, Department of Human Genetics, Ann Arbor, MI

Spermatogenesis is a highly choreographed developmental process by which diploid spermatogonial stem cells terminally differentiate to produce mature, haploid spermatozoa within the testis. This process is continuous throughout adult life in mammals. In humans, male-derived factors account for up to 50% of infertile couples; however, the etiology of male infertility remains unclear. For decades, our knowledge of the functional heterogeneity of cells involved in spermatogenesis was based on histological/morphological staging, and focused analyses of known cells and their cell surface proteins. Recently we performed single-cell RNA sequencing of ~35,000 cells from the adult mouse testis and identified not only all known germ and somatic cells, but also two unexpected somatic cell types not known to be present in the testis (PM30146481). The data revealed a continuous developmental trajectory from spermatogonia to spermatids, producing a new atlas of 29,522 germ cells for understanding the biology of meiosis. Focused analyses delineated four subtypes of spermatogonia and nine subtypes of Sertoli cells; the latter linked to histologically defined developmental stages over the seminiferous epithelial cycle. We have now extended the analyses to adult humans (n=4, 13,837 cells) and rhesus macaques (n=5, 22,015 cells), identifying seven somatic cell types in humans, and an overlapping set of seven somatic cell types in monkey. We compared the pseudotime of continuous germ cells trajectory across the three species (10,115 germ cells in humans and 19,678 in monkeys) and observed striking differences in cell state occupancy across species. At the start of the trajectory rodent germ cells are "earlier", more stem-like, than the cells in primate. Toward the end of the trajectory the monkey spermatids extend into further differentiated states than the mouse and human spermatids. The dis-synchrony of germ cell maturation program also leads to variable "clock rate" within the trajectories, providing important clues of evolutionary divergence of mammalian spermatogenesis. Genes known to be implicated in male infertility often show stage-specific dynamic patterns corresponding to, or preceding, the observed stage of developmental arrest for that gene. Unraveling the regulatory mechanisms governing the elastic pseudotime during and after meiosis is expected to inform new strategies for *in vitro* gametogenesis and better treatment of infertility.

(Supported by NIH/NICHD R21HD090371, NIH New Innovator Award DP2OD022405 (SSH), Michigan Institute for Data Science, Open Philanthropy Foundation)

# DELVING INTO CHROMATIN ON THE ENCODE PORTAL

Khine Z Lin, Idan Gabdank, Jason Hilton, Yunhai Luo, Jennifer Zamanian, Philip Adenekan, Ulugbek K Baymuradov, Keenan Graham, Jennifer P Jou, Jin W Lee, Casey R Litton, Zachary A Myers, Emma O'Neill, Paul Sud, Forrest Tanaka, J. Seth Strattan, Benjamin C Hitz, J. Michael Cherry

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

The Encyclopedia of DNA Elements (ENCODE) project is an international collaboration of research groups funded by the NHGRI. The main goal of the project is to identify the functional elements of the human genome. The ENCODE portal (https://www.encodeproject.org/), established and maintained by the ENCODE Data Coordination Center (DCC), provides unrestricted access to the data generated by the members of the ENCODE consortium and other related projects such as Model organism ENCODE (modENCODE), Model organism Encyclopedia of Regulatory Networks (modERN), Genomics of Gene Regulation (GGR), and Roadmap Epigenomics Project (Roadmap). The DCC develops and updates the data model and uniform processing pipelines that allow transparency and reproducibility of the experimentally collected and computationally generated data. The categories of metadata and their implementation in a structured, machine-readable data model ensure that the users of the ENCODE portal can search and identify related experiments, understand how assays were performed, and track the provenance of the data. Currently the ENCODE portal hosts data from more than 14 thousand datasets, representing experimental results of more than 40 diverse assay types, performed in multiple cell lines and tissues under various conditions. The portal provides free access to over 640 terabytes of raw experimental, processed, and analysis data files with the corresponding metadata.

Richness, diversity, and structure of the data on the ENCODE portal make it an invaluable resource for the scientific community. The presence of curated and standardized experimental results from ChIP-seq, DNase-seq, ATAC-seq, HiC and ChIA-PET assays, that were uniformly processed, makes the portal into a one-stop shop for a chromatin investigator. Here we show how the ENCODE portal can be used as a multifaceted lens for chromatin exploration.

# DEVELOPING *DROSOPHILA MELANOGASTER* AS A MODEL FOR THE GENETIC BASIS OF EARLY-STAGE REPRODUCTIVE ISOLATION

Matthew J Lollar, John E Pool

University of Wisconsin-Madison, Genetics, Madison, WI

A critical question in evolutionary biology is how distinct populations of the same species progress into reproductively isolated and distinct species. Many researchers have attempted to address these uncertainties by studying largely-isolated Drosophila species. However, there are also intrinsic barriers to gene flow that exist between geographically distant *D. melanogaster* populations, particularly those between ancestral African and cosmopolitan European groups. This finding suggests that *D. melanogaster* may be a particularly good model for studying the genetic basis of early stage reproductive isolation within a species. Here, we highlight quantitative and population genetic approaches to finding genes underlying epistatic incompatibilities between natural populations. First, we provide preliminary results from genomic mapping experiments aimed at uncovering the underlying genetic architecture of F2 lethality, which are consistent with multiple incompatibilities per cross. Second, we extend the population genetic analysis of Pool (2015), which revealed strong evidence of ancestry disequilibrium between unlinked loci in the (admixed) Drosophila Genetics Reference Panel. One explanation of this observed deficiency of African-European allelic combinations may be Bateson-Dobzhansky-Muller incompatibilities, and we thus sought to increase the genomic resolution of these results to reveal new insights into potential epistatic incompatibilities. Single nucleotide gametic disequilibrium between previously identified interactors revealed local peaks of correlation between multiple unlinked neuronal genes, often localizing to specific exonic or intronic gene regions. Many of these loci also displayed high FST between the African and European source populations, consistent with functional evolution following these populations' divergence that may have contributed to the origin of incompatibilities. Taken together, these approaches will yield insights into the genetic basis of reproductive isolation while simultaneously elucidating underlying epistatic interactions at incompatible loci.

# LESS-IS-MORE AND MORE-IS-LESS IN INTEGRATING MULTI-SCALE ‘OMICS WITH POLYGENIC PHENOTYPE PREDICTORS

Shengjie Lu[1], Chen Cao[2], Zhi Xiong[3], Xuewen Lu[1], Quan Long[1,2,4]

[1]University of Calgary, Mathematics and Statistics, Calgary, Canada, [2]University of Calgary, Biochemistry and Molecular Biology, Calgary, Canada, [3]Shantou University, Computer Science, Shantou, China, [4]University of Calgary, Medical Genetics, Calgary, Canada

A goal of genomics is to develop a genome-based phenotype predictor. Empowered by the advancement of high-throughput technologies, in addition to genomes, researchers can assess transcriptome, methylome, and etc. How to statistically integrate them in a phenotype predictor is not trivial. This is because, due to overfitting, more data may cause an increased model complexity that offsets the benefit of the extra information. We develop a novel approach that leverages ‘omics to reduce, instead of increase, the model complexity. In our method, the role of ‘omics is to provide priori knowledge that helps group genomic variants effectively, leading to fewer coefficients in the predictor. Testing it on benchmark datasets in plants and humans demonstrates that it increases the power of prediction substantially. This research leads to a new path towards integration of multi-scale data with polygenic predictors along the line of the “less-is-more” principle in the circumstance of “more-is-less”.

# SCIAPPS.ORG: A CLOUD-BASED PLATFORM FOR REPRODUCIBLE BIOINFORMATICS WORKFLOWS

Zhenyuan Lu[1], Liya Wang[1], Peter Van Buren[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY

Due to the rapid growth of both sequence and phenotype data generated by high-throughput methods, there have been increasing needs to store and analyze data on distributed storage and computing systems. Meanwhile, a workflow management system is needed to ensure efficient data management across heterogeneous systems, simplify the task of analysis through automation, and make large-scale bioinformatics analysis accessible and reproducible. To address these needs, we have developed SciApps, a cloud-based platform for reproducible bioinformatics workflows. The system is integrated with CyVerse Cyber-infrastructure (CI) through the Agave platform for job management and iRODS-based CyVerse Data Store for data management. To create a workflow, each analysis job is submitted, recorded, and accessed through the web portal. Part or all of a series of recorded jobs can be saved as reproducible, shareable workflows for future execution using the original or modified inputs and parameters. The platform is designed to automate the execution of modular Agave apps and make it easy to bring reproducible workflows to cloud-based computing systems.

# PUTATIVE NON-CODING REGULATORY DRIVER DISCOVERY USING UNMATCHED TUMOR-NORMAL SAMPLES

N M R Lykoskoufis, H Ongen, E T Dermitzakis

University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva, Switzerland

We had previously developed a methodology using allele specific expression (ASE) as a proxy to discovering Genes with Allelic Dysregulation (GADs) impacted by non-coding somatic mutations, which requires matched healthy-tumor samples. Here, we describe a novel methodology to discover GADs by using unmatched samples where we can compare tumor RNAseq ASE with reference RNAseq ASE (e.g. GTEx).

It is expected that somatic mutations under selection for tumorigenesis falling in non-coding regulatory regions of a gene tend to affect transcription in an allelic fashion (somatic mutations are in heterozygosity), which is observed as an altered allelic imbalance between the normal and tumor samples. In addition to that, we also expect this gene to have an altered gene expression in order for selection to occur. Hence, in order to detect non-coding regulatory drivers in tumorigenesis and discriminate genes under selection or under an eQTL effect, we devised three criteria ensuring that the discovered genes are true drivers.

To test our method, we used Colorectal cancer data from the SYSCOL project and treated tumor and healthy samples as unmatched. We discovered 702 genes passing all filtering criteria. Of them, 197 overlapped with the 373 Genes with Allelic Dysregulation (GADs) previously discovered with matched samples analysis. We further tested our methodology in a real setting using SYSCOL tumor and GTEx samples from colon sigmoid and colon transverse as normal and discovered 612 and 689 GADs, respectively. Of these, 480 genes overlapped among all three unmatched analyses passing all criteria of which 152 overlapped with the previously discovered GADs indicating that our methodology is able to detect regulatory drivers. Many of these 480 genes have already been implicated in various cancers or have biological properties relevant to cancer progression. Future studies are being planned for validating the method using Prostate adenocarcinoma and other cancers. This new approach provides a great opportunity to learn more about the understudied non-coding regulatory drivers in a large number of cancers where the cell type or tissue of origin has been profiled in a reference sample.

# SOMATIC MUTATIONAL PROFILING REVEALS THE CELLULAR HISTORY OF NORMAL HUMAN LYMPHOCYTES

Heather Machado[1], Emily Mitchell[1], Nina Øbro[2], Henry Lee-Six[1], Megan Davies[2], Kourosh Saeb-Parsy[3], Dan Hodson[2], Anthony Green[2], David Kent[2], Peter Campbell[1]

[1]Wellcome Trust Sanger Institute, Cancer, Ageing and Somatic Mutation, Hinxton, United Kingdom, [2]University of Cambridge, Wellcome-MRC Cambridge Stem Cell Institute and Department of Haematology, Cambridge, United Kingdom, [3]University Department of Surgery and Cambridge NIHR Biomedical Research Centre, Addenbrooke's Hospital, Cambridge, United Kingdom

Mature blood cells harbor a mixture of mutations inherited from ancestral hematopoietic stem cells (HSCs) and mutations accumulated after maturation. The landscape of these somatic mutations in normal blood is poorly mapped, with questions as simple as "how many mutations does a memory T cell accumulate throughout life?" remaining unanswered. In order to characterize normal somatic mutation in lymphocytes, we performed single-cell expansion and whole genome sequencing of over 600 T and B lymphocytes and 200 HSC and progenitor cells across 5 individuals (ages 0-85). We find a striking variation in mutation burden within and between cell subsets. Micro-environment specific mutational processes dominate the observed differences. Examples of this include germinal center ("non-canonical AID") mutations in memory B cells and UV-like mutations in memory T cells (putatively skin resident cells). Naive B and T cells show a lack of variation in discreet mutational patterns relative to their memory counterparts, and have mutational profiles more similar to that of HSCs. Our work highlights an under-appreciated genetic diversity in normal lymphocytes, with some cells accumulating thousands of mutations on top of those inherited from the HSC compartment. These mutations can be used to describe the life history of each individual lymphocyte including their exposure to specific micro-environments. Our findings shed light on the biology of these cells and will help differentiate between normal and disease processes.

# PROFILING IMMUNOGLOBULIN REPERTOIRES BY RNA SEQUENCING ACROSS 8555 SAMPLES FROM 53 GTEX TISSUES

Serghei Mangul, Noah Zaitlen, Eleazar Eskin

UCLA, Computer Science, Los Angeles, CA

Assay-based approaches provide a detailed view of the adaptive immune system by profiling immunoglobulin (Ig) receptor repertoires. However, these methods carry a high cost and lack the scale of standard RNA sequencing (RNA-Seq). In this study, we developed ImReP, a novel alignment-free computational method for rapid and accurate profiling of the Ig repertoire from regular RNA-Seq data. We applied it to 8,555 samples across 544 individuals from 53 tissues obtained from Genotype-Tissue Expression study (GTEx v6)11. The data was derived from 38 solid organ tissues, 11 brain subregions, whole blood, and three cell lines. ImReP is able to efficiently extract Ig-derived reads from the RNA-Seq data and accurately assemble the complementarity determining regions 3 (CDR3s). A number of methods were designed to assemble Ig repertoires and have been applied across various public RNA-Seq datasets. Existing methods that are capable of assembling Ig repertoires from bulk RNA-Seq data produce low-accuracy results (f-score<.2) relative to ImReP. We investigated which portion of the total immune repertoire RNA-Seq is capable of capturing. Using RNA-Seq ImReP was able to capture on average 53.3% of the IGH repertoire, estimated as the sum of the detected BCR-seq-confirmed clonotypes. ImReP was able to accurately estimate the relative frequencies of assembled clonotypes (r=0.97, p-value=10-40). We have demonstrated the feasibility of applying RNA-Seq to study the adaptive immune repertoire. Although RNA-Seq lacks the sequencing depth of targeted sequencing (Rep-Seq), it can compensate by examining a larger sample size. Using ImReP, we have created a systematic atlas of 3.6 million Ig sequences across a broad range of tissue types, most of which have not been studied for Ig receptor repertoires. We also compared the GTEx tissues to track the flow of Ig clonotypes across immune-related tissues, including secondary lymphoid organs and organs encompassing mucosal, exocrine, and endocrine sites, and we examined the compositional similarities of clonal populations between these tissues. The Atlas of Immunoglobulin Repertoires (TAIR), is freely available at https://github.com/smangul1/TheAIR/wiki, is one of the largest collection of CDR3 sequences and tissue types. We anticipate this recourse will enhance future immunology studies and advance the development of therapies for human diseases. Using RNA-Seq to study immune repertoires has some advantages compared to Rep-Seq, including the ability to simultaneously capture clonotype populations from all the chains during a single run. Given a large number of large-scale RNA-Seq datasets becoming available, we look forward to scaling up the atlas of immune receptors in order to provide valuable insights into immune responses across various autoimmune diseases, allergies, and cancers.

# SIGNET: IDENTIFYING MUTATIONAL PROCESSES IN CANCER USING NEURAL NETWORKS

Lasse Maretty, Søren Besenbacher

Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark

Cancer is caused by the accumulation of mutations arising from multiple processes such as exposure to different mutagens and defects in DNA repair mechanisms, and information about such processes carries both etiological and prognostic relevance. The mutational processes contributing to a set of cancer samples can be identified and characterised by factoring a mutation matrix, which contains the number of observed mutations for each mutation type for every sample, using non-negative matrix factorization (NMF). In the NMF approach, the mutation type is typically defined using only the immediate sequence context and substitution type as the number of parameters grows exponentially in the number of genomic features used to define the type. However, the activity of different processes is known to vary across a range of additional genomic features such as longer range sequence context, transcription level, and replication timing, and hence potentially valuable information for discriminating different processes is lost in the NMF approach.

We here propose a new model in which the (non-negative) mutational susceptibility of a genomic site for a given mutational process is modelled as a nonlinear function of the features (parameterized by a neural network), which in turn renders the number of parameters linear in the number of features (for a fixed size neural network and number of processes). In a Bayesian variant of our model, we further introduce the horseshoe prior as a prior over the sample exposures to simultaneously model that many mutational processes are only rarely active and the wide range of mutation counts observed for active processes. Finally, we propose to use the marginal likelihood estimates obtained in the Bayesian approach to estimate the number of mutational processes. The model has been implemented in a software package called SigNet.

We estimate the model's performance on a number of large whole genome and exome sequencing datasets and compare with the classical NMF approach using different metrics. Finally, we also demonstrate how the model further enables inference of which features affect a given mutational process and hereby also represents a new framework for discovering such features.

# A CUSTOM BIOINFORMATICS PIPELINE FOR ANALYZING STRAND-SEQ DATA.

Carl-Adam Mattsson[1], Victor Gurjev[2], Peter M Lansdorp[1,2,3]

[1]Terry Fox Laboratory, BC, Cancer Research Centre, Vancouver, BC, Canada, [2]European Research Institute for the Biology of Ageing, University Medical Center Groningen, Groningen, Netherlands, [3]Department of Medical genetics, University of British Columbia, Vancouver, BC, Canada

Structural Variants (SVs) contribute greater diversity at the nucleotide level between two human genomes than any other form of genetic variation and are three-fold more likely to correlate with a genome-wide association study signal than single-nucleotide variants (SNVs) (Chiang et al. 2017). Using short-read, high-throughput sequencing technologies to uncover such variation has proven to be troublesome and the methods to detect SVs are independent on indirect inferences (e.g read-depth and discordant read-pair mapping). However, while larger (>5kb) copy number variants (CNVs) could be characterized by using read-depth-based algorithms, this approach often fails for smaller and balanced events, such as inversions. Another fundamental problem for SV detection from short-read sequencing is inherent to the predominant data type and the typical SV detection algorithm can thus be effective in unique sequences but fails within complex genomic regions (e.g repetitive and low map-ability regions), which are highly enriched for SVs. In addition, most SV discovery methods do not indicate the haplotype-origin for a given SV or requires parental sequencing information to achieve this. For a more complete description and interpretation of human genomic information in relation to phenotypes and e.g. cancer predisposition and response to therapies, it will therefore, be necessary to arrange sequence data into parental haplotypes and ascertain polymorphic inversions with respect to such haplotypes. All this (!) can be achieved using a single cell template strand sequencing technique called Strand-Seq (Sanders et al. 2017). Strand-seq complements other sequencing approaches by providing crucial information about the genetic make-up of individuals that cannot be obtained in any other way. To make Strand-Seq available for human studies worldwide is an immense challenge. Library construction, as well as data analysis, needs to be further developed, integrated and made user-friendly to allow accurate and rapid interpretation of results. Here we present a custom bioinformatics pipeline for analyzing Strand-seq data that streamlines the work-flow of raw sequence read alignment and putative variant calling to allow for accurate whole-genome haplotype (WGH) assembly and high confidence variant call-sets. In addition, relevant metric data are compiled and visualized, ensuring and reinforcing the potential of Strand-Seq as a robust sequencing method for uncovering clinically significant SVs and the assembly of WGH without additional parental genomic data.

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Consortium GT et al. 2017. The impact of structural variation on human gene expression. Nat Genet 49: 692-699.

Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. 2017. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat Protoc 12: 1151-1176.

# DEVELOPING A MOUSE MODEL FOR IDENTIFYING NEOANTIGEN-SPECIFIC T CELLS FOR ADOPTIVE CELL TRANSFER

Laura H McCulloch[1], David M Truong[1], Justin Mastroianni[2], Michelle Krogsgaard[2], Jef D Boeke[1]

[1]NYU Langone Health, Institute for Systems Genetics, New York, NY, [2]NYU Langone Health, Perlmutter Cancer Center, New York, NY

Adoptive cell transfer (ACT) has shown great potential as a cancer immunotherapy. In ACT, the patient is given T cells targeting a component of his tumor, allowing the immune response to attack the tumor and thus treat the patient. Tumor-specific antigens, or neoantigens, serve as promising targets for ACT. Whole-exome tumor sequencing enables prospective identification of neoantigens. These neoantigens can then be expressed on antigen-presenting cells and screened against T cells for their ability to induce responses in neoantigen-specific T cells (NSTCs). However, it is difficult to identify NSTCs, limiting the therapeutic use of ACT targeting neoantigens. Screening a patient's own peripheral blood lymphocytes (PBLs) restricts T cell identification to the patient's own T cell repertoire, which the tumor can often already evade due to continuous neoantigen expression. Alternatively, healthy donor PBLs can be used, but they must be matched to the patient's human leukocyte antigen (HLA) alleles and ABO blood group. Here, we will create a humanized mouse containing several synthetic immune genes to improve NSTC identification. This murine platform will couple the T cell diversity of the healthy donor PBL approach to the self-matched HLA typing of the patient PBL approach, enabling screening of a wide range of T cells produced in a living animal and identification of new NSTCs that will specifically target the tumor when introduced to the patient. As the first step towards model development, we have designed synthetic chromosome segments bearing the human TCR alpha and TCR beta gene loci, a set of common human major histocompatibility complex allelic variants, beta-2-microglobulin light chain, and CD8. We have begun to assemble these loci in yeast from bacterial artificial chromosomes and synthesized blocks of DNA, and plan to assess their expression in cultured mammalian cells in which the endogenous versions of our genes of interest have been knocked out. As part of this process, we have also designed constructs to investigate the regulatory DNA required for proper expression of each locus. Following successful assembly of each segment and appropriate gene expression in cultured mammalian cells, we will generate and characterize a novel humanized mouse model containing these assembled synthetic chromosome segments as a potential NSTC screening platform for cancer therapy.

# ISOLATING PROTEIN-OCCUPIED OPEN CHROMATIN REGIONS

Torrin McDonald[1], Rebecca Hurto[2], Peter Freddolino[2,3], Alan Boyle[1,3]

[1]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [2]University of Michigan, Department of Biological Chemistry, Ann Arbor, MI, [3]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI

Modern technology and recent advances in genomics have given rise to a variety of high-throughput methods that interrogate regulatory function of the human genome. In particular, the advent of techniques that measure open chromatin have enabled the identification of loci with active regulatory elements. These methods, in addition to mapping open chromatin, have been widely used to characterize footprints of transcription factors (TFs). The ability to infer TF binding motifs from footprints is dependent upon the lack of sequencing coverage over the footprint (strength of binding of TF) within open chromatin regions. This presents an unmet challenge for critical TFs that do not associate as strongly or as frequently with their binding sites. Indeed, it has been well documented that some TFs present very clear footprints (e.g. CTCF) while others lack any discernable difference in chromatin accessibility and footprinting techniques for these perform no better than a basic motif match. Here we present an orthogonal open chromatin assay that isolates TF binding footprints to be made into sequencing libraries. Briefly, we use micrococcal nuclease to extensively digest crosslinked chromatin to mononucleosomal size and extract the low molecular weight fragments. These fragments represent protein-occupied open chromatin regions protected from nuclease digestion. Analysis of these regions will allow us to finely map functional units within open chromatin and better understand how these regions are utilized in eukaryotic gene regulation.

# IDENTIFYING NOVEL REGULATORY ELEMENTS USING RELICS, A STATISTICAL FRAMEWORK FOR THE ANALYSIS OF CRISPR REGULATORY SCREENS

Patrick C Fiaux[1,2], Hsiuyi Chen[1], Ishika Luthra[3], Carolyn O'Connor[4], Graham McVicker[1]

[1]Salk Institute for Biological Studies, Integrative Biology Laboratory, La Jolla, CA, [2]UCSD, Bioinformatics and Systems Biology Program, La Jolla, CA, [3]Simon Fraser University, Biomedical Engineering, Vancouver, Canada, [4]Salk Institute for Biological Studies, Flow Cytometry Core, La Jolla, CA

High-throughput CRISPR/Cas9 screens are a powerful new tool for the systematic discovery of regulatory elements in the human genome. In these regulatory screens, thousands of guide RNAs (gRNAs) are delivered to cells to target potential regulatory sequences for mutation, activation or inhibition. The cells are then sorted into high- and low-expression pools based on the expression of a target gene. While, these screens have the potential to perform unbiased discovery of regulatory elements, they generate noisy data and the performance of analysis methods has not been rigorously assessed. Here we describe RELICS, a statistical framework for Regulatory Element Identification in CRISPR Screens. RELICS models the observed guide counts in different expression pools with a generalized linear mixed model. This approach is very flexible, can jointly model multiple expression pools (beyond just high and low), incorporate variability across guides, and accommodate over-dispersion. RELICS outperforms existing analysis methods on simulated data and we have applied it to identify regulatory elements in several published datasets. In addition, we have applied RELICS to data from a paired-guide regulatory screen that we performed for GATA3 in Jurkat T cells. We identify a total of 44 putative regulatory elements (FDR < 0.1) within the 2MB targeted region surrounding GATA3. Notably 26 of the identified elements lie within the same topological associating domain as GATA3, but only 2 overlap enhancers predicted by ChromHMM.

# POPULATION GENOMICS OF COPY NUMBER VARIATION IN A NATURAL POPULATION OF MAIZE WILD RELATIVE TEOSINTE

Wenbin Mei[1,2,7], Simon Renny-Byfield[1], Anne Lorant[1], Arun S Seetharam[3], John Doebley[4], Matthew B Hufford[5], Jeffrey Ross-Ibarra[1,2,6]

[1]University of California, Department of Plant Science, Davis, CA, [2]University of California, Center for Population Biology, Davis, CA, [3]Iowa State University, Genome Informatics Facility, Ames, IA, [4]University of Wisconsin, Department of Genetics, Madison, WI, [5]Iowa State University, Department of Ecology, Evolution, and Organismal Biology, Ames, IA, [6]University of California, Genome Center, Davis, CA, [7]Current address: Inari Agriculture, Cambridge, MA

Understanding genetic diversity in natural populations has long been a central theme for biologists. However, so far most efforts have focused on single nucleotide polymorphisms (SNPs) to characterize genetic diversity, demography, introgression and gene flow. Yet considerable evidence suggests an important evolutionary role for structural variation in shaping genetic and phenotypic diversity. Here, we use high depth whole genome sequencing data from a single natural population of the maize wild relative teosinte to study the population genomics of copy number variants (CNVs). We find about 6.7% of maize genome space shows CNVs (either deletion or duplication) in a single population, more than 95% of which is still segregating in the population. Our results show lower frequency CNVs in genic regions comparing to non-coding regions suggestion purifying selection on CNVs in genic regions. The SFS of deletions and duplications are skewed towards singletons. In addition, we model the CNV density and frequency using genomic features including DNase I hypersensitive sites base pair (bp), exonic bp, GERP elements, recombination rate and TE bp.We found that CNVs also bias Tajima's D calculated from SNPs in the regions containing CNVs. Although genome-wide CNVs appear largely neutral, we found that some CNVs may affect fitness and are enriched in selective sweep regions. Taken together, our results begin to shed light on the importance of structural variation in plant evolution and adaptation.

# DECODING THE TRANSCRIPTIONAL REGULATORY ATLAS OF THE MAIZE LEAF

Maria K Mejia-Guerra[2], Xiaoyu Tu[3], Jose A Valdes Franco[4], Pinghua Li[1], Edward S Buckler[2,4,5], Silin Zhong[3]

[1]College of Agronomy, Shandong Agricultural University, Tai'an, China, [2]Cornell University, Institute for Genomic Diversity, Ithaca, NY, [3]School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, China, [4]Cornell University, School of Integrative Plant Sciences, Section of Plant Breeding and Genetics, Ithaca, NY, [5]United States Department of Agriculture, Agricultural Research Service, Ithaca, NY

Complex traits related to the structure and physiology of the leaf are determinants of maize (*Zea mays* [*L.*]) yield. Important genes have been identified through transcriptomics, quantitative genetic studies, and evolutionary comparative analysis. Yet, the architecture of the gene regulatory network controlling the functioning of the maize leaf is unknown. In this study we use a novel ChIP-seq approach to catalog the binding of 104 transcription factors (TFs) that are highly expressed in leaves. This constitute the largest map of *in-vivo* regulatory regions in a given plant, resulting in 2,162,941 highly reproducible peaks that represent a non-overlapping set of 144,890 regulatory loci.

Peak information is used to (1) generate machine learning models of TF-binding, and (2) for network analysis to characterize the gene regulatory network that governs the phenotype of the maize leaf. We trained machine learning (ML) models to characterize the sequence preferences for the binding of each TF individually (accuracy ~85%) and to model TF-partnerships and co-binding (accuracy > 90%). Clustering of the TFs binding sequence preferences obtained with the ML models, closely recapitulates the evolutionary similarity among TFs. We observed some TFs being differential predictors of proximal and distal binding, and identified a group of TFs highly predictive across the co-binding contexts of the other TFs. Mutants of *Arabidopsis* - the reference plant - orthologs for these hubs show enrichment in observable phenotypes. A community analysis revealed a highly modular regulatory network, with the communities presenting non-random enrichment of GO terms. We observed a continuous range of community specialization, ranging from some including a large number of TFs, to communities with few/one TF highly specialized in few biological processes.

TF binding regions highly overlap with open chromatin regions (*i.e.*, 99% of ATAC-seq peaks), and shows enrichment for eQTLs and GWAS SNPs. The atlas and models presented here will be instrumental to interpreting the deluge of sequence variation information available for maize, in the light of its regulatory function, a key step to exploit non-coding variation in breeding and genome editing approaches.

# MULTIPLE WAVES OF PEOPLING NORTHEAST EUROPE: AND ANCIENT DNA STUDY THROUGH MESOLITHIC TO THE MIDDLE AGES.

Lehti Saag[1,2], Kristiina Tambets[1], Mait Metspalu[1]

[1]Institute of Genomics, University of Tartu, Estonian Biocentre, Tartu, Estonia, [2]Institute of Cell and Molecular Biology, University of Tartu, Department of Evolutionary Biology, Tartu, Estonia

Studies of ancient DNA have resketched the peopling story of Europe highlighting the major role of three distinct mass immigrations in the Mesolithic, Neolithic and Early Bronze Age periods. Here we take a closer look at these processes in the northeast corner of Europe, concentrating mostly on what is now Estonia. One key difference of this region from the rest of Europe is high (30-40%) frequency of chr Y haplogroup N3a, which is otherwise spread in Siberia. Genome wide the Siberian component makes up only a few percent of the genepool of Northeast Europeans. We sample each major cultural transition in the region and follow the accompanying demographic changes. The earliest Mesolithic samples cluster together with the Western hunter-gatherers of Europe (WHG). The arrival of Comb Ceramics Culture (CCC) is coupled with Eastern hunter-gatherer (EHG) genetic ancestry. The arrival of Corded Ware Culture (CWC) in Late Neolithic/Early Bronze Age, is part of the massive immigration to Europe from the Steppes of contemporary Russia, and installs a new genepool. We then turned to study the genetic ancestry of individuals (N=56) from more recent but as yet genetically unstudied cultural traditions in Estonia: i)the Late Bronze Age (EstBA) stone-cist graves (1200–400 BC), ii) the Pre-Roman Iron Age (EstIA) tarand cemeteries (800/500 BC–50 AD), iii) Pre-Roman to Roman Iron Age Ingrian (IngIA) (500 BC–450 AD) and iv) Middle Ages samples from Estonia (EstMA) (1200–1600 AD). Compared to the preceding CWC people the genetic ancestry of the EstBA samples shifts back toward that of WHG. While it is the EstIA where we see minute fraction of the Siberian ancestry component for the first time. Moreover, the respective chr Y variant – N3a – is also detected in EstIA samples and not in bearers of the older cultures. This eastern/Siberian ancestry reached the coasts of the Baltic Sea no later than the mid-first millennium BC; i.e. in the same time window as the diversification of west Uralic/Finnic languages. Furthermore, phenotypic characteristics often associated with modern Northern Europeans like light eyes, hair and skin as well as lactose tolerance can be traced back to the Bronze Age in the Eastern Baltic. During the Middle Ages the Estonian population is socially structured into subpopulations of landlords of likely mostly German origin and the common people. In the pilot phase of the study we have confirmed, that burials from high class cemetery resemble genetically contemporary Germans, while rural and urban cemeteries of common people yield aDNA samples genetically reminiscent of the current Estonian population.

# GENOMIC ANALYSIS OF CONTAGIOUS CANCERS OF CLAMS

Marisa Yonemitsu[1], Brian Beal[2], Brian W Davis[3], Michael J Metzger[1]

[1]Pacific Northwest Research Institute, PNRI, Seattle, WA, [2] Univ. of Maine at Machias, School of Marine Sciences, Machias, ME, [3]Texas A&M University, Department of Veterinary Integrative Biosciences, College Station, TX

Cancer is normally an evolutionary dead-end—neoplastic cells that arise and evolve within an organism either regress or they kill their host, and the death of the host marks the death of the cancer lineage. However, in some cases, neoplastic cells develop the ability to spread from individual to individual, turning from conventional cancers into clonal contagious cancer lineages. The natural transmission of cancer cells has been observed in two mammals (Tasmanian devils and dogs), and we have found that a leukemia-like disease in soft-shell clams (*Mya arenaria*) is due to the horizontal spread of a clonal cancer lineage. We also analyzed mussels (*Mytilus trossulus*), cockles (*Cerastoderma edule*), and carpet shell clams (*Polititapes aureus*) and found that the neoplasias in all three of these species are due to independent transmissible cancer lineages. We have assembled an initial draft reference genome for *M. arenaria* using 10X Chromium methods (Phaseblock N50: 66 kb), and we are generating a new reference genome using 10X methods combined with HiC data for greater contiguity. Using these references, we are investigating genomic changes in the evolution of this unique cancer lineage, including SNPs, structural variation, and copy number variation. In particular, we found a retrotransposon, *Steamer*, which is expressed and amplified in genomic DNA of the contagious cancer lineage, expanding from 2-10 copies per haploid genome in normal animals to >100 in neoplastic cells. Genomic analysis of cancer samples from isolated clam populations in Maine and Prince Edward Island confirm initial qPCR analysis, and show that at least 130 sites are found in cancer cells from both populations. These common sites likely integrated early in the evolution of the cancer lineage, and have been conserved either as passenger or driver mutations. We also found many insertions that are unique to only one subgroup (494 and 144, unique insertions in Maine and PEI, respectively), showing either continued amplification or deletion after divergence of the cancer in the two populations. These new integration events and other genomic changes have likely played a role in oncogenesis and continued evolution of the cancer with its hosts.

# IMPROVING CLINICAL ASSESSMENT OF WHOLE GENOME SEQUENCING DATA: THE CAGI SICKKIDS CHALLENGES

Stephen Meyn[1,3], Jesse Hunter[2], Rupa Udani[2]

[1]University of Wisconsin, Center for Human Genomics, Madison, WI, [2]University of Wisconsin, State Laboratory of Hygiene, Madison, WI, [3]Hospital for Sick Children, Department of Paediatrics, Toronto, Canada

Most children suspected of a rare genetic disorder remain undiagnosed, in large part due to our limited ability to correctly classify genomic variants. The Critical Assessment of Genome Interpretation (CAG() challenges are designed, in part, to spur efforts to improve variant interpretation. We now report the results of two CAGI challenges that focused on using novel methods to analyse whole genome sequencing (WGS) data in order to diagnose children suspected of genetic disease.

Both challenges involved children suspected of a genetic disorder who had undergone WGS but remained undiagnosed after evaluation by the SickKids Genome Clinic project (see Stavropoulos et al., npj Genomic Medicine 2016 and Lionel, et al., Genetics in Medicine 2017). CAGI teams were supplied with the WGS data and clinical phenotypic data, including HPO terms, for each patient.

Four teams attempted to identify diagnostic variants in the first cohort, which consisted of 25 patients suspected of genetic disorders who had also undergone microarray analyses. Each team's approach was unique. Exomizer was used by two teams, dcGO and fathmm by one, and the fourth used conventional clinical diagnostic tools. Of the 187 diagnostic variants proposed by the four teams, most were nominated by only one team and did not fully explain the patient's phenotype. However, two of the ten variants proposed by two teams were deemed to be diagnostic.

Eight teams analysed the second cohort, which consisted of six patients with eye diseases, seven with neurologic symptoms, and eleven with connective disorders. This challenge was more difficult, as the genomes were not linked to the patients' clinical descriptions. As with the first cohort, each teams' approach utilized a unique combination of bioinformatic filtering and in silico classifier tools. The teams correctly matched the genomes to the three disease categories in the majority of cases. Matching genomes to patients proved difficult, although three teams performed significantly better than chance, in part because they attempted to relate declared ethnicity to genomic variants. Of note, there was only a weak correlation between complexity of the clinical phenotype and the ability to match phenotype to genotype. Despite partial success in matching genomes to patients, no candidate variants were deemed to be diagnostic. As with the first challenge, variants nominated by the teams were frequently not clinically plausible.

In summary, CAGI challenges involving clinical WGS data offer a unique opportunity to improve variant classification, but the competitions could be improved by reducing the difficulty of the challenges and adding clinical expertise to the teams.

# HOMOTYPIC COOPERATIVITY AND COLLECTIVE BINDING ARE DETERMINANTS OF BHLH SPECIFICITY AND FUNCTION

Christian A Shively[1,2], Jiayue Liu[1,2], Xuhua Chen[1,2], Kaiser Loell[1,2], Robi D Mitra[1,2]

[1]Washington University School of Medicine in St. Louis, Department of Genetics, St Louis, MO, [2]Washington University School of Medicine in St. Louis, Center for Genome Sciences and Systems Biology, St Louis, MO

Eukaryotic cells often express transcription factors (TFs) that bind to very similar DNA sequences in vitro but bind at different genomic loci and perform different functions in vivo. To understand how this in vivo specificity is achieved, we analyzed two yeast bHLH proteins, Cbf1 and Tye7, which have nearly identical binding preferences in vitro, but bind at almost completely non-overlapping target loci in vivo. We dissected the determinants of specificity for these two proteins by making a number of chimeric TFs where we swapped different domains of Cbf1 and Tye7 and determined the effects on in vivo binding and cellular function. From these experiments, we learned that the Cbf1 dimer achieves its specificity by binding cooperatively with other Cbf1 dimers bound nearby. This type of interaction, known as homotypic cooperativity, allows Cbf1 to outcompete Tye7 at most promoters that contain their shared consensus binding site. In contrast, we found that Tye7 achieves its specificity by binding cooperatively with three other DNA binding proteins, Gcr1p, Gcr2p, and Rap1p. Tye7 binds in an "all-or-none" fashion with its partner TFs, and most promoters (63%) that are bound by Tye7 do not contain a consensus Tye7 binding site. Instead, different Tye7-bound loci contain different subsets of Gcr1/2, Rap1, and Tye7 motifs, and these motifs display no fixed spacing or orientation. These features are the hallmarks of a "TF Collective", a recently described but poorly understood model of TF cooperativity. Using this information, we were able to build simple models to accurately discriminate bound and unbound genomic loci for both Cbf1p and Tye7p (AUROCs of 0.92 and 0.96, respectively) and to assign bound targets to the correct TF. To demonstrate the generality of these findings, we show that Usf1p, a human homologue of Cbf1 that is a therapeutic target for obesity and metabolic disease, also utilizes homotypic cooperativity to achieve its function. Finally, we describe new methods for the high throughput analyses of homotypic cooperativity and collective binding. These results demonstrate that the genome-wide binding targets of paralogous TFs can be discriminated using sequence information alone, and provide lessons about TF specificity that can be applied across the phylogenetic tree.

# IMPACT OF ALTERNATIVE POLYADENYLATION ON GENE AND PROTEIN EXPRESSION LEVELS IN HUMANS

Briana E Mittleman[1,2], Shane F Warland[2], Kristen Patterson[2], Zheng Tony[2], Pott Sebastian[3], Gilad Yoav[2,3], Yang I Li[2,3]

[1]University of Chicago, Genetics, Genomics, and Systems Biology, Chicago, IL, [2]University of Chicago, Genetic Medicine, Chicago, IL, [3]University of Chicago, Human Genetics, Chicago, IL

Over half of all human protein coding genes possess two or more polyadenylation sites, resulting in the production of a diverse number of mRNAs with alternative 3' UTRs. Yet, the function of alternative polyadenylation (APA), if any, is unknown for most genes. Here, by mapping genetic variants associated with APA in a panel of 55 Yoruba lymphoblastoid cell lines, we carefully dissected the impact of APA on gene regulation at multiple levels.

To this end, we first evaluated our ability to infer polyadenylation site usage (PAU) directly from RNA-seq data. We observed that computational identification of PAU using RNA-seq drastically underestimates the amount of APA compared to experimental approaches that directly measure polyadenylation sites such as 3'-prime RNA-seq. This finding therefore motivated us to use 3' RNA-seq to directly measure PAU in total RNA extracted from Yoruba LCLs (n = 55).

We also performed 3'-seq to measure PAU in RNA extracted from the nuclear fraction of the same LCLs. Analysis of PAU in the nuclear and total mRNA fractions revealed that 2424 polyadenylation sites in 1877 genes showed differential usage across the two fractions (10% FDR, $\Delta$PAU > .2). This implies a pervasive role of APA in controlling mRNA decay and/or mRNA export from the nucleus to the cytoplasm. Therefore, APA can profoundly shape the mRNA content of the translated transcriptome.

By correlating genetic variation to polyadenylation site usage in the 55 individuals, we identified 363 and 623 polyadenylation QTLs (apaQTLs) in the total and nuclear fractions, respectively (10% FDR). As expected, apaQTLs are highly enriched near the 3'UTRs and in sequence that contain polyadenylation signals, suggesting that a large number of genetic variants directly affect PAU. We also found that a substantial number of apaQTLs colocalize with eQTLs that show no effect on enhancer or promoter activity as measured by H3K27 acetylation or DNase-I hypersensitivity. Remarkably, a subset of the apaQTLs are associated with protein expression levels, but not mRNA expression or ribosome occupancy levels.

Thus, our study establishes APA as an important regulatory mechanism that underlies the genetic regulation of gene and protein expression levels.

# GENETIC AND ENVIRONMENTAL INFLUENCES ON PROTEIN VARIABILITY IN THE HUMAN IMMUNE SYSTEM

Michael D Morgan[1], Etienne Patin[2], Bernd Jagla[3], Milena Hasan[3], Massimo Mangino[4], Tim D Spector[4], Lluis Quintana-Murci[2], John C Marioni[1,5,6]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [2]Human Evolutionary Genetics Unit, Institut Pasteur, Paris, France, [3]Cytometry and Biomarkers UTechS, Institut Pasteur, Paris, France, [4]2Department of Twin Research & Genetic Epidemiology, Kings' College London, London, United Kingdom, [5]EMBL-EBI, Wellcome Genome Campus, Hinxton, United Kingdom, [6]Cancer Research UK - Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, United Kingdom

Cellular heterogeneity is a hallmark of multicellular organisms and tissues. Recently developed single-cell measurement technologies enable the quantification of heterogeneity at the level of individual mRNAs or proteins. Additionally, collecting single-cell measurements across population health cohorts allows us to link such heterogeneity to important health factors, including age and genetic variation.

Following this principle, we have analysed single-cell protein measurements across various flow cytometry-defined populations of immune cells in identical and non-identical twins, in the TwinsUK cohort, in order to quantitate the contribution of genetic variation to protein variability traits. Importantly, partitioning of heritability demonstrates that trans factors dominate the heritable component of variation in protein variability. Focussing on variability traits with a non-zero additive genetic component we have identified specific genetic loci that influence protein variability, called variability-QTLs. By highlighting several examples we demonstrate how variability-QTL mapping has the capacity to discover novel contributions of genetic variation to immune cell behaviour and function.

# RECOMPLETING THE *CAENORHABDITIS ELEGANS* GENOME

Jun Yoshimura[1], Kazuki Ichikawa[1], Massa J Shoura[2], Karen L Artiles[2], Idan Gabdank[3], Lamia Wahba[2], Cheryl L Smith[2,3], Mark L Edgley[4], Ann E Rougvie[5], Andrew Z Fire[2,3], Shinichi Morishita[1], Erich M Schwarz[6]

[1]The University of Tokyo, Computational Biology and Medical Sciences, Tokyo, Japan, [2]Stanford University, Pathology, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA, [4]University of British Columbia, Zoology and Michael Smith Laboratories, Vancouver, Canada, [5]University of Minnesota, Genetics, Cell Biology, and Development, Minneapolis, MN, [6]Cornell University, Molecular Biology and Genetics, Ithaca, NY

*Caenorhabditis elegans* has the first multicellular eukaryotic genome sequenced to apparent completion. Although this assembly employed a standard *C. elegans* strain (N2), it used sequence data from several laboratories, with DNA propagated in bacteria and yeast. Thus, the N2 assembly has many differences from any *C. elegans* available today. To provide a more accurate *C. elegans* genome, we performed long-read assembly of VC2010, a modern strain derived from N2. Our VC2010 assembly has 99.98% identity to N2, but with an additional 1.8 Mb including tandem repeat expansions and genome duplications. For 116 structural discrepancies between N2 and VC2010, 97 structures matching VC2010 (84%) were also found in two outgroup strains, implying deficiencies in N2. Over 98% of N2 genes encoded unchanged products in VC2010; moreover, we predicted ≥53 new genes in VC2010. The recompleted genome of *C. elegans* should be a valuable resource for genetics, genomics, and systems biology.

# PREDICTION OF NEURONAL PHENOTYPES USING RNA MODIFICATIONS AND GENOMIC DISSECTION OF CELL COMPARTMENTS IN SINGLE NEURONS

Andrea B Kohn[1], Leonid L Moroz[1,2]

[1]University of Florida, Whitney laboratory, St. Augustine, FL, [2]University of Florida, Neuroscience, Gainesville, FL

Virtually all neurons are remarkably different! Starting with pioneering work of Ramon-y-Cajal to the recent tsunami of single-cell genomics – the scope of recognized neuronal diversity and specificity is only increasing exponentially. Clustering scRNA-seq approaches can group neurons unbiasedly but failed to predict neuronal phenotype de novo. However, the single-cell epitranscriptome has far greater predictability of phenotype than scRNA-seq. Specifically, we developed an ultra-sensitive (UPLC-MS/MS) method to measure RNA modifications in individually identifiable neurons, which allowed prediction of their phenotypes with nearly 100% accuracy. Epitranscriptome data also allowed us to distinguish left and right homologs of the same neuronal class. Second, by performing RNA-seq from the nucleus and cytoplasm of the very same identified neurons, we identified more than 600 transcripts differentially expressed including those encoded RNA binding proteins and RNA modification. We validated, by in situ hybridizations, the transcripts for the enzymes that perform these modifications and found most are expressed in the nucleus. These nuclear-retained transcripts and RNA modifying enzymes can act as master regulators controlling the fate and polarized distribution of RNA in neurons. We conclude that nuclear transcriptomes alone have a limited predictive-power of neuronal cell-type phenotypes, but this first unbiased cell-compartment specific genomic portrait of identified neurons suggests that most neurons are epitranscriptomically unique - they learn and age differently, opening a new perspective toward analysis of neuronal individuality and personalized medicine.

# SEQUENCING OF ENTIRE BRAINS AT THE SINGLE-CELL RESOLUTION: PRINCIPLES OF MOLECULAR CLASSIFICATION AND PERIODIC SYSTEM OF NEURONS

Leonid L Moroz

University of Florida, Neuroscience, St. Augustine/Gainesville, FL

There is more than one way to develop neuronal complexity, and animals frequently use different molecular toolkits to achieve similar functional outcomes (=convergent evolution). As the first step toward the phylogenetical classification of cells and the deciphering evolutionary history of neuronal lineages, we employed microfluidics, single-neuron capture and sequenced >1.2 million cells from 14 reference species across the entire animal kingdom. Second, for selected animals, including Aplysia, we performed scRNA-seq of all individual neurons composing the entire brain (Brain-seq). By combining tools of modern phylogenomics, single-neuron sequencing, machine learning, single-cell proteomics and metabolomic across Metazoa, it is possible to reconstruct the genealogies of neurons tracing them to ancestral secretory cells. 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 scale from ctenophores and cnidarians to cephalopods and primates. Comparative data suggest that neurons, and perhaps synapses, evolved at least three times during ~600 million years of animal evolution. Also, there were several independent events of the nervous system centralization either from a common ancestor with diffuse, nerve-net type nervous system. From the 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 34 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 for synthetic Neuroscience of the future. Such phylogenetic framework of Neuronal Systematics might be a conceptual analog of the Periodic System of Chemical Elements.

# CONNECTING CHROMATIN STRUCTURE TO REGULATORY FUNCTION: EXTENDING SPLIT-POOL RECOGNITION OF INTERACTIONS BY TAG EXTENSION (SPRITE) TO ASSOCIATE STRUCTURAL COMPLEXES WITH TRANSCRIPTIONAL REGULATION

Rachel Murphy*[1], Pranitha Vangala*[1], Patrick McDonel[1], Sofia A Quinodoz[2], Mitchell Guttman[2], Manuel Garber[1]

[1]University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA, [2]California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA

Current knowledge of gene regulation is limited due to the complexities by which non-coding elements of the genome interact to control transcription. We have found that Split-Pool Recognition of Interactions by Tag Extension (SPRITE), which uniquely tags interacting DNA molecules, can capture the complexity of multi-dimensional interactions at a high resolution. The sequencing depth required to map genome wide spatial interactions with SPRITE is prohibitive, we developed SPRITE-immunoprecipitation (SIP) to enrich for interactions that include promoter regions.

Here we show that SIP provides a cost effective way to study the structural interactions associated with promoters. We applied SIP to uncover the dynamics of regulatory complexes in response to immune stimulation. We found that a majority of genes have simple enhancer-promoter interactions and only a subset of genes have complex regulatory architecture. Additionally, we found that a subset of these complex structures are transient, and change configuration with immune stimulation. To understand how these simple and complex configurations result in transcriptional regulation, the transcriptional output of every configuration must be measured.

The technology used in SIP is versatile and provides opportunities to extend SIP in ways that directly capture the transcriptional outputs resulting from a single configuration. In addition, 5'EU labeling, prior to crosslinking, will allow us to isolate nascent RNA and identify correlations between transcriptional output and the transient enhancer-promoter interactions. We have constructed methods to bridge the gap between our understanding of the multi-dimensional structural configuration of chromatin and its regulatory nature of the transcriptome.

# GENOME-WIDE CHROMATIN STRUCTURAL ANALYSIS REVEALS INTER-CHROMOSOMAL TRANSLOCATIONS IN B-CELL LYMPHOMA CELL LINE

Luis A E Nagai[1], Sung-Joon Park[2], Kenta Nakai[1,2]

[1]University of Tokyo, Department of Computational Biology and Medical Science, Kashiwanoha, Chiba, Japan, [2]Human Genome Center, Institute of Medical Science - University of Tokyo, Shirokanedai, Minato, Japan

The genome is non-randomly organized into three-dimensional structure inside the nucleus. Alterations of this structure are commonly associated with disease development. In particular, the disruption of humoral immune system presents abnormal gene regulation accompanied by chromatin reorganizations. In B-cell lymphomas, a hallmark is the diverse chromosomal translocations, mostly involving transcriptionally deregulation in oncogenes. Here, we dissect the chromatin dynamics in normal and abnormal mouse cells. Although the chromatin compartments were highly preserved among several cells, 5.2% of B-cell genes in repressive compartment switched to permissive compartment in lymphoma with increased expression levels. These genes are involved in B-cell lymphoma related biological processes, such as V(D)J recombination. By assessing pairs of regions that have more reads between them than expected by chance, significant interactions (p-value < 0.001) were selected and classified in intra- and inter-interaction. An inexpressive amount of significant inter-chromosomal interactions was observed in normal cells. However, in B-cell lymphoma, 34% of the significant interactions were between two different chromosomes. Surprisingly, around 70 thousand interactions (23% of all significant interactions) were between chromosome 7 and chromosome 17. Our findings suggest the presence of intricate crosstalk between the higher-order chromatin structure and cancer development. In addition, Hi-C chromatin interaction maps can provide means to identify chromosomal translocations in cancer cell samples and may contribute to poorly diagnostics in this type of lymphoma.

# LINE-1 ACTIVATION IN CANCER AND ITS IMPACT ON CANCER IMMUNOTHERAPY

Fabio Navarro, George Gayed, Mark Gerstein

Yale University, Molecular Biophysics and Biochemistry, New Haven, CT

Long interspersed nuclear element 1 (LINE-1) is a primary source of genetic variation in humans and other mammals. Despite its importance, LINE-1 activity remains difficult to study because of its highly repetitive nature.

Here, we developed and validated a method called TeXP to gauge LINE-1 activity accurately. TexP builds mappability signatures from LINE-1 subfamilies to deconvolve the effect of pervasive transcription from autonomous LINE-1 activity. In particular, it apportions the multiple reads aligned to the many LINE-1 instances in the genome into these two categories. Using our method, we evaluated well-established cell lines, cell-line compartments and healthy tissues and found that the vast majority (91.7%) of transcriptome reads overlapping LINE-1 derive from pervasive transcription. We validated TeXP by independently estimating the levels of LINE-1 autonomous transcription using ddPCR, finding high concordance. Next, we applied our method to comprehensively measure LINE-1 activity across healthy somatic cells, while backing out the effect of pervasive transcription. Unexpectedly, we found that LINE-1 activity is present in many normal somatic cells. This finding contrasts with earlier studies showing that LINE-1 has limited activity in healthy somatic tissues, except for neuroprogenitor cells. Interestingly, we found that the amount of LINE-1 activity was associated with the with the amount of cell turnover, with tissues with low cell turnover rates (e.g. the adult central nervous system) showing lower LINE-1 activity.

Revealing autonomous activity of LINE-1 allowed us to uncover the role of LINE-1 in tumor progression. We used TeXP to analyze datasets from bladder cancer to investigate the role of LINE-1 in patients treated with Atezolizumab (anti-PD-L1). When analyzing tumors from over 300 patients that have undergone checkpoint blockade immunotherapy we found that tumors with low LINE-1 expression have a significantly better prognosis than patients with high LINE-1 activity. Further investigation of these patients indicates that bladder tumors with low LINE-1 expression have significantly higher chances of responding to CBI and are enriched for immune-active tumors. A potential mechanism is perhaps LINE-1 promoting tumor genomic instability through the creation of INDELs, which in turn promotes tumor recognition by the immune system. This would, therefore, indicate that the persistent expression of LINE-1 in tumors acts as a marker of immune-dysfunction, and ultimately of poorer outcomes of cancer immunotherapy treatments.

Altogether, our results show how accounting for pervasive transcription is critical to accurately quantify the activity of highly repetitive regions of the human genome.

# EXAMINING THE CONSISTENCY BETWEEN SURVEY RESPONSES AND GENETIC INFORMATION OF ANCESTRYDNA CUSTOMERS

Luong T Nguyen[1], Natalie Telis[1], Harendra Guturu[1], Julie M Granka[1], Jake Byrnes[1], Keith Noto[1], Eurie L Hong[1], Ken G Chahine[2], Catherine A Ball[1]

[1]Ancestry.com DNA, LLC, San Francisco, CA, [2]Ancestry.com DNA, LLC, Lehi, UT

AncestryDNA has the largest genomics database in the world, with more than 10 million DNA samples and a large number of genetic matches and survey responses for those individuals. In this research, we aim to examine the consistency between survey responses and genetic information of research consented genotyped individuals. The similarity between individuals is characterized by IBD matches and available connections in Genetic Communities (Han [1]). Publicly available methods for collaborative filtering is then used to investigate whether the similarities between genetic information of individuals are in agreement with the similarities of their survey responses. A portion of the current survey answers (20%, 40%, and 60%) is randomly hidden and the survey answers by similar consented individuals are used to infer the hidden data. The correlation between inferred answers and ground truth is used to evaluate the performance of different collaborative filtering methods. Overall, there is a high level of consistencies between survey responses and genetic information of these individuals, but the levels of consistencies vary slightly across different survey questions.

[1] https://www.ncbi.nlm.nih.gov/pubmed/28169989

# GENOMIC PREDICTIONS FOR SELECTING LOCALLY ADAPTED BEEF CATTLE

Sara M Nilson[1], Troy N Rowan[1], Robert D Schnabel[1,2], Jared E Decker[1,2]

[1]University of Missouri, Division of Animal Sciences, Columbia, MO,
[2]University of Missouri, Informatics Institute, Columbia, MO

Genotype-by-environment interactions play a significant role in the lifetime productivity of an animal. We hypothesize that by including loci responsible for environmental adaptation in prediction models the accuracy and speed of region-specific genomic selection will increase. To develop environment-specific genomic predictions, 30-year normals for temperature, precipitation, and elevation were analyzed with K-means clustering to divide the United States into 9 distinct regional environments. Over 12,000 Gelbvieh, 15,000 Simmental, and 17,000 Red Angus cattle distributed across the environmental regions were included in the analyses. All genotypes were phased with Eagle v2.4 and imputed with Minimac3 on ARS-UCD1.2 up to ~850,000 SNPs. Breed and region-specific Bayesian Sparse Linear Mixed Models were fit in GEMMA v0.94 for 3 pre-adjusted phenotypes: birth weight, weaning weight, and yearling weight. Separate models were fit for each region, thus a possible 27 models (3 phenotypes by 9 regions) were fit for each breed. In addition, we used sommer v3.6 to fit multivariate models for each phenotype by breed, in which the phenotypes from each region were fit as separate dependent variables. Regions with less than 300 samples were removed from the analyses. Accuracy of the models were assessed with a three-fold cross-validation and by calculating the square of the correlation between the phenotype and the predicted breeding value. A predicted breeding value was calculated for every animal for all three phenotypes across the environmental regions to investigate reranking across environments. These prediction models will serve as a valuable tool for producers to identify how well an animal and their progeny will perform in their environment before making selection or breeding decisions.

# IDENTIFYING SUBSETS OF CELL-SPECIFIC REGULATORY SEQUENCE FEATURES FROM HETEROGENEOUS TISSUE-SPECIFIC CHROMATIN STATE DATA USING PREDICTIVE GAPPED K-MER FEATURES

Jin Woo Oh[1], Michael A Beer[1,2]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, Mckusick-Nathans institute of Genetic Medicine, Baltimore, MD

Chromatin accessibility of genomic regions is highly associated with regulatory activity, but cell and tissue accessible regions are often functionally diverse with distinct combinations of transcription factor binding sites (TFBS). The ability to recognize shared patterns of TFBS combinations among subsets of chromatin open regions is crucial for annotating regulatory regions potentially coming from distinct cell types within the tissue. Thus clustering chromatin open regions by their functional similarity is an important, yet computationally difficult problem for understanding diverse modes of regulation, mainly because TFBS tend to be short and flexible. Our previous work has shown that cell specific regulatory regions can be accurately predicted with support vector machine using gapped-kmer features (gkm-SVM). Since transcription factor binding is usually cell-specific, the effectiveness of gapped-kmer feature to predict cell- specific regulatory region indicates that regulatory regions with different TFBS combinations can be discriminatively characterized by gapped kmer content. Thus, clustering sequences by gapped kmer features can potentially identify subpopulations of regulatory sequences that share common TFBS combinations. Here we present a new sequence clustering method using predictive gapped kmer features collected through gkm-SVM training. To validate our method, we have applied our algorithm to synthetic DNA sequences seeded with different combinations of known TFBS motif PWM and demonstrate that our method can be effective in discovering combinatorial regulatory vocabularies. Our method is expected to have abroad applications such as identification of cell-specific regulatory regions from tissue-level DHS data.

# TBROWSE: A MODULAR GENE TREE VIEWER FOR EXPLORING FUNCTIONAL CONSERVATION

Andrew J Olson[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

A gene tree reconstructs the evolutionary history of a gene family, revealing speciation and duplication events based on sequence homology. The Ensembl team at the EBI uses its Compara pipeline to build gene trees for various sets of species (vertebrates, plants, fungi, protists, metazoa, bacteria, and a pan-taxonomic sampling of reference genomes.) These trees can be visualized on the corresponding Ensembl web sites and are also available via the Ensembl REST API for use in 3rd party applications.

We are developing a generic tree API (www.genetrees.org) and a modular viewer (www.tbrowse.org) to support interactive exploration of gene families and associated data. The API is backed by Apache Solr and has been populated with trees from recent Ensembl Compara database builds as well as a set of Panther gene trees. Each tree node is stored as a document with a predetermined set of fields to identify the node and its parent node. Dynamic fields in the Solr schema allow us to incorporate various gene and protein annotations such as InterPro domains. The API has a /search endpoint for finding genes of interest and a /tree endpoint that retrieves the nodes in a given tree and assembles them into a nested JSON response. If a pruned tree is required, the Solr GraphQueryParser can retrieve ancestral nodes for a subset of genes in a tree identified by an arbitrary set of filters.

TBrowse is a React and Redux based JavaScript application that lets users search for and visualize trees in a customizable layout. The code is designed to support plugins to fetch and visualize data associated with nodes in the tree. Core components render the visible tree nodes and associated data such as the multiple sequence alignment, InterPro domain architecture, neighborhood conservation, and plain text or numeric fields included in the API response.

This work is supported by an NSF grant IOS-1127112, and from USDA-ARS (1907-21000-030-00D).

# BENCHMARKS FOR 'DARK MATTER' OF THE HUMAN GENOME.

Nathan D Olson, Justin Wagner, Jennifer McDaniel, Justin M Zook, Genome In A Bottle Consortium

National Institute of Standards and Technology, Biosystems and Biomaterials Division, Gaithersburg, MD

The human genome "dark matter" is comprised of large low-complexity and highly repetitive regions as well as large duplications that are challenging to characterize with short-read sequencing methods. The Genome In A Bottle (GIAB) consortium was established to authoritatively characterize genomes for use in method validation and benchmarking. There are seven genomes in the GIAB portfolio including commonly used NA12878 and two openly-consented trios, of Ashkenazi Jewish (HG002 - HG004) and Han Chinese ancestry (HG005 - HG007). GIAB has developed small variant benchmark sets for these 7 genomes by integrating data from multiple sequencing platforms and variant calling methods, accounting for the different methods' strengths and biases. Due to the limitations of short-reads sequencing methods the current benchmark sets are limited to easier regions of the genome. However, the more challenging "dark matter" includes clinically relevant small variants as well as structural variants, therefore expanding the GIAB characterization efforts is a priority. Advances in DNA sequencing technologies provide the opportunity to characterize parts of the genome "dark matter". Specifically, linked reads from 10x Genomics enable mapping of short reads to difficult to map regions, as well as long reads from PacBio circular consensus sequencing (CCS), which generates high accuracy reads (99%) with read-lengths ~15kb, and ultra-long Oxford Nanopore (ONT) data with > 15X coverage with reads longer than 100kb enable mapping across repetitive regions. While the current benchmark small variant set covers ~90% of the reference genome with over 3.5 million variants, adding the CCS and linked-read data preliminarily added ~150,000 new variants and 80 million bp of the reference genome, over half in regions that are difficult to map with short reads. Structural variants are challenging to call using short-read data but make up a majority of bases altered by genome variation. To build upon the small variant characterization efforts long-read, linked-read, along with optical and electronic mapping data in addition to short-read data were used to generate a draft benchmark structural variant set for the Ashkenazi Jewish trio son (HG002) with plans to develop benchmark structural variant sets for all seven genomes. The HG002 SV benchmark set includes ~12,000 insertions and deletions along with 2.69 Gbp benchmark regions supported by diploid assemblies. We have shown that comparing to this benchmark set reliably identifies both false positives and false negatives across multiple technologies. By incorporating new linked and long-read with existing datasets we have expanded the authoritatively characterized regions of the GIAB genomes to enable robust benchmarking of more of the genome "dark matter".

# COMBINED DNASE FOOTPRINTING AND MOTIF MATCHING MODEL FOR IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES

Ningxin Ouyang[1], Alan P Boyle[1,2]

[1]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Department of Human Genetics, Ann Arbor, MI

Transcription is tightly regulated by cis-regulatory DNA elements where transcription factors (TFs) can bind. Thus, identification of transcription factor binding sites (TFBSs) is key to understand gene expression and transcriptional regulation. The standard approaches for TFBSs prediction such as Position weight matrices (PWMs) and ChIP-seq are widely used but have their drawbacks such as high false positive rate and difficulty creating specific antibodies respectively. The traditional DNase I footprinting assay is a classical method to investigate *in vivo* protection of DNA by individually bound proteins. It has been adapted for high-throughput sequencing (DNase-seq), which allows genome-wide footprinting in single nucleotide resolution. Here we integrate both DNase I hypersensitivity and genome sequence information in a multivariate Hidden Markov Model to detect binding sites for specific TFs. To search for footprint-like regions with a depletion of DNase I cuts surrounded by regions with high numbers of cuts, the basic structure of our model includes hidden states of TFBSs and their surrounding small peaks. By incorporating DNase-seq data and PWM information, our model is able to detect footprints with the desired DNase digestion pattern and matching motifs. Since different motifs are treated as separate states in the model, our method has the potential of targeting multiple TFs in a single model. This new method annotates binding sites for desired TFs across input regions automatically with no pre-generated candidate binding sites required. Also, as an unsupervised model, its application is not limited to TFs with available ChIP-seq data. Compared to other supervised and unsupervised footprinting algorithms, our method has the best overall performance as it has the best ranking in both precision recall and operating characteristic curve area under the curve across 101 tested TFs.

# COMPUTATIONAL APPROACHES TO ESTIMATE THE KINETICS OF mRNA BIOGENESIS

Nida Javeed, Kevin Fortier, Alexander Park, Leslie Torres Ulloa, Athma A Pai

University of Massachusetts Medical School, RNA Therapeutics Institute, Worcester, MA

It is becoming increasingly clear that alternative mRNA processing is a major factor underlying human phenotypes, disease etiology, and cellular responses to external stimuli. Though we have begun to unravel the genetic factors involved in regulating alternative isoform usage, we do not yet understand the temporal dynamics of RNA processing decisions. The development of high-throughput sequencing approaches to systematically capture and analyze nascent pre-mRNA molecules at defined time-scales has opened the door for genome-wide kinetic profiling of RNA maturation. Many steps of RNA maturation – including mRNA splicing and cleavage – are thought to occur primarily co-transcriptionally, with substantial interplay between the rate of transcription elongation and choices involved in isoform variability. Thus, measuring the kinetics of transcriptional elongation, mRNA splicing, and mRNA cleavage are central to understanding the timing of initial gene regulatory decisions. Previous methods to do so have imaged single loci, studied dynamics in mutated or heavily altered genes, or used chemical treatments that affect cellular metabolisms. Here, we present three computational approaches to estimate genome-wide kinetic parameters for transcription elongation, mRNA splicing, and mRNA cleavage rates. All three approaches utilize high-throughput sequencing data from nascent mRNAs isolated after a time course of 4sU metabolic labeling to mathematically model the rate of mRNA maturation across time. Nascent RNAs are sequenced using a combination of short-read Illumina sequencing and Oxford Nanopore Minion long read sequencing in order to leverage the increased coverage afforded by short-reads and molecular completeness present in long reads. Using these methods, we estimate locus-specific PolII transcription elongation rates across human and fly transcriptomes and inter-molecular variability in PolII traveling distance. We are also able to precisely estimate intron-specific rates of mRNA splicing and polyadenylation site-specific rates of mRNA cleavage. We use our estimates to evaluate existing hypotheses that transcription slows down or pauses around critical sites involved in splicing or cleavage decisions, linking PolII dynamics to rates of splicing or cleavage at alternative sites. Our findings have many implications for the regulation of alternative exon usage, splicing efficiency, and 3’ UTR length or composition.

# SYSTEMS BIOLOGY ANALYSIS OF SINGLE-CELL TRANSCRIPTOMICS IDENTIFIES HIDDEN DRIVERS IN TUMOR MICROENVIRONMENT FOR METASTASIS OF BREAST AND LIVER CANCERS

Qingfei Pan[1], Bensheng Ju[1], Liyuan Li[2], Yanyan Wang[3], Long Nguyen[3], Hongbo Chi[3], John Easton[1], Myriam Labelle[4], Liqin Zhu[2], Jiyang Yu[1]

[1]St. Jude Children's Research Hospital, Department of Computational Biology, Memphis, TN, [2]St. Jude Children's Research Hospital, Department of Pharmaceutical Sciences, Memphis, TN, [3]St. Jude Children's Research Hospital, Department of Immunology, Memphis, TN, [4]St. Jude Children's Research Hospital, Department of Developmental Neurobiology, Memphis, TN

Metastasis accounts for ~90% of cancer deaths. Network rewiring in tumor microenvironment (TME) plays vital roles in cancer metastasis. The emergence of scRNA-seq has presented an unprecedented opportunity to dissect the TME heterogenicity and TME–tumor interactions. However, no robust computational tools are available to systemically decipher network rewiring and underpinning hidden drivers of cancer metastasis from single-cell omics data.

We have developed a systems biology algorithm, scMINER, to reverse engineer networks and expose hidden drivers from scRNA-seq omics data. Here, we used scRNA-seq to profile tumor and TME cells at early/late stages and before/after metastasis of an immunodeficient liver-2-lung and an immunocompetent breast-2-lung model. We applied scMINER to investigate the interaction and reprogramming of heterogeneous TME that facilitates lung metastasis. We profiled 67,498 tumor and TME cells in the liver-2-lung model and 52904 in the breast-2-lung model. Striking changes from early to late stages occurred to several TME cell types, and some of our identified TME reprogramming is organ-dependent. Our systems biology approaches have reconstructed intracellular gene regulatory and signaling networks specific to each of TME cell types, and have identified promising hidden drivers in reprogrammed TME cell types, which are being validated by functional assays.

Our study provides a computational framework to elucidate the dynamic TME molecular landscapes in cancer metastasis at the single-cell resolution, and lay the foundation to dissect the underlying mechanisms of TME reprogramming modulated by hidden drivers, for development of novel therapeutics to treat cancer metastasis. Our framework can be broadly applied to other complex biological problems.

# MACROMAP: A SYSTEMATIC MAP OF CONDITION-SPECIFIC GENETIC EFFECTS ON EXPRESSION IN IMMUNE RESPONSE BASED IN iPSC-DERIVED MACROPHAGES

Nikolaos I Panousis, Andrew Knights, Maria Imaz, Alice Barnett, Lorena B Vilarino, Anthi Tsingene, Celine Gomez, Daniel Gaffney

Wellcome Sanger Institute, Cellular Genetics, Hinxton, United Kingdom

A fundamental role of innate immune cells is to respond to external cues, such as pathogen challenge. Genetic differences between individuals are known to modulate these responses, and is likely to play an important role in the development of immune related diseases. However, identification and characterisation of immune-modulating variants remains challenging because many have no observable impact on gene expression in naive cells. Mapping the effects of variants across large numbers of cellular environments is extremely challenging due to the logistical problems in obtaining sufficient number of cells from a large cohort of individuals. Here, we use a high-throughput in vitro system based on a panel of human induced pluripotent stem cell-derived macrophages from 180 donors derived by the HIPSCI Consortium to study the effects of common variation on gene expression in 24 different cell states. Our stimulus panel included multiple immunologically active compounds, including agonists for a range of Toll Like Receptors (TLRs), compounds that mimic macrophage T-cell cross talk, chronic and acute inflammatory and metabolic stimuli, and multiple myeloid differentiation timepoints. We mapped hundreds of previously hidden response expression quantitative trait loci (eQTLs) and use these to identify candidate causal genes for a wide range of diseases. We observed a high degree of sharing of genetic effects (86%) across conditions and that “single condition” eQTLs are rare. We also identified cases where a broad spectrum of condition specific eQTLs are shared for only a handful of conditions based on stimulus or time. Finally, we show how disease risk alleles clustered in particular response pathways and examine how variant effects propagate over time during cell activation. Our study highlights for the true diversity of condition-specific effects on expression across a broad range of states for a single cell type and provides a rich resource for dissecting the functions of common immune disease risk alleles.

# HOST-UNMAPPED NGS READS SHAPE THE PREVALENCE OF MICROBIAL CONTAMINATION

Sung-Joon Park, Kenta Nakai

The University of Tokyo, The Institute of Medical Science, Human Genome Center, Tokyo, Japan

Contemporary biology routinely manipulates cell resources with various techniques under a range of conditions. During the course of cell manipulations, eukaryotic cells are potentially exposed to microorganisms that cause prominent morphological and physiological changes in host cells, which often yield erroneous conclusions. Despite a community-wide effort to take precautions, recent studies have pointed the prevalence of unexpected microbial contaminants in publications. This quality disruption is due in part to the intrinsic difficulty of contamination assaying, e.g., window period, primer dependency and drug resistance. As an alternative solution, next-generation sequencing (NGS) has successfully demonstrated its applicability, and NGS-based computational approaches offer promising diagnostics to assess the presence of contaminants. Since biological resources contaminated by multiple microorganisms are not uncommon, the approaches require careful attention to intra- and interspecies sequence similarities for preventing misinterpretations, which has not been fully addressed yet.

Here, we tackle this issue related to the detection certainty by proposing a novel method that explores the likelihood of reads mapped to a specific microbial genus and exploits the reads mapped to multiple microbial genera. Through the analysis of simulated and spike-in datasets, we achieved the robust performance compared to the typical approaches. We then applied it to 432 NGS datasets, 5% of which were detected as Mycoplasma-infected samples, to catalogue potential contaminants. As a result, we estimated that an RNA-seq assay includes 1,000−100,000 contaminant reads when one million host reads are sequenced, and that the contaminant reads consist of multiple genera with 10−10,000 reads. Cutibacterium was determined as one of predominant contaminants in public NGS data. We furthermore incorporated a joint non-negative matrix factorization to systematically infer the functional impact of contamination, and revealed the intricate alteration of host defense pathways, involving tumour necrosis factor receptors and humanin-like proteins, in severe Mycoplasma-infected cell lines.

Throughout this study, we have confirmed the widespread prevalence of microbial contamination in research, and deepened our understanding of the difficulty of inferring its functional impact. We present our recent effort that facilitates the profiling of contamination landscape with reduced detection uncertainty, promoting research quality assurance.

# IMPROVED GKM-SVM REGULATORY ELEMENT AND TFBS DETECTION INCORPORATING MULTIPLE TISSUE/CELL SAMPLES AND DATA TYPES

Ayoti Patra[1,2], Dustin Shigaki[1], Michael A Beer[1,2]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

Recently, several machine learning methods have been developed to classify cell-specific regulatory sequences and identify TFBS. Our gkm-SVM [1] modeling was originally trained on one data type (typically DHS or ATAC chromatin accessibility data) and only one cell or tissue sample at a time, but many regulatory elements have common patterns of accessibility across multiple ENCODE samples. We are developing techniques to identify the common patterns of epigenomic state variation across multiple samples data types (ATAC, DHS, histone-modification and TF ChIP-seq) and use gkm-SVM to detect the TFs whose modulated activity is responsible for these patterns of accessibility. We have developed clustering methods to detect these sets of co-accessible elements, and have found that using k-means or NMF techniques can detect clear regulatory signals within these elements, but with variable predictive accuracy (as assessed by AUROC). In addition, through our work in the CAGI5 regulation saturation challenge to predict enhancer and promoter activity, we have observed that training on multiple data types can significantly improve the ability of sequence based models to accurately predict the impact of mutations in MPRAs used by many ENCODE functional characterization centers. These improvements allow gkm-SVM to more accurately identify the combinations of binding sites in distal regulatory elements and promoters.

[1] Enhanced regulatory sequence prediction using gapped k-mer features. Ghandi, Lee, Mohammad-Noori, and Beer, PLOS Comp Bio 2014.

# deTS: TISSUE-SPECIFIC ENRICHMENT ANALYSIS TO DECODE TISSUE SPECIFICITY

Guangsheng Pei[1], Yulin Dai[1], Zhongming Zhao[1,2,3], Peilin Jia[1]

[1]The University of Texas Health Science Center at Houston, Center for Precision Health, School of Biomedical Informatics, Houston, TX, [2]The University of Texas Health Science Center at Houston, School of Public Health, Human Genetics Center, Houston, TX, [3]Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN

Genome-wide association studies (GWAS) and next-generation sequencing technologies have identified hundreds of thousands of disease-associated variants and genes. Unfortunately, there are many diseases where the causal tissues or cell types remain unknown. Interpretation of these variants, however, remains an open challenge. With the recent success of the Genotype-Tissue Expression (GTEx), systematically investigate tissue-specific gene expression and regulation has been made available. Leveraging these non-disease, reference data could open new avenues to infer causal tissues for diseases and to unveil underlying biological mechanisms. On the other hand, in most biomedical transcriptome data studies, heterogeneous tissues are often collected, which reduce the power in the identification of disease-associated variants and gene expression profiles. Here, we present a convenient R package, deTS: Tissue-Specific Enrichment Analysis, to identify the most relevant tissues for candidate genes or for gene expression profiles. deTS builds on two pre-processed reference panels: the GTEx data and the ENCyclopedia Of DNA Elements (ENCODE) data. Based on both panels, we first developed a statistical method based on t-statistics or z-score to identifying TSGs while controlling potential confounding factors, then we implemented two statistic tests for different forms of query data (gene list or gene expression profiles data). To validate the reliability, we conducted a across validation between GTEx and ENCODE panels, and results showed high concordance (91.5–93.2%). In addition, we demonstrated deTS using 26 traits GWAS data and 14 cancer RNA-seq data. These applications showed that deTS could effectively identify the most relevant tissues for each query trait or sample, which can not only be to identify novel relationships between gene expression and phenotypes, but also helpful to determine the tissue purity and composition in a gene expression dataset.

# THE LANDSCAPE OF INTERACTIONS BETWEEN GERMLINE AND SOMATIC VARIANTS IN CANCER.

Bernard Pereira, Catarina D Campbell, Wendy Winckler

Novartis Institutes for Biomedical Research, Oncology, Cambridge, MA

Somatic alterations in the cancer genome, most of which are associated with changes in gene expression, have been characterized in multiple studies across diverse cancer types. Similarly, large-scale association studies have uncovered associations between inherited variants and disease risk. However, less is known about germline variants that influence tumor biology by modulating expression of driver genes. Furthermore, the role of common polymorphisms in shaping the tumor microenvironment (TME) is not well understood.

To address these questions, we have performed cis-eQTL analyses to identify germline variants that influence the expression levels of 833 established cancer driver genes and 933 relevant to TME composition. We used germline polymorphism, gene expression, and somatic mutation data from over 5,500 samples spanning 19 cancer types from The Cancer Genome Atlas project. To identify germline variants associated with gene expression, we used a model that accounted for somatic mutations and copy number alterations.

We identified 602 eQTL-associtated genes (eGenes) across the combined dataset, ranging from 37 eGenes for esophageal carcinoma to 317 for thyroid carcinoma (THCA). Whereas there was an association between sample size and detected eGenes, higher numbers of eGenes were observed for THCA and prostate adenocarcinoma even after performing downsampling. Across cancers, 345 eGenes were shared between at least two cancer types, with the largest number of unique eGenes identified in THCA (77/317). Differences in the eGene distributions for ER-positive and ER-negative breast cancer were also observed. Finally, filtering by effect size and presence in GTEx revealed an enrichment of tumor-specific eVariants in TME-related genes, including *ICOSLG*, *ERAP2*, *HLA-DQA2* and *GBP2*. These genes are involved in tissue inflammation, and may influence the effects of cancer immunotherapy.

Our results highlight the extent to which germline variants can influence tumor biology and the TME. Moreover, we can now integrate these results with functional data including in vitro drug sensitivity data from projects such as the Cancer Cell Line Encyclopedia. This integrative approach has the potential to reveal prognostic and predictive germline genetic biomarkers for cancer therapy.

# FROM GENE CO-EXPRESSION TO BRAIN CIRCUITS: PREFRONTAL GENE CO-EXPRESSION NETWORKS PREDICT DRUG RESPONSE IN PATIENTS WITH SCHIZOPHRENIA

Giulio Pergola[1,2], Pasquale Di Carlo[1,2], Andrew E Jaffe[1,3,4,5], Marco Papalino[2], Qiang Chen[1], Thomas M Hyde[1,6,7], Joel E Kleinman[1,7], Joo Heon Shin[1], Antonio Rampino[2,8], Giuseppe Blasi[2,8], Alessandro Bertolino[2,8], Daniel R Weinberger[1,7,9,10]

[1]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, [2]University of Bari Aldo Moro, Basic Medical Sciences, Neuroscience and Sense Organs, Bari, Italy, [3]Johns Hopkins Bloomberg School of Public Health, Mental Health, Baltimore, MD, [4]Johns Hopkins Bloomberg School of Public Health, Biostatistics, Baltimore, MD, [5]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [6]Johns Hopkins School of Medicine, Neurology, Baltimore, MD, [7]Johns Hopkins School of Medicine, Psychiatry and Behavioral Sciences, Baltimore, MD, [8]Azienda Ospedaliero-Universitaria Consorziale Policlinico, Institute of Psychiatry, Baltimore, MD, [9]Johns Hopkins School of Medicine, Neuroscience, Baltimore, MD, [10]Johns Hopkins School of Medicine, 9McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

**Background.** Gene co-expression networks are relevant to functional and clinical translation of schizophrenia risk genes. We hypothesized that schizophrenia risk genes converge into co-expression pathways which may be associated with gene regulation mechanisms and with response to treatment in patients with schizophrenia.
**Methods.** We identified gene co-expression networks in two prefrontal cortex post-mortem RNA sequencing datasets (N=688) and replicated them in four more datasets (N=227). We identified and replicated (p-values<.001) a single module enriched for schizophrenia risk loci (13 risk genes in 10 loci). In silico screening of potential regulators of the schizophrenia risk module via bioinformatic analyses identified two transcription factors and three miRNAs associated with the risk module. To translate post-mortem information into clinical phenotypes, we identified polymorphisms predicting co-expression and combined them to obtain an index approximating module co-expression (Polygenic Co-expression Index: PCI).
**Results.** The PCI-co-expression association was successfully replicated in two independent brain transcriptome datasets (N=131; p-values<.05). Finally, we tested the association between the PCI and short-term treatment response in two independent samples of patients with schizophrenia treated with olanzapine (N=167). The PCI was associated with treatment response in the positive symptom domain in both clinical cohorts (p-values<.05).
**Conclusions.** In summary, our findings in a large sample of human post-mortem prefrontal cortex show that co-expression of a set of genes enriched for schizophrenia risk genes is relevant to treatment response. This co-expression pathway may be co-regulated by transcription factors and miRNA associated with it.

# CHROMATIN ACCESSIBILITY DURING ADIPOCYTE DIFFERENTIATION IDENTIFIES REGULATORY ELEMENTS AT CARDIOMETABOLIC TRAIT GWAS LOCI

H J Perrin[1], K W Currin[1], C K Raulerson[1], S Vadlamudi[1], J P Davis[1], M Wabitsch[2], P Pajukanta[3], M Laakso[4], K L Mohlke[1]

[1]University of North Carolina, Genetics, Chapel Hill, NC, [2]University of Ulm, Pediatrics, Ulm, Germany, [3]University of California, Los Angeles, Genetics, Los Angeles, CA, [4]University of Eastern Finland, Medicine, Kuopio, Finland

GWAS have identified thousands of cardiometabolic loci, but most lie in noncoding regions with mechanisms that remain poorly understood. At these loci, chromatin accessibility (CA) in relevant cells and contexts can guide identification of regulatory elements and mechanisms. To identify elements that influence adipocyte differentiation, we performed replicates of ATAC-seq in SGBS human adipocytes at days 0, 4, and 14 of differentiation. On average, we identified 149,945 CA regions per sample (MACS2, FDR<5%). We evaluated differential CA between timepoints (DESeq2, log2 fold change (LFC)>1, FDR<5%) and identified regions more accessible in adipocytes after 4 (30,038 regions) or 14 (20,158 regions) days of differentiation compared to preadipocytes (day 0); conversely we identified regions more accessible in preadipocytes compared to adipocytes after 4 (30,177 regions) or 14 (29,734 regions) days of differentiation. Fewer than 3,000 regions were differentially accessible between 4 and 14 days of differentiation. Regions more accessible in adipocytes were enriched for 54 motifs including CEBP, involved in differentiation and immune response, and PPAR, involved in differentiation and metabolism, while regions more accessible in preadipocytes were enriched (HOMER P<1E-5) for 70 transcription factor motif families including AP, involved in differentiation control, and GATA, involved in inhibiting adipogenesis. To identify potential regulatory variants at cardiometabolic GWAS loci, we examined 259 loci with a colocalized adipose eQTL (FDR<5%). Of 14,938 candidate variants (r2>0.8), 370 variants at 130 loci were located in a region of differential CA. For example, at a GWAS locus for phospholipids colocalized with an adipose eQTL for SCD (P=8.4E-23), we observed an adipocyte CA region (LFC=2.72) containing rs603424. This variant exhibited stronger allelic differences in transcriptional reporter activity in 3T3-L1 differentiated adipocytes (FC=2.09, P=0.01) than in SGBS preadipocytes (FC=1.25, P=0.09). These results demonstrate that CA differences between preadipocytes and adipocytes reflect biological cell context and, although further validation is needed, that epigenomic annotation in different cell contexts can guide discovery of regulatory variants at GWAS loci. Identification of CA regions in other disease-relevant contexts could aid identification of additional regulatory mechanisms.

# CHARACTERIZATION OF ACTIVATED ION-ELECTRON TRANSFER DISSOCIATION (AI-ETD) REACTION KINETICS

Trenton M Peters-Clarke[1], Benton J Anderson[1], Jean M Lodge[2], Dain R Brademan[1], Kevin L Schauer[2,3], Michael S Westphall[2], Joshua J Coon[1,2,4,5]

[1]University of Wisconsin-Madison, Department of Chemistry, Madison, WI, [2]Genome Center of Wisconsin, Madison, WI, [3]Thermo Fisher Scientific, West Palm Beach, FL, [4]University of Wisconsin-Madison, Department of Biomolecular Chemistry, Madison, WI, [5]Morgridge Institute for Research, Madison, WI

Electron Transfer Dissociation (ETD) has been widely implemented and has become a routine method. In ETD, multiply protonated polypeptides react with small molecule radical anions, resulting in peptide backbone dissociation. For molecules of low charge density, seen often in intact proteins, gas phase structure can prevent dissociation even after an electron transfer event has occurred, a process termed ETnoD. During AI-ETD, the ETD reaction within the ion trap is concurrently irradiated with IR photons from a $CO_2$ laser, which disrupts structure, reduces ETnoD, and increases the generation of sequence informative fragment ions of both modified and unmodified intact proteins. Here we characterize the ion-ion reaction kinetics that occur during AI-ETD. Considering the pseudo-first order kinetic dependence of ETD, we allow for an excess of reagent anions. The precursor intensity to product ion signal-to-noise ratio is monitored as a function of AI-ETD reaction duration. We examine the reaction kinetics of ETD and AI-ETD to characterize whether concurrent irradiation with the IR photons changes kinetic properties of the reaction. AI-ETD is shown to increase product ion formation relative to ETD alone. Preliminary investigations indicate subtle differences in the time required to reach peak c- and z-ion formation and to achieve optimal signal-to-noise. To evaluate these differences, we have begun to collect kinetic data where ion-ion reaction time is varied for a set of precursor ions with and without concurrent photo-activation. We modulate the reagent anion AGC target while reacting the precursor with reagent anion for varied duration to identify corresponding reaction constants, *k*. The ratio of the remaining precursor population after ETD to the initial population is linearly correlated with reagent ion AGC target. Mirrored preliminary investigations with concurrent photo-activation via AI-ETD also give a linear relationship but show subtle differences in *k* values. Ultimately, we will use these data to optimize AI-ETD reactions on a Fusion Lumos instrument. This work advances understanding of AI-ETD reaction kinetics to optimize AI-ETD experimentation and spectral quality for top-down proteomics.

# SPDI: DATA MODEL FOR VARIANTS AND APPLICATIONS AT NCBI

J. Bradley Holmes, Eric Moyer, Lon Phan, Donna Maglott, Brandi Kattman

National Institutes of Health, NCBI/NLM, Bethesda, MD

Normalizing diverse representations of sequence variants is critical to the elucidation of the genetic basis of disease and biological function. NCBI has long wrestled with integrating data from multiple submitters to build databases such as dbSNP and ClinVar. Inconsistent representation of variants among variant callers, local databases, and tools results in discrepancies and duplications that complicate analysis. Current tools are not robust enough to manage variants in different formats and different reference sequence coordinates.

The SPDI (pronounced “speedy”) data model defines variants as a sequence of 4 operations: start at the boundary before the first position in the sequence S, advance P positions, delete D positions, then insert the sequence in the string I, giving the data model its name, SPDI. The SPDI model can thus be applied to both nucleotide and protein variants, but the services discussed here are limited to the nucleotide. Current services convert representations between HGVS, VCF, and SPDI and provide two forms of normalization. The first, based on the NCBI Variant Overprecision Correction Algorithm, returns a unique, normalized representation termed the “Contextual Allele” for any input. The SPDI name, with its four operations, defines exactly the reference subsequence potentially affected by the variant, even in low complexity regions such as homopolymer and dinucleotide sequence repeats. The second level of normalization depends on alignment dataset (ADS). SPDI services perform remapping (AKA lift-over) of variants from the input reference sequence to return a list of all equivalent Contextual Alleles based on the transcript or genomic sequences that were aligned. One of these contextual alleles is selected to represent all, usually, that based on the latest genomic assembly such as GRCh38 and is designated as the unique “Canonical Allele”. ADS includes alignments between non-assembly RefSeq sequences (prefixed NM, NR, NG), as well inter- and intra-assembly-associated genomic sequences (NCs, NTs, and NWs) and this allows for robust remapping and normalization of variants across sequences and assembly versions.

# MODELING GENETIC VARIANTS PREDISPOSING FOR SCHIZOPHRENIA IN HUMAN NEURONS

Olli Pietilainen[1,2], Ralda Nehme[1,2], Andrea Ganna[2,3], Mykyta Artomov[2,4], Aditi Trehan[1,2], Lindy Barrett[1,2], Aarno Palotie[2,3,4], Mark Daly[2,3,4], Kevin Eggan[1,2]

[1]Harvard University, Department of Stem Cell and Regenerative Biology, Cambridge, MA, [2]The Broad Institute, Stanley Center for Psychiatric Research, Cambridge, MA, [3]Institute for Molecular Medicine, Finland, Helsinki, Finland, [4]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA

Human induced pluripotent stem cell (iPSC) derived neurons are a promising platform for studying the biological effects of genetic variants associated with severe psychiatric disorders, such as schizophrenia. Genomic CNVs constitute an extreme class of regulatory variants imposing a high risk for psychiatric disorders. One such variant is the 22q11.2 deletion syndrome that forms a 1%, genetically identifiable sub-group of schizophrenia patients. Despite its prominence, the biological mechanisms leading to the increased risk for schizophrenia remains unknown. We hypothesized that the 22q11.2 deletion exerts its effects via transcriptional regulation of cis and trans genes. We used RNA sequencing to record the transcriptional landscape of 51 human iPSC cell lines (20 carrying the 22q11.2 deletion) during neuronal differentiation by overexpression of Ngn2 transcription factor coupled with dual-Smad inhibition. We recorded mRNA levels at three cellular stages: stem cells (Day 0), neural progenitor-like cells (Day 4), and functional neurons (Day 28) for all 51 lines in three experimental replicates (total 459 experiments). Our initial power analysis suggest that our study sample is 80% powered to detect fold-changes >1.6 in a median expressed gene. The transcriptional data revealed major genome-wide transcriptional changes between the cell stages and confirmed high reproducibility between the biological replicates. The data revealed altogether 432 differentially expressed (DE) genes in patient lines (FDR<5%) 26 of which were shared across all three cellular stages. Majority (80%) of the DE genes were located outside the deletion region. We identified one gene (CAB39L) that was up-regulated in patient lines in all three cellular stages. We used LD score regression to assess if mRNA levels of genes implicated in genetic risk for schizophrenia were enriched among the DE genes. We identified enrichment for schizophrenia heritability particularly in genes that had higher mRNA abundances ($p<0.05$) among the lines carrying the deletion ($p=4.2*10-6$). This enrichment was only present in the neurons, suggesting cell-type specific effects for the 22q11.2 deletion. To assess causality between the deletion and transcriptional changes, we generated an isogenic cell line with equivalent 3Mb 22q11.2 deletion. RNA sequencing of the isogenic lines confirmed the aggregate cis effect of the deletion in genes overlapping with the deletion ($p=6*10-61$) and with up-regulated genes ($p=2.39*10-26$). Our results suggest that the 22q11.2 deletion induces genome-wide transcriptional changes that are cell type specific. Furthermore, overlap with genes implicated in schizophrenia heritability suggest a regulatory effect of the deletion on biological pathways involved in cerebrocortical functioning.

# DETERMINING ACQUISITION OF PLURIPOTENCY WITH SINGLE-CELL TRANSCRIPTOMICS

Stefan J Pietrzak*, Khoa A Tran*, Nur Zafirah Zaidan*, Alireza Siahpirani, Sunnie Grace McCalla, Gopal Iyer, Sushmita Roy, Rupa Sridharan

University of Wisconsin-Madison, Wisconsin Institute for Discovery, Madison, WI

**Background:** Effective investigation into the mechanisms regulating the reprogramming of somatic cells to induced pluripotent stem cells (iPSCs) is marred by heterogeneity due to low efficiency (~5%) and variability in kinetics, obscuring the transcriptional dynamics essential for successful reprogramming.

**Methods:** Reprogramming experiments were performed on mouse embryonic fibroblasts (MEFs) in two different culture conditions: regular serum-based media and media enriched with a combination of signaling (kinase inhibitors) and epigenetic (ascorbic acid and H3K79 methylation inhibitor) modifiers that synergistically increase reprogramming efficiency (~35%). Single-cell transcriptomics were used to profile cell populations along the reprogramming timeline.

**Results:** Single-cell analysis revealed that downregulation of MEF-associated genes and upregulation of cell cycle, epithelial, and pluripotency-associated genes occur as independent events, a phenomenon not previously observed in bulk RNA-seq experiments. Key pluripotency genes displayed high levels of co-expression with unique gene sets, each group behaving in an independent manner. A reprogramming trajectory identified cells that failed to adhere to the appropriate transcriptional pattern and the associated genes whose mis-regulation acted as a barrier to successful reprogramming. Finally, regulatory network analysis revealed a role for each of the high-efficiency components. Signaling inhibitors were key in somatic gene downregulation while the epigenetic modifiers were important for pluripotency activation.

**Significance:** The high-efficiency system coupled with single-cell analytics shows that the steps involved in reprogramming are not incremental in nature and are largely independent events. Future studies will utilize single-cell technology to investigate facets such as chromatin accessibility and epigenetics to determine how these factors coordinately operate during reprogramming.

# STREAM: SINGLE-CELL TRAJECTORIES RECONSTRUCTION, EXPLORATION AND MAPPING OF OMICS DATA

Huidong Chen[1,2,3,4], Luca Albergante[5], Jonathan Y Hsu[1,6], Caleb A Lareau[1,7], Giosue` Lo Bosco[8], Jihong Guan[4], Shuigeng Zhou[9], Alexander N Gorban[10], Daniel E Bauer[11], Martin J Aryee[1,3,7], David M Langenau[1], Andrei Zinovyev[5], Jason D Buenrostro[7], Guo-Cheng Yuan[2,3,12], Luca Pinello[1,7]

[1]Massachusetts General Hospital Research Institute and Harvard Medical School, Molecular Pathology Unit & Cancer Center, Boston, MA, [2]Dana-Farber Cancer Institute, Biostatistics and Computational Biology, Boston, MA, [3]Harvard T.H. Chan School of Public Health, Biostatistics, Boston, MA, [4]Tongji University, Computer Science and Technology, Shanghai, China, [5]Institut Curie, PSL Research University, Paris, France, [6]Massachusetts Institute of Technology, Biological Engineering, Cambridge, MA, [7]Broad Institute of MIT and Harvard, Cambridge, MA, [8]University of Palermo, Mathematics and Computer Science, Palermo, Italy, [9]Fudan University, Shanghai Key Lab of Intelligent Information Processing, and School of Computer ScienceShanghai, China, [10]University of Leicester, Mathematics, Leicester, United Kingdom, [11]Boston Children's Hospital&Dana-Farber Cancer Institute, Division of Hematology/Oncology&Department of Pediatric Oncology, Boston, MA, [12]Harvard Stem Cell Institute, Cambridge, MA

Single-cell transcriptomic assays have enabled the de novo reconstruction of lineage differentiation trajectories, along with the characterization of cellular heterogeneity and state transitions. Several methods have been developed for reconstructing developmental trajectories from single-cell transcriptomic data, but efforts on analyzing single-cell epigenomic data and on trajectory visualization remain limited. Here we present STREAM (Single-cell Trajectories Reconstruction, Exploration And Mapping), an interactive pipeline capable of disentangling and visualizing complex branching trajectories from both single-cell transcriptomic and epigenomic data. STREAM has four innovations compared to other existing methods: 1) a novel density-level trajectory visualization useful to study subpopulation composition and cell-fate genes along branching trajectories, 2) a documented end-to-end pipeline to reconstruct trajectories from chromatin-accessibility data, 3) the first interactive database focused on single-cell trajectory visualization for several published studies,and 4) a trajectory mapping procedure to readily map new cells to precomputed structures without pooling data and recomputing trajectories. This last innovation allows facile analysis of data from genomic perturbation studies or to assign cancer cells to a normal developmental trajectory. STREAM is available as a user-friendly interactive website at http://stream.pinellolab.org/ .

# TRANSCRIPTIONAL PATHWAYS OF PARTURITION AT SINGLE-CELL RESOLUTION

R Pique-Regi[1,3], R Romero[1,2,3], A L Tarca[2,3], E Sendler[1], Y Xu[2,3], O Erez[2,3,4], F Luca[1,2], S Hassan[2,3], N Gomez-Lopez[2,3]

[1]Wayne State University, Genetics, Detroit, MI, [2]NICHD/NIH/DHHS, Perinatology Research Branch, Detroit, MI, [3]Wayne State University, Ob/Gyn, Detroit, MI, [4]Ben Gurion University of the Negev, Ob/Gyn, Beer Sheva, Israel

The placenta is a complex heterogeneous organ composed of cells of both maternal and fetal origin enabling the cooperation between these two genomes required for reproductive success. Any insult during gestation may disrupt this collaboration leading to clinical manifestations of disease. It is therefore imperative to develop means to assess the maternal-fetal dialogue in order to better understand the molecular mechanisms underlying the physiologic and pathologic conditions that can affect pregnancy. Recent microfluidic and droplet-based technological advances have enabled characterization of gene expression at single-cell resolution (scRNA-seq), including in the placenta. However, studies to date have been limited in terms of the number of cases included and placental compartments that were interrogated. Furthermore, information regarding the cell-type-specific transcriptional regulation involved in human parturition and in the pathophysiology of preterm labor is lacking.
For the first time we describe a systematic comparison of single-cell transcriptional profiles from three different placental compartments (villous tree, basal plate, and chorioamniotic membranes) delivered either by: spontaneous labor at term, spontaneous preterm labor, or by cesarean section without labor at term (25 samples total). 1) Significant differences in cell-type composition were detected between the three placental compartments, including a new decidual endothelial cell type in the chorioamniotic membranes, distinct from the decidual cell types detected in the basal plate. 2) Differential gene expression was determined for specific cellular subsets between labor at term and preterm labor, in which the largest number of differentially expressed genes was observed in maternal macrophages. 3) The magnitude of cell type-specific expression changes with labor varied significantly across compartments. For instance, the increase in maternal macrophage expression of NFKB1 in both term and preterm labor was higher in the chorioamniotic membranes than in the other placental compartments. 4) Lastly, the average expression of mRNAs highly specific to certain cell types changes with gestational age and with labor at term in maternal whole blood, suggesting that placental cellular states could be monitored by mRNA analysis in maternal circulation. This study provides an important resource of single-cell transcriptomic profiles in normal and pathologic labor that will be useful for interpreting the molecular underpinnings of obstetrical complications in the maternal and fetal compartments.

# UNDERSTANDING METHYLATION VARIATION WITHIN NATURAL LINES IN *ARABIDOPSIS THALIANA*

Rahul Pisupati[1,2], Magnus Nordborg[1]

[1]Gregor Mendel Institute, Nordborg, Vienna, Austria, [2]Institute of Population Genetics, VetMed Uni, Vienna, Austria

Epigenetics is fascinating as it can make Lamarckian adaptation possible with the inheritance of acquired characteristics. In *Arabidopsis thaliana*, a model plant system, experiments suggests DNA methylation marks are faithfully inherited over generations and they affect important traits. Genome wide association studies in natural populations suggests that most of the epigenetic variation has a genetic basis. We try to determine different sources of variation for global pattern of methylation. We are analysing intercross populations and performing coalescent simulations to determine genetic or epigenetic nature of methylation patterns.

# THE IOBIO ARCHITECTURE - HIGH-PERFORMANCE GENOMIC DATA ANALYSIS IN THE WEB BROWSER

Anders Pitman[1], Chase A Miller[1,2], Yi Qiao[1], Gabor Marth[1], Tonya DiSera[1], Alistair Ward[1,2]

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

Modern web browsers provide many benefits for interactive genomic data analysis. Browsers are installed by default on every major operating system and device, and eliminate the need for complicated software installation or upgrades. They also provide a comprehensive platform for creating rich and engaging user interfaces and visualizations. However, with these advantages come many challenges. The technologies on which web browsers are built were not originally designed for interactive applications, so usability and performance must be carefully considered when analyzing and visualizing genomic data. One of the greatest challenges is handling large data files, such as sequence alignments, that are often orders of magnitude larger than typical common media format files such as MP3 music files and videos. The size of the files makes it impractical to download all the data to the user's computer. Even if the data were downloaded, the complexity of the analysis precludes the possibility of running all of the computation in the browser.

*IOBIO* (*http://iobio.io*) is an open source suite of web applications for real-time, visual, interactive genomic data analysis. *IOBIO* is architected to circumvent or mitigate the limitations of the web platform. Most processing is done on remote servers. Since a majority of datasets are stored on servers, this architecture allows for a very high-speed connection between the computational resources and the data, which is the bottleneck for most types of analyses. A user interface presents the results to the user, and receives commands for additional tasks. *IOBIO* consists of a series of software components for coordinating this process. Stochastic downsampling is used to reduce the amount of data the web browser must retrieve from the servers. A tiered approach is implemented, starting with a small sample across the dataset to give immediate visual information to the user, followed by increasingly precise sampling with the visualization updating smoothly in real-time. Data transfers are executed via streaming, where the data is broken into small chunks that are processed immediately as they arrive, without the user having to wait for the entire dataset to download before seeing results. Together, these techniques and others result in a platform that enables the creation of fast, intuitive, visually engaging software tools for genomic analysis, exploration, and discovery. This presentation will demonstrate how these techniques enable our existing web tools to solve specific, commonly encountered data analysis and visualization tasks in genomic analysis.

# INVERSION HOTSPOTS DRIVE GREAT-APE GENOME AND GENE EVOLUTION

David Porubsky[1], Ashley D Sanders[2], Shwetha C Murali[1,3], David Gordon[1,3], Stuart Cantsilieris[4], Arvis Sulovari[1], Mario Ventura[5], Jan O Korbel[2], Tobias Marschall[6,7], Evan E Eichler[1,3]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]European Molecular Biology Laboratory, Genome Biology, Heidelberg, Germany, [3]University of Washington, Howard Hughes Medical Institute, Seattle, WA, [4]University of Melbourne, CERA Ltd, Melbourne, Australia, [5]Universita degli Studi di Bari, Dipartimento di Biologia, Bari, Italy, [6]Saarland University, Center for Bioinformatics, Saarbrucken, Germany, [7]Max Planck Institute for Informatics, Algorithms for Computational Genomics, Saarbrucken, Germany

Inversions are considered to play an important role in speciation since they have the ability to suppress recombination in heterozygous conformation and have been shown to be implicated in complex neurological disorders and cancer. They are, however, difficult to resolve with standard sequencing technologies. We performed a genome-wide analysis to discover and sequence resolve all great-ape inversions using long-read sequencing and Strand-seq—a single cell sequencing technology that tracks directionality of single-stranded DNA allowing unambiguous identification of inversions. Analyzing great ape genomes,we have detected total of 758 simple inversions (chimpanzee n=151, bonobo n=140, human n=128, gorilla n=146, orangutan n=193) and 379 inverted duplications—which represent lineage-specific duplications accompanied by a change in orientation with respect to the ancestral sequence. Using multiple orthogonal technologies, we have, to date, validated >84% of the simple inversions. This analysis expands the number of previously reported inversions by ~ 7-fold including 24 human-specific inversions of which 19 were previously unknown. Our analysis of the great ape genomes, identifies 34 inversion hotspots where breakpoints for multiple inversions cluster. Nearly 1/3rd of these regions are also sites of human inversion polymorphism. Among the simple inversions, we detect 156 protein-coding genes that are likely disrupted by inversion breakpoints. Notably, both inversion hotspots and polymorphic inversions are significantly enriched for genes specifically expressed in radial glial cells and excitatory neurons (p-value <0.001). Interestingly, our analysis shows that lineage-specific duplications occur predominantly (83%) in the inverted orientation. Thus, segmental duplications and the genes therein are 4.5 times more likely to originate as inverted duplications providing fundamental insight into their mechanistic origin. Combining Strand-seq, long-read PacBio and large-insert BAC clones, we are in the process of resolving the sequence structure of all inversions and their effect on differentially expressed genes among the African great apes.

# DETECTION OF LOCAL SPLICING DISRUPTIONS IN CANCER

Laurie Prelot, Andre Kahles, Natalie Davidson, Gunnar Rätsch

ETH Zurich, Department of Computer Science, Zurich, Switzerland

Alternative splicing (AS) generates RNA and protein diversity and is known to play a role in several diseases, including cancer, where it promotes apoptotic resistance, tumor growth and metastasis. In cancer, dysregulation can occur via deleterious mutations which modify either regulatory sequences or genes of the splicing machinery. Splicing changes have been directly measured using sequencing or were inferred through sequence-based prediction algorithms. However, often the link between variants and splicing remains unclear. Analyses in cancer have found an increased amount of AS in tumors, and have identified genes which are differentially spliced between tumor and normal samples. As such approaches rely both on classical definitions of AS events (e.g., exon skipping, intron retention, alternative 5' and 3' sites) and percent spliced-in (PSI), they lack the necessary generalisation when it comes to understanding local splicing patterns in cancer. Therefore, we aim at characterizing the role of local splicing changes in cancer at a higher level of abstraction and investigate how this is influenced by genetic sequence context.

We used RNA-sequencing data of breast cancer patients from the Cancer Genome Atlas (TCGA) to generate splicing graphs per sample and gene. We merged the sample graphs to create a joined graph per gene. Next, we quantified the connectivity of the graph at each acceptor or donor position in individual samples. Finally, we ran a differential analysis on 96 patients with available tumor and matched normal samples to identify cancer specific splicing patterns. We used a generalized linear model to assess splice changes per gene, i.e. site creation, deletion or change in site strength at each position of the graph. We also assessed differences in local pattern between donors and acceptors sites.

Together, this analysis will allow us to find genes with recurrent changes in splicing patterns from a more general point of view than previous studies. This will increase our sensitivity to identify mutations and genomic contexts that influence splicing. Furthermore, we will also gain a deeper understanding of the role of acceptor, and donor variability and AS dysregulation in cancer. In conclusion, we believe that local characterisation of splicing in cancer will allow for better stratification of patients based on refined splicing patterns. Furthermore, these local changes could be used in a therapeutic setting, as novel drug targets or as sources for neoepitopes for personalized immunotherapy.

# DIVERGENT NEURONAL DNA METHYLATION PATTERNS ACROSS HUMAN CORTICAL DEVELOPMENT: CRITICAL PERIODS AND A UNIQUE ROLE OF CPH METHYLATION

Amanda J Price[1,2], Leonardo Collado-Torres[1,3], Nikolay A Ivanov[1], Wei Xia[1], Emily E Burke[1], Jooheon Shin[1], Ran Tao[1], Liang Ma[1], Yankai Jia[1], Thomas M Hyde[1,4,5], Joel E Kleinman[1,5], Daniel R Weinberger[1,2,4,5], Andrew E Jaffe[1,2,3,7]

[1]Lieber Institute for Brain Development, Baltimore, MD, [2]Johns Hopkins School of Medicine (JHSOM), McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [3]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [4]JHSOM, Dept. of Neurology, Baltimore, MD, [5]JHSOM, Dept. of Psychiatry, Baltimore, MD, [6]JHSOM, Dept. of Neuroscience, Baltimore, MD, [7]Johns Hopkins Bloomberg School of Public Health (JHSPH), Dept. of Mental Health, Baltimore, MD, [8]JHSPH, Dept. of Biostatistics, Baltimore, MD

Assessing DNA methylation (DNAm) remodeling during neuronal development identifies changes in DNA regulation as neurons mature and is particularly interesting because neurons uniquely exhibit high levels of DNAm in the non-CG context (mCH, where H=A, T or C). Here we characterize the landscape of DNAm across the first two decades of human neocortical development in neurons and non-neurons isolated from postmortem tissue using whole-genome bisulfite sequencing. We show that the rate of DNAm change is greater during the first five years of postnatal life, a time marked by dramatic synaptic remodeling, than during the entire remaining period. We further refined global patterns of increasingly divergent neuronal mCG and mCH into six unique developmental trajectories each associated with unique biological processes. Within these trajectories, neighboring mC levels—including both mCG and mCH, independent of dinucleotide context—were highly correlated, in contrast to genome-wide patterns where mCG levels were correlated but mCH levels were not. This pattern suggests that in these regions marked by cell type-specific developmental remodeling there may be functional convergence between mCG and mCH, a unique signature of the neuronal epigenome.

We then integrated paired RNA-seq data to assess the relationship between changing DNAm and expression and found that increases in both mCG and mCH were generally associated with decreasing expression. Interestingly, we also identified direct regulation of hundreds of transcripts and their splicing events by both mCG and mCH across postnatal human cortical development, including many exclusively associated with mCH, independent of mCG levels. In addition to expanding the relationship between mCH and gene expression, these splicing-associated cytosines and developmentally dynamic DNAm regions were associated with neuropsychiatric disease risk sequence, providing insight into the active cell type and timing of dynamic epigenomic remodeling in known disease risk genes and loci.

# APPROACHING THE MINIMUM COPY NUMBER GENERATION PROBLEM IN CANCER GENOMICS

Letu Qingge[1], Conor Nodzak[1], Xinghua Shi[1], Binhai Zhu[2]

[1]University of North Carolina at Charlotte, Department of Bioinformatics and Genomics, Charlotte, NC, [2]Montana State University, Department of Computer Science, Bozeman, MT

Evolutionary process of cancer is characterized by the rapid accumulation of somatic structural variants, including segmental duplications and deletions. These genomic rearrangements lead to changes in the genetic content along a chromosome, which defined as an individual copy number profile. An improved understanding of the parsimonious route of copy number profile changes through disease progression can assist in prognosis and be useful to explain drug resistance.

In this paper, we revisit the Minimum Copy Number Generation (MCNG) problem which attempts to solve the minimum number of duplications and deletions necessary to transform a given genome, G, into another genome, H, with a target copy number profile. Specifically, we investigate the complexity surrounding different types of MCNG problems. We show here that the MCNG problem is NP-hard for generic tandem duplications. Hence, we design a new 2-step greedy method to solve the MCNG problem when arbitrary segment duplications and deletions are allowed. Our experimental results on a whole genome sequencing dataset show that our algorithm performs well to detect the minimum copy number changes acquired between primary and metastatic tumors.

# DISCORDANT MONOZYGOTIC TWINS PAIRS TO COMPLETE THE CATALOG OF GENES MUTATIONS WHICH CAUSE DISEASES.

Christophe Ambroise, Guillaume Charpentier, Jean-François Deleuze, Sylvia Franc, Michele Granier, Marion Hetzel, Paul Jabert, Francis Quétier, Guillem Rigaill, Béatrice Segurens, Marie-Victoire Sénat, Isabelle Sudre, Yves Ville

University of Evry, The JUMOGEN Consortium, Evry, France

For 2-3,000 human genes, we don't know which diseases are triggered by the mutation(s) that they carry. Pairs of Discordant Monozygotic Twins (DMZT), of which one is affected by a genetic disease while the other does not display–or not yet- the clinical signs, constitute an extraordinarily powerful source to identify the genes involved in the triggering of the disease. Indeed it is already known since 2013 that the 2 genomes of a Monozygotic Twins pair only differ by a few tens of SNPs mutations and rare CNVs, all ones under a mosaïcism status because of their post-zygotic origin.

Progresses in the sequencing of human genomes today allows rapid omics alignements of DMZT pairs of WGS, of RNA-Seq from PBMC +disease-specific target tissues/organs (when accessible), of gut metagenomes, and of epigenomes (RRBS for methylomes and ATAC-Seq for chromatin accessibility).The identification of the sequence differences, and of the genes which carry those mutations will likely be impressively facilitated in this genetic context whose genetic variability is at its minimum. The very low number of SNP mutations in a pair of DMZT makes unlikely the probability of having 2 causal genes in one pair, as well as finding multiple causal CNVs in a DMZT pair.

While the GWAS strategy, too biased by the presence of heterozygotes in the control cohort, showed its limits for the multifactorial diseases, the power of the approach by DMZT benefits from the best couple healthy control/affected patient, for the genome, the transcriptome and the epigenome. The other advantage, very important, concerns the enormous financial cost of the GWAS, the cohorts of which exceed thousands of individuals, versus couples of DMZT in restricted numbers. By using DMZT, the real handicap is to end up in a gene already known to be involved in the multifactorial disease; this end could be avoided by a previous genotyping of the twins for the set of causal genes already reported for the disease.

This project aims to take benefit of the power offered by DMZT to reduce the wanderings of diagnoses which are yet too large in the frequent multifactorial diseases and to extend the catalog of gene mutations with their corresponding diseases. Validation of this strategy will be carried out by a pilot demonstrator on a few pairs of DMZT affected by TD1, Obesity, Morbid Obesity and TD2; we shall start owing to the 20.000 twins pairs already registred by the “Fédération Jumeaux et Plus” in France.

# NATURAL POPULATIONS OF FOXES, WOLVES, AND OTTERS REVEAL THE IMPACT OF DEMOGRAPHY ON DELETERIOUS VARIATION AND THE GENETIC BASIS OF INBREEDING DEPRESSION

Jacqueline Robinson*[1,2], Annabel Beichman*[1], Caitlin Brown[1], Jannikke Räikkönen[3], Robert Wayne*[1], Kirk Lohmueller*[1]

[1]UCLA, Ecol. & Evol. Biol., Los Angeles, CA, [2]UCSF, Inst. of Hum. Genet., San Francisco, CA, [3]Swedish Museum of Natural History, Envir. Res. & Monit., Stockholm, Sweden

*Contributed equally

Deleterious mutations are ubiquitous in genomes and they affect evolution. First, I will describe how we use morphological assessment, complete genomes, and population genetic simulations to contrast the Isle Royale gray wolf with the California Channel island fox to reveal the mechanistic basis of inbreeding depression. Isle Royale wolves are recently descended from a historically large population of gray wolves, but currently have a small population size and suffer from inbreeding depression. However, California's Channel Island foxes appear healthy and have persisted for thousands of generations despite their small population sizes. We find that although island foxes show no canonical signs of inbreeding depression, long-term small population size has led to an accumulation of deleterious alleles and a dramatic loss of genetic variation. In contrast, Isle Royale wolf genomes contain large runs of homozygosity, interspersed with regions of high variation, consistent with recent and intense inbreeding since their founding less than a century ago. Furthermore, the number of deleterious alleles per genome is equivalent between inbred wolves from Isle Royale and outbred wolves from the nearby mainland. Simulations under different demographic models suggest individuals recently descended from historically large populations, like Isle Royale wolves, will carry more heterozygous deleterious variants than individuals descended from historically smaller populations, like island foxes. As such, recent inbreeding can create more homozygous recessive deleterious genotypes in Isle Royale wolves than in island foxes. More generally, our findings support a model where inbreeding depression is caused by strongly deleterious recessive alleles. Second, I will describe our work on generating a de novo genome of the southern sea otter to reconstruct its evolutionary history and detect how decreases in population size have affected deleterious variation. The 18th-20th century fur trade devastated sea otter populations and nearly resulted in the extinction of the species. We show that the southern sea otter has extremely low genetic diversity and a demographic history marked by population declines that resulted in an increase in putatively deleterious variants. In sum, we find that in each of these systems, the manner in which demography impacts deleterious variation depends on the strength of selection and the specifics of the demographic history.

# WIDESPREAD HAPLOID-BIASED GENE EXPRESSION IN MAMMALIAN SPERMATOGENESIS ASSOCIATED WITH FREQUENT SELECTIVE SWEEPS AND EVOLUTIONARY CONFLICT

Kunal Bhutani*, Katherine Stansifer*, Simina Ticau*, Lazar Bojic, Joanna Slisz, Claudia Cremers, Christian Roy, Jerry Donovan, Brian Fiske, Robin Friedman

Ohana Biosciences, Cambridge, MA

Mendel's first law dictates that alleles segregate randomly during meiosis and are distributed to offspring with equal frequency, requiring sperm to be functionally independent of their genetic payload. Developing mammalian spermatids have been believed to accomplish this by freely sharing RNA from virtually all genes through cytoplasmic bridges, equalizing allelic gene expression between sperm genotypes. Applying single cell RNA sequencing to developing spermatids, we now identify a large class of mammalian genes whose allelic expression ratio is informative of the haploid genotype, which we call genoinformative markers (GIMs). 27% of spermatid-expressed genes in mice and 42% in non-human primates are not uniformly shared, but instead show at least a 2-fold allelic expression bias towards the haploid genotype. This property of these GIMs was significantly conserved between individuals and between rodents and primates. Consistent with the interpretation of specific RNA localization resulting in incomplete sharing through cytoplasmic bridges, we observe a strong depletion of GIM transcripts from chromatoid bodies, structures involved in shuttling RNA across cytoplasmic bridges, and an enrichment for 3′ UTR motifs involved in RNA localization. If GIMs are translated and functional in the context of fertility, they would be able to violate Mendel's first law, leading to selective sweeps through a population. Indeed, we show that GIMs are localized to polysomes and enriched for signatures of positive selection, accounting for 68 recent mouse selective sweeps and 5 recent human selective sweeps. Intense selection at the sperm level risks evolutionary conflict between germline and somatic function, and GIMs show evidence of avoiding this conflict by evolving more testis-specific paralogs and isoforms than expression-matched controls. The widespread existence of GIMs suggest that selective forces acting at the level of individual mammalian sperm are much more frequent than commonly believed.

# ASSOCIATING MOLECULAR EVOLUTIONARY RATES WITH CONVERGENT METABOLIC TRAITS IN OVER 200 YEAST GENOMES

Wynn K Meyer[1], Xing-Xing Shen[2], Chris T Hittinger[3,4], Antonis Rokas[2], Nathan Clark[1]

[1]University of Pittsburgh, Computational and Systems Biology, Pittsburgh, PA, [2]Vanderbilt University, Biological Sciences, Nashville, TN, [3]University of Wisconsin, Laboratory of Genetics, Genome Ctr of Wisconsin, Wisconsin Energy Inst, J.F. Crow Inst for the Study of Evolution, Madison, WI, [4]University of Wisconsin, DOE Great Lakes Bioenergy Research Center, Madison, WI

Adaptation to different selective regimes can involve molecular mechanisms ranging from subtle changes in a few loci to widespread gain or loss of functional genes throughout the genome. Convergent evolution provides an opportunity to assess which types of molecular changes typically accompany adaptation to similar environments by generating biological replicates of these adaptive processes. Recently developed methods have enabled the association of environmental or phenotype transitions with changes in molecular evolutionary rates or with the functionality of genes, while accounting for underlying phylogenetic relationships. These methods have provided insights into several convergent traits occurring on small numbers of branches of the mammalian phylogeny. We here apply these methods to investigate the evolution of 45 metabolic traits across a phylogeny of 274 yeast experiencing repeated gains and losses of these traits across different clades. We identify widespread convergent evolutionary rate changes associated with transitions in lactose fermentation and nitrite growth phenotypes, as well as individual loci associated with both transition and presence/absence of other individual metabolic traits and with sparse principal components of the trait variation. We also characterize the evolution of these traits as primarily involving amino acid changes in functional genes, or the gain or loss of genes, or both, and query which characteristics of genes and traits predict these patterns. We use the large number of traits and species within this dataset to improve methods to correct for correlation across traits and among adjacent branches of the phylogeny. This work provides a clearer picture of the molecular mechanisms by which eukaryote genomes evolve in response to changing environmental selective pressures.

# GENOMIC ARCHITECTURE AND EVOLUTIONARY CONFLICT AS DRIVERS OF EXPRESSION PATTERNS IN A SOCIAL SUPERGENE

Carlos Martinez-Ruiz, Rodrigo Pracana, Eckart Stolle, Richard A Nichols, Yannick Wurm

Queen Mary U London, Biology, London, United Kingdom

Supergenes regions - where recombination is suppressed over large parts of the genome which can contain hundreds of genes - have recently been shown to be pervasive in defining and maintaining complex polymorphisms in a wide range of organisms. We still know little, however, about how such regions come to be or the evolutionary forces underpinning their evolution. Here, we use the young “social chromosome” supergene system of the invasive red fire ant Solenopsis invicta to understand the role of intra-genomic conflict in young supergene evolution.
In this species, two variants of a the “social chromosome”, SB and Sb, inversions. Recombination between SB and Sb is severely repressed and, since Sb homozygotes are lethal recessives, there is no effective recombination within Sb. This genetic architecture is very similar to that of sex chromosomes. The social chromosome determines whether a colony has one or multiple reproductive queens, two fundamentally different froms of social organisation.We hypothesise that the evolution of the social chromosome system has been shaped by social antagonism between the two social forms, in the same way that sexual antagonism can shape the evolution of sex chromosomes.
To test whether this is the case, we perform extensive analyses of gene expression between single- and multiple-queen colonies, and in heterozygous individuals found only in multiple-queen colonies, between SB and Sb alleles of the hundreds of genes of the supergene region. Intriguingly, we find that the majority of gene expression differences between supergene variants are driven by the genomic architecture of the social chromosome - i.e. they are due to neutral and deleterious processes. In contrast, a much smaller portion of gene expression differences are likely to be adaptive and due to evolutionary conflict between social forms. Overall, our results demonstrate that supergene evolution is a complex process due to a combination of neutral, deleterious and adaptive mechanisms.

# EARLY ORIGIN OF DEVELOPMENTAL ENHANCERS IN ANIMALS

Emily S Wong[1], Siew Z Tan[2], Gaiti Federico[1], Victoria Garside[3], Ethan Scott[1], Edwina Mcglinn[3], Bernard Degnan[1], Mathias Francois[2]

[1]University of Queensland, School of Biological Sciences, St Lucia, Australia, [2]University of Queensland, Institute for Molecular Biosciences, St Lucia, Australia, [3]Monash University, Australian Regenerative Medicine Institute, Clayton, Australia, [4]University of Queensland, School of Biomedical Sciences, St Lucia, Australia

A vast array of animal body plans has arisen from a common set of regulatory transcription factor (TF) genes that interact with cis-regulatory sequences such as promoters and enhancers. However, the origin and evolution of enhancers is difficult to trace due to their fast evolution. Here, we take five candidate enhancers in sponges, a very early diverging animal phylum, and test their function in zebrafish and mouse transgenic assays. Despite diverging over 700 MYA and poor sequence identity, the enhancers are able to drive highly cell type-specific reporter gene expression in vertebrates. Activity of sponge enhancers is not only restricted to regions of conserved gene arrangement in metazoans, but enhancers near sponge-specific genes also activate transcription in vertebrates. We found regulatory networks were largely stable within the same tissue type – the same cis-regulatory element can be inserted into mouse and fish and direct similar tissue-specific expression. By characterizing TF binding motif density in the sponge Islet enhancer, we were able to identify its orthologs in human and mouse. Despite the mammalian enhancers lacking significant similarity to the sponge enhancer, they drove very similar reporter gene expression patterns as the sponge in the zebrafish. These results reveal that the origin of deeply conserved developmental enhancers can be traced back to origin of metazoan multicellularity. Despite this deep functional conservation, orthologous enhancers have highly flexible syntax and are unrecognizable using standard methods. Our result reveals metazoans not only share common TF families but also a common grammar that can be interpreted across different phlya. Together, these results suggest that developmental and cell-type enhancers evolved early in metazoan evolution and have been maintained independently in lineages that diverged over 700 million years ago, and suggest animal enhancers are as deeply conserved as many metazoan TF families. Our results provide insights into the evolution of animal diversity by the rewiring of deeply conserved regulatory modules.

# CLUSTERS OF DIVERGENT ENHANCERS IN THE ADAPTIVE EVOLUTION OF NATURAL STICKLEBACK POPULATIONS

Stanley J Neufeld[1], Jonas Schwickert[2], Jukka-Pekka Verta[3], Muhua Wang[1], Malina John[1], Domenico Scionti[1], Felicity C Jones[1]

[1]Max Planck Society, Friedrich Miescher Laboratory, Tuebingen, Germany, [2]German Cancer Research Center, Division of Stem Cells and Cancer, Heidelberg, Germany, [3]University of Helsinki, Organismal and Evolutionary Biology Research Programme, Helsinki, Finland

The rapid adaptation of natural populations to new and changing environments is critical for their survival. Genome wide association studies frequently identify non-coding regions as drivers of adaptation, suggesting gene regulatory elements play a major role. Here, we use complementary genomic, molecular, and genetic techniques in recently diverged marine and freshwater stickleback fish to identify divergent enhancers, pinpoint causal regulatory mutations, and show that clusters of regulatory elements may underlie rapid adaptation to different environments. Chromatin profiling experiments in multiple organs reveal numerous divergent putative regulatory elements between ecotypes, with regulatory regions showing molecular signatures of selection. Allele-specific ATACseq allows us to define regulatory elements under cis-control, and we intersect these data with RNAseq to identify potential enhancer-gene relationships underlying adaptive divergence. By analyzing the genomic distribution of chromatin marks, we find clusters of divergent chromatin regions spanning the size of several genes and forming potential 'adaptive regulatory cassettes'. Using in vivo enhancer assays, we functionally dissect divergent chromatin regions on an adaptive chromosomal inversion. We find the inversion contains multiple enhancers that have divergent regulatory activity in marine and freshwater ecotypes and we identify causal mutations driving this regulatory divergence. Altogether, these data suggest that divergence in gene regulatory elements is extensive, often driven by natural selection, and that clusters of divergent enhancers may drive adaptive divergence.

# CONSERVED TRANSCRIPTIONAL REGULATORY PROGRAM BRINGS ABOUT TISSUE DIFFERENTIATION DURING DEVELOPMENT

Marina Ruiz-Romero[1], Cecilia C Klein [1,3], Sílvia Pérez-Lluch[1], Alessandra Breschi[1,4], Amaya Abad[1,5], Emilio Palumbo[1], Roderic Guigó[1,2]

[1]Centre for Genomic Regulation (CRG), Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Universitat de Barcelona, Departament de Genètica, Microbiologia i Estadística, Barcelona, Spain, [4]Stanford University, Department of Genetics, Stanford, CA, [5]Center for Applied Medical Research (CIMA), Pamplona, Spain

During development, most tissues undergo striking changes in order to develop into functional organs. All along this process, the identity of each tissue arises from the particular combination of master regulatory transcription factors that specifically control the expression of relevant genes for growth, pattern formation and differentiation. In this scenario, regulation of gene expression turns to be essential to determine cell fate and tissue specificity. Although many studies aim to decipher tissue signature through the analysis of their transcriptome profiles, most lack temporal information during development and, in consequence, many differentiation events are poorly understood.

To characterize dynamic transcriptional profiles during differentiation, we tracked down the transcriptome of committed cells throughout differentiation of eye, leg and wing of Drosophila melanogaster. Our analyses indicate that transcriptome profiles of different tissues at a specific stage of differentiation are more similar to each other, than to their own lineage in the following stage. We identified a gene regulatory program shared among tissues, which suggests a tight and coordinated regulation during differentiation. Interestingly, in other fly tissues, for instance brain, this mechanism is also conserved. To describe deeply the regulatory program underlying differentiation, we studied the dynamics of chromatin accessibility surrounding transcriptional changes, the differential expression of associated Transcription factors (TFs), as well as the diversity of splicing events in this process. Overall, we designed a gene regulatory network to model the transcriptional program of tissue specification in flies. Besides, we inspected how the orthologs of these genes behave in diverse differentiation processes in higher organisms. Our results indicate that, indeed, tissue specification in other species, such as mouse and human, could follow a similar regulatory program than in the fly. Finally, functional annotation provided new insights into tissue specialization, supporting a model in which epigenetic regulators trigger cell differentiation boosting a metabolic change in pre-determined undifferentiated cells.

In conclusion, we have seen that, although differences among tissues increase over time, a common gene regulatory network is leading the differentiation process in all tissues. Our data supports a novel vision of the Waddington's landscape model, in which paths are more similar between tissues, and common and coordinated mechanisms could be responsible for organ and cell differentiation. Our investigation may help to clarify the understanding of cell differentiation, confidently allowing cell-reprogramming technologies to get closer to future personalized therapies.

# CRISPR SINGLE-CELL SEQUENCING: TOWARD FUNCTIONAL BIOLOGY IN HIGH THROUGHPUT

Christoph Bock[1,2,3]

[1]CeMM Research Center for Molecular Medicine, Austrian Academy of Sciences, Vienna, Austria, [2]Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria, [3]Max Planck Institute for Informatics, Max Planck Society, Saarbrücken, Germany

International consortia have mapped the human transcriptome and epigenome in hundreds of cells types (ENCODE, IHEC, BLUEPRINT, etc.) and thousands of individuals (GTEx, GoDMC etc.). These maps are being refined by ongoing single-cell sequencing projects (HCA, HuBMAP, etc.), which will eventually give rise to a comprehensive catalog of all cells in the human body. However, we are lagging behind with our ability to assign biological functions to the observed gene regulatory patterns.

CRISPR-based genetic screens have the potential to accelerate functional and gene-regulatory studies, but current methods have inherent limitations. Widely used pooled screens are restricted to simple readouts including cell proliferation and sortable marker proteins. Arrayed screens allow for comprehensive molecular readouts such as transcriptome profiling, but operate at much lower throughput.

To overcome these limitations, we combined pooled CRISPR screening with single-cell RNA sequencing into a broadly applicable workflow, directly linking guide RNA expression to transcriptome responses in thousands of individual cells (Datlinger et al. 2017 Nature Methods). Our CROP-seq method enables pooled CRISPR screens with single-cell transcriptome resolution in a way that scales well in the number of cells (up to millions using combinatorial indexing) and target genes (up to genome-wide).

We and many other labs have found CROP-seq broadly useful for studying biological mechanisms that are difficult to reduce to a simple readout needed for classical pooled screens. Applied to heterogeneous cell populations and ex vivo tissue (e.g., in combination with Cas9-expressing mice), CROP-seq can identify cell-type-specific regulation in entire organs and uncover cellular differentiation hierarchies.

Given the increasing throughput of single-cell transcriptomics and the advent of single-cell multi-omics technology (reviewed in: Bock et al. 2016 Trends in Biotechnology), CROP-seq enables deep profiling of large CRISPR libraries and constitutes a powerful method for functional biology at scale.

*Funding: C.B. is supported by an ERC Starting Grant (n° 679146) and by a New Frontiers Group award of the Austrian Academy of Sciences.*

# PAIRWISE AND HIGHER-ORDER GENETIC INTERACTIONS DURING THE EVOLUTION OF A tRNA

Júlia Domingo[1,2], Guillaume Diss[1,2], Ben Lehner[1,2,3]

[1]EMBL-CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain, [2]Universitat Pompeu Fabra, (UPF), Barcelona, Spain, [3]Institució Catalana de Recerca i Estudis Avançats, (ICREA), Barcelona, Spain

A central question in genetics and evolution is the extent to which mutations have outcomes that change depending on the genetic context in which they occur. Pairwise interactions between mutations have been systematically mapped within and between genes, and contribute substantially to phenotypic variation amongst individuals. However, the extent to which genetic interactions themselves are stable or dynamic across genotypes is unclear. We quantified >45,000 genetic interactions between the same 87 pairs of mutations across >500 closely related genotypes of a yeast tRNA. Strikingly, all pairs of mutations interacted in at least 9% of genetic backgrounds and all pairs switched from interacting positively to interacting negatively in different genotypes. Switches in sign were also frequent for higher-order epistatic interactions. The pairwise and higher-order epistasis in this molecule mean that all individual mutations switch from detrimental to beneficial in even closely-related genotypes. These results add to a growing body of evidence that evolution is highly contingent at the molecular level. As a consequence, accurate genetic prediction requires models that incorporate both pairwise and higher order epistatic terms.

# ENHANCER FUNCTION OF ULTRACONSERVED ELEMENTS DOES NOT REQUIRE PERFECT SEQUENCE CONSERVATION

Valentina Snetkova[1], Quan T Pham[1], Momoe Kato[1], Stella Tran[1], Ingrid Plajzer-Frick[1], Catherine S Pickle[1], Anne N Harrington[1], Jennifer A Akiyama[1], Veena Afzal[1], Diane E Dickel[1], Axel Visel[1,2,3], Len A Pennacchio[1,2,4]

[1]Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division, Berkeley, CA, [2]U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA, [3]University of California Merced, School of Natural Sciences, Merced, CA, [4]University of California Berkeley, Comparative Biochemistry Program, Berkeley, CA

Ultraconserved elements are extended stretches of genome sequence showing perfect conservation across phylogenetically distant mammalian species. Many ultraconserved elements are noncoding and act as tissue-specific transcriptional enhancers during embryogenesis. In line with this general function, complete loss of individual ultraconserved enhancers can cause developmental phenotypes. However, the molecular constraints causing the extreme negative selection on every single base pair within these enhancers remain unknown. To explore the *in vivo* robustness of ultraconserved enhancers to mutations, we subjected 23 different ultraconserved enhancers to different levels of targeted mutagenesis. We generated 75 enhancer variants carrying a total of 2,073 mutations of ultraconserved base pairs in various combinations and tested the *in vivo* impact of each set of mutations in transgenic reporter mouse assays. The majority of ultraconserved enhancers (19/23, 83%) showed remarkable robustness to sequence changes, with up to 5% of mutated base pairs causing no apparent changes in the tissue specificity or reproducibility of enhancer reporter activity. In two cases, even enhancers in which 20% of ultraconserved base pairs had been mutated showed residual tissue-specific activity. These results indicate that negative selection associated with the known embryonic enhancer activity of these elements alone is insufficient to explain their extreme evolutionary conservation. We also explored the hypothesis that ultraconserved enhancers may be particularly susceptible to gains of function, which would cause mutations to result in aberrant expression of developmental genes. However, such gains of function resulting from point mutations were even less common than loss of function (2/23, 9%). Taken together, our results demonstrate that the embryonic enhancer function of many ultraconserved elements is preserved even at substantial degrees of mutation, supporting the possibility that additional, hitherto unrecognized, molecular functions or evolutionary dynamics contribute to their extreme degree of conservation.

# HUMAN PANCREATIC ISLET 3D CHROMATIN ARCHITECTURE PROVIDES INSIGHTS INTO THE GENETICS OF TYPE 2 DIABETES

Inês Cebola*[1], Irene Miguel-Escalada*[2], Silvia Bonàs-Guarch*[2], Goutham Atla[2], Julen Mendieta-Esteban[3], Biola M Javierre[4], Irene Farabella[3], Marc A Marti-Renom[3], Peter Fraser[5], Jorge Ferrer[1,2]

*Contributed equally

[1]Imperial College London, Department of Medicine, London, United Kingdom, [2]Centre for Genomic Regulation (CRG), Bioinformatics and Genomics, Barcelona, Spain, [3]CNAG-CRG, Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, [4]Josep Carreras Leukaemia Research Institute, 3D chromatin organization, Badalona, Spain, [5]Florida State University, Department of Biological Science, Tallahassee, FL

Genome-wide association studies (GWAS) have identified over 200 loci that carry genetic variants associated with Type 2 Diabetes (T2D) susceptibility. However, most of the identified variants reside within non-coding sequences, which makes their interpretation difficult. As such, the molecular mechanisms underlying most T2D genetic association signals remain elusive.

Previously, we observed that there is an enrichment of T2D GWAS variants in super-enhancers that are selectively active in pancreatic islets, the tissue in which the insulin-producing beta cells reside. It is thus likely that some T2D susceptibility SNPs act by affecting islet regulatory activity rather than by altering protein-coding sequence, but the identification of the target genes of such regulatory variants is challenging.

In this study, we have used promoter-capture Hi-C (pcHi-C) to generate an interactome map for human pancreatic islets, identifying all promoter-anchored 3D chromatin interactions. This allowed us to identify the target genes of >80% of all human pancreatic islet enhancers and revealed >1000 groups of islet enhancers, super-enhancers and active promoters that form tissue-specific, higher-order 3D hubs - *enhancer hubs*.

Using CRISPR/Cas9 in pancreatic beta cells, we validated enhancer-promoter assignments of >10 T2D-associated enhancers, including assignments to genes located hundreds of kilobases away, and multiple genes within enhancer hubs that were affected by perturbation of single enhancers.

These results indicate that human islet 3D chromatin architecture provides a framework for interpretation of T2D GWAS signals. The combination of pcHi-C with CRISPR/Cas9 validations shed light into unexpected regulatory links that may be affected by T2D susceptibility variants, and brought to our attention potential new players in T2D aetiology.

PRECISE EDITING AT DNA REPLICATION FORKS ENABLES MULTIPLEX GENOME ENGINEERING IN EUKARYOTES

Farren Isaacs

Yale University, Dept. of Molecular, Cellular and Developmental Biology, Systems Biology Institute, New Haven, CT

I will describe a multiplex genome engineering technology in *Saccharomyces cerevisiae* based on annealing of synthetic oligonucleotides at the lagging strand of DNA replication. The mechanism is independent of Rad51-directed homologous recombination and avoids the creation of double-strand DNA breaks, enabling precise chromosome modifications at single base-pair resolution with efficiencies >40% without unintended mutagenic changes at the targeted genetic loci. We observed the simultaneous incorporation of up to 12 oligonucleotides with as many as 60 targeted mutations in one transformation. Iterative transformations of a complex pool of oligonucleotides rapidly produced large combinatorial genomic diversity $>10^5$. This method was used to diversify a heterologous β-carotene biosynthetic pathway that produced genetic variants with precise mutations in promoters, genes, and terminators, leading to altered carotenoid levels. Our approach of engineering the conserved processes of DNA replication, repair, and recombination could be automated and establishes a general strategy for multiplex combinatorial genome engineering in eukaryotes.

# MITOCHONDRIAL EVOLUTION AMONG DIVERSE YEAST SPECIES AND SYNTHETIC MITOCHONDRIAL GENOMICS TOOLS FOR FUNCTIONAL CHARACTERIZATION

John F Wolters[1], Dana A Opulente[1], Jacek Kominek[1], Abbe L LaBella[2], Xing-Xing Shen[2], Xiaofan Zhou[2,3], Clete Kurtzman[4], Julie Trolle[5], Sudarshan Pinglay[5], Jef D Boeke[5], Antonis Rokas[2], Chris T Hittinger[1]

[1]UW-Madison, Laboratory of Genetics, Madison, WI, [2]Vanderbilt University, Department of Biological Sciences, Nashville, TN, [3]South China Agricultural University, Integrative Microbiology Research Center, Guangzhou, China, [4]USDA Agricultural Research Service, Applied Microbiology Research Unit, Peoria, IL, [5]New York University, Langone Health, New York, NY

Mitochondrial DNA (mtDNA) variation has functional effects in diverse conditions even in organisms for which respiration is not mandatory. These consequences range from impacts on unicellular growth, organismal phenotypes such as longevity, and complex diseases including cancer. The mechanisms underlying this functional variation and the evolutionary patterns shaping mtDNA variation broadly are poorly understood. Studying functional mitochondrial evolution requires diverse models which are amenable for laboratory manipulation. We suggest that the budding yeast subphylum Saccharomycotina provides an ideal group. Studying mitochondrial evolution in budding yeasts benefits from combining the powerful mitochondrial genetics tools developed in Saccharomyces cerevisiae with the evolutionary genomics approaches made possible by comparing species spanning over 400 million years of evolution. Yeast mtDNAs deviate from animal mtDNA in several ways including larger genomes, introns, and recombination between divergent haplotypes. Lack of mtDNA resources for the vast majority of these species has hindered comparative study. Here, we present new sequences that more than double the number of existing complete mtDNAs for budding yeast species. We test two major hypotheses regarding mitochondrial evolution: do yeast species expanding into novel temperature regimes exhibit corresponding signatures of positive selection on nuclear and mitochondrial respiratory genes, and do yeast species that can grow without respiration have reduced purifying selection on these genes. To functionally test the significance of mtDNA variants we are developing novel methods to transform S. cerevisiae with synthetic mtDNAs. The flexibility offered by synthetic mtDNA will ease the verification of functional variants, allow elimination of confounding factors such as introns that cause cytonuclear incompatibilities, and allow us to test novel hypotheses regarding the impacts of genome architecture on mitochondrial function.

# CRISPRESSO2: ALLELE-SPECIFIC QUANTIFICATION AND VISUALIZATION OF GENOME EDITING EVENTS FROM CRISPR/CAS9, CAS12A (CPF1) AND BASE EDITORS.

Kendell Clement[1,2,3], Holly Rees[4,5,6], Matthew Canver[1,2,3], Jason M Gehrke[2,3], Rick Farouni[1,2,3], Jonathan Y Hsu[1,2,3], Mitchel A Cole[7], David R Liu[4,5,6], Keith Joung[2,3], Daniel E Bauer[1,7,8,9], Luca Pinello[1,2,3]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]Massachusetts General Hospital, Molecular Pathology Unit, Boston, MA, [3]Harvard Medical School, Department of Pathology, Boston, MA, [4]Broad Institute of MIT and Harvard, Merkin Institute of Transformative Technologies in Healthcare, Cambridge, MA, [5]Howard Hughes Medical Institute, Harvard University, Cambridge, MA, [6]Harvard University, Department of Chemistry and Chemical Biology, Cambridge, MA, [7]Boston Children's Hospital, Division of Hematology/Oncology, Boston, MA, [8]Harvard Medical School, Department of Pediatrics, Boston, MA, [9]Harvard Stem Cell Institute, Cambridge, MA

The field of genome editing is rapidly advancing, and technologies to modify the genome are becoming increasingly more accurate, efficient and versatile. These technologies are accelerating discovery and interpretation of functional genetic variants from GWA and eQTL studies by enabling targeted perturbations, even of a single base pair. Amplicon sequencing is the gold standard for studying editing outcomes, but the technical expertise required to analyze amplicon sequencing data is a critical barrier to correct interpretation of editing outcomes. In addition, current software toolkits are not designed for modern genome editing tools such as base editors which are capable of inducing single-nucleotide changes. Here, we introduce CRISPResso2, a comprehensive analysis toolkit that introduces five key innovations for the analysis of genome editing data: (1) Comprehensive analysis of sequencing data from base editors; (2) Allele specific quantification of heterozygous reference sequences; (3) A novel biologically-informed alignment algorithm; (4) Ultra-fast processing time; and (5) Batch mode for analyzing and comparing multiple editing experiments.

We generated experimental data for six modern genome editing agents: Cas9, four C→T base editor variants, and one A→G base editor using guides targeting five well-studied genomic loci. To demonstrate the utility of CRISPResso2, we analyzed this dataset and provide an in-depth characterization of the editing properties of these agents. Using the intuitive graphics produced by analyzing multiple samples simultaneously, the editing characteristics of each agent can be distinguished and compared. When comparing the four C→T base editor variants, our plots specifically designed for base editors highlight shifted activity windows and increased editing purity resulting from modification to the deaminase enzyme. We also used CRISPResso2 to recover alleles important for fetal hemoglobin induction by analyzing targeted perturbation of key enhancer elements harboring trait-associated variants.

CRISPResso2 enables accurate and intuitive analysis of coding and noncoding elements in CRISPR based perturbation studies and is available online at: http://crispresso2.pinellolab.org.

# BIG BRAINS: HIGH-THROUGHPUT ENHANCER DELETIONS AND THE GENETICS OF BECOMING HUMAN

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, Department of Ecology and Evolutionary Biology, New Haven, CT

Gene regulatory elements active during early human cortical development are thought to contribute to the expansion of the neocortex. We hypothesized that identification of regulatory elements which alter human neural stem cell (hNSC) proliferation will provide insight into the genetics of cortical expansion in humans. Here we implement a non-coding sgRNA-Cas9 screening method to interrogate more than 25,000 individual binding sites across thousands of embryonic cortical enhancers and more than 10,500 genes. We utilized active sgRNA-Cas9 to 'delete' conserved elements within enhancers and to knockout protein-coding genes. Proliferation phenotypes were modeled across two replicates and a simple machine-learning algorithm (linear-discriminant analysis) was used to classify each enhancer or gene deletion into three categories: decreased proliferation, neutral, or increased proliferation.

We identified 1,270 unique deletions within enhancers and 2,266 gene knockouts that alter neural stem cell proliferation. By 12 cell divisions, relatively few sites within enhancers decrease proliferation (3.9%) and even fewer increase proliferation (0.2%), suggesting that disruption of most binding sites in enhancers does not substantially alter hNSC proliferation. Intriguingly, binding sites that decrease proliferation are enriched for specific transcription-factor motifs, including E2F4, SP1, and KLF5, providing evidence that critical sites function as downstream effectors of cell cycle regulation, recruit factors responsible for local chromatin remodeling, and respond to developmental cues through growth factor binding.

We identified 16 deletions in human-accelerated regions and 47 deletions in human-gain enhancers that alter hNSC proliferation. Our findings provide direct evidence that these regulatory elements impact hNSC biology and suggests a role in human neurogenesis. The list of regulatory elements we provide can be utilized to explore human-specific aspects of these loci during cortical development. As expected, we identified many gene knockouts that decrease proliferation (22.8%) and a smaller set of gene knockouts that increase proliferation (<0.1%). Genes that decrease proliferation are enriched for essential biological processes including chromosome segregation and mRNA processing. Proliferation decreasing genes are notably enriched for ASD and developmental disorder risk genes, providing insight into how loss-of-function mutations in these genes may contribute to disease risk.

Our results provide quantitative maps of regulatory elements required hNSC proliferation and reveal novel insights into the critical sites underlying enhancer function during evolution and development. We expect these findings to inform future research on human-specific biology and to aid the interpretation of non-coding genetic variation linked to developmental disorders.

# ANALYSING THE JOINT EFFECT OF COMMON AND RARE VARIANTS ON HUMAN TRAITS, USING WHOLE-GENOME SEQUENCING DATA AND A GENE-BASED APPROACH

Nima Rafati[1], Marcin Kierczak[2], Julia Höglund[1], Mathias Rask-Andersen[1], Weronica Ek[1], Diana Ekman[3], Torgny Karlsson[1], Åsa Johansson[1]

[1]Science for Life Laboratory, Uppsala University, Immunology, Genetics and Pathology, Uppsala, Sweden, [2]National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Cell and Molecular Biology, Uppsala, Sweden, [3]National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Biochemistry and Biophysics, Stockholm, Sweden

Despite the success of genome-wide association studies (GWAS) in revealing the genetic architecture of various traits and diseases, much of the genetic contribution to complex diseases and phenotypes remains unexplained. Characterising rare variants can unravel additional effects that explain phenotypic variability and disease risk. Whole genome sequencing (WGS) of large cohorts selected from isolated population provides an opportunity to study the combined effects of common and rare variants. In this study, we have generated high coverage (30 - 55x) WGS data from 1,021 Swedish individuals and measured expression of more than 400 disease-related biomarkers. We developed a pipeline for exploring the combined effects of common and rare single nucleotide variants (SNVs) and short insertions/deletions (INDELs). The pipeline was used to analyse the combined effect of genetic variants in both coding and regulatory regions, on protein biomarker levels in a gene-wise manner. We scored SNVs and INDELs in coding regions for estimated deleteriousness based on Combined Annotation Dependent Depletion (CADD) scores and SNVs in regulatory regions based on Eigen scores. For coding variants, we identified 158 independent signals, of which 52 were in trans. Among these hits, we identified 91 novel genes to be associated with 81 biomarkers, which were not detected using a standard GWAS with genotyped/imputed SNPs in the same cohort. For regulatory regions, only cis-regulatory elements were analysed and for as many as 191 (44% of the biomarkers) we identified at least one regulatory element to be associated with the expression of the biomarker. We also compared our results with another region-based method, MONSTER, which is an extension of the well-established gene-based test, SKAT. In this study, we have shown that combination of using WGS data and a gene-based approach that evaluates the combined effect of rare and common variants, increases the power in detecting novel genetic effects.

# INTEGRATIVE GENOMICS IDENTIFIES A CONVERGENT MOLECULAR SUBTYPE THAT LINKS EPIGENOMIC WITH TRANSCRIPTOMIC DIFFERENCES IN AUTISM

Gokul Ramaswami, Daniel H Geschwind

UCLA, Neurology, Los Angeles, CA

Autism spectrum disorder (ASD) is a phenotypically and genetically heterogeneous neurodevelopmental disorder. Despite this vast heterogeneity, previous studies have shown patterns of molecular convergence at the epigenomic and transcriptomic level in post-mortem brain tissue from autistic subjects. However, to date these molecular datasets have only been analyzed independently. There is no clear understanding of how epigenetic changes directly regulate expression of their cognate genes and the order in which these processes occur. In this study, we integrate genome-wide measures of mRNA expression, miRNA expression, DNA methylation, and histone acetylation from ASD and control brains to identify a convergent molecular subtype of ASD with shared dysregulation across both the epigenome and transcriptome. By focusing on this more homogeneous subtype, we are able to increase power over previous efforts, identifying nearly 3-fold more (n = 4599) differentially expressed genes. Consistent with previous studies, we find an upregulation of genes with gene ontology annotations of immune response and a downregulation of genes involved in synaptic transmission and neuronal ion transport. Additionally, we identify 4187 and 2415 differentially methylated promoters and gene bodies, respectively. A few hundred hypomethylated promoters and gene bodies are localized to immune genes exhibiting upregulated expression. But, overall, differential methylation is not a prominent driver of differential gene expression. We identify 6894 differentially acetylated regions across the genome. In contrast to methylation, differential acetylation at gene promoters is strongly correlated with differential expression ($R=0.4$; $p=3x10^{-21}$). We utilize eQTL and chromosome conformation datasets from adult brain tissue to link noncoding differentially acetylated regions with their cognate genes and identify numerous hyperacetylated regions regulating critical hub genes of co-expression modules. Furthermore, using stratified LD score regression we find an enrichment of ASD genetic risk variants in hyperacetylated noncoding regulatory regions linked to neuronal genes, providing a potential mechanistic understanding of how non-coding ASD risk variants impact gene regulation. Overall, the findings in this study help elucidate how diverse genetic risk factors converge onto specific molecular processes in ASD.

# CANDIDATE CANCER DRIVER MUTATIONS IN DISTAL REGULATORY ELEMENTS AND LONG-RANGE CHROMATIN INTERACTION NETWORKS

Helen Zhu*[1,2], Liis Uuskula-Reimand*[3,4], Keren Isaev*[1,2], Lina Wadi[1], Azad Alizada[3], Shimin Shuai[1,5], Vincent Huang[1], Dike Aduluso-Nwaobasi[1], Marta Paczkowska[1], Diala Abd-Rabbo[1], Oliver Ocsenas[1,2], Minggao Liang[3,5], Jared T Simpson[1,7], Mathieu Lupien[2,6], Lincoln D Stein[1,5], Paul C Boutros[1,2], Michael D Wilson[3,5], Juri Reimand[1,2]

[1]Ontario Institute for Cancer Research, Computational Biology Program, Toronto, Canada, [2]University of Toronto, Department of Medical Biophysics, Toronto, Canada, [3]SickKids Research Institute, Program in Genetics and Genome Biology, Toronto, Canada, [4]Tallinn University of Technology, Department of Gene Technology, Tallinn, Estonia, [5]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [6]Princess Margaret Cancer Centre, Toronto, Canada, [7]University of Toronto, Department of Computer Science, Toronto, Canada

A comprehensive catalogue of driver mutations is essential for understanding tumor biology and treatments. Driver mutations are positively selected in tumor genomes and thus appear statistically over-represented in cancer genes across large patient cohorts. Exome-sequencing studies have mapped many protein-coding drivers, yet few non-coding drivers are known because genome-wide discovery is challenging and limited in statistical power. To fill this gap, we analyzed 150,000 cis-regulatory regions (CRMs) in 1,844 whole cancer genomes from the largest uniformly processed whole genome sequencing dataset compiled by the ICGC-TCGA Pan-cancer Analysis of Whole Genomes (PCAWG) project. We developed a new driver discovery method, ActiveDriverWGS, validated it on protein-coding genome and found 48 genes of which >90% are known cancer drivers. We then applied the model to cis-regulatory regions (CRMs) from ENCODE and found 30 candidate drivers enriched in non-coding SNVs and indels. These frequently mutated regulatory elements (FMREs) were also enriched in genomic rearrangements, pan-tissue enhancers, long-range chromatin interactions and motif-rewiring mutations, suggesting that these non-coding regions and mutations control transcription via three-dimensional genome organization. The network of FMREs, chromatin loops and target genes revealed a subset of candidate drivers where the mutations associated with differential mRNA abundance of known and novel cancer genes in matched patient transcriptomes. Pathway analysis of CRMs revealed many sub-significant mutated FMREs at known cancer genes, indicating that a further landscape of infrequent non-coding driver mutations remains to be characterized in future well-powered datasets.

# PSYCHOSOCIAL EXPERIENCES AND THEIR GENETIC INTERACTIONS ARE ASSOCIATED WITH TRANSCRIPTIONAL DYSREGULATION IN ASTHMA

Justyna A Resztak[1], Allison K Farrell[2], Henriette E Mair-Meijers[1], Adnan Alazizi[1], Samuele Zilioli[2], Roger Pique-Regi[1], Richard B Slatcher[2], Francesca Luca[1]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]Wayne State University, Department of Psychology, Detroit, MI

The psychosocial environment has major effects on human health. Intrapersonal experiences like stress and emotion, interpersonal environments like families, and broader structural environments like neighborhoods and socioeconomic status (SES) are all known to affect physical health and disease outcome. People living in impoverished cities, with high exposure to pollutants and negative psychosocial experiences, have high prevalence and mortality rates from asthma. Allergic asthma is a chronic disease characterized by inflammation of the airways and hyperreactivity to allergens. Genetic studies have identified transcriptional dysregulation in immune cells as an important mechanism for the genetic risk of asthma. Additionally, psychosocial stress is known to affect gene expression, but the specific biological pathways and mechanisms linking psychosocial experiences and asthma are poorly defined.

To identify transcriptional signatures of psychosocial experiences that may affect asthma symptoms, we analyzed 23 psychosocial variables and 16 asthma traits in leukocytes from an ethnically-diverse sample of 119 youth from Detroit, MI. We identified significant transcriptional signatures of 9 psychosocial variables and 4 asthma traits. Transcriptional signatures of forced expiratory volume (FEV) and nightly asthma symptoms were highly correlated with those of socio-economic status, percent unoccupied houses in the neighborhood and self-disclosure. Gene co-expression analysis identified 6 modules significantly associated with these traits (10% FDR), including a module that is enriched for genes in the type I interferon signaling pathway. Over 30 genes in this module are differentially expressed for nightly asthma symptoms and self-disclosure (10% FDR), and this transcriptional response cannot be explained by inter-individual variation in blood cell composition. Using the transcriptional signatures of psychosocial variables as environmental proxies, we identified 28 interaction eQTLs, thus demonstrating that negative psychosocial experiences in humans contribute to interindividual variation in gene expression.

These results demonstrate that social genomics approaches in humans can uncover potential molecular mechanisms underlying health disparities.

# DEVELOPMENTAL ROLES OF FRIZZLED RECEPTORS IN BUTTERFLY EMBRYOS

Teomie S Rivera-Miranda[1], Joseph Hanly[2], Jose J Hermina-Perez[2], Christopher R Day[2], Riccardo Papa[1], Arnaud Martin[2]

[1]University of Puerto Rico Rio Piedras, Biology, San Juan, PR, [2]George Washington University, Biology, Washington DC, DC

Frizzled (fz) genes encode receptors of the Wnt signaling pathway and play important developmental roles, such as head and leg formation, in all animals. In insects, both frizzled receptors and Wnt ligands have been shown to pattern embryos, and the Wnt pathway is also associated with the formation of color patterns at various life stages. The roles of fz1, fz2, and fz4 have been specified for Tribolium, but no specific function for fz3 has been described (Beerman et al., 2011). We hypothesized, that fz1, fz2, and fz4 in Vanessa cardui butterflies will share function and expression domains with Tribolium at the embryo stage. To address this, we used CRISPR somatic mutagenesis to test the roles of four frizzled receptors fz1-fz4, in developing embryos (26 hours old) and 5th instar larval wing disks and compare their expression patterns using in situ hybridizations (ISH) of wild-type individuals. In situ hybridization results indicated that fz1 is ubiquitously expressed in the embryos, fz2 is segmentally expressed at 26 hours, fz3 seemed weakly expressed on 26 hours embryos but showed chevron-like expressions on the wing disks. CRISPR knockouts of the four fz receptors revealed interesting results even though fz1 and fz3 did not show any apparent effects on embryonic development. fz2 knockout effects ranged from head and mouthpart deformation to completely impeding embryo development. On the other hand, fz4 knockouts interfered with head and limb development. ISH results observed for fz1 coincide with what has been seen in Tribolium and fz4 results are still not clear for both embryos and 5th instar wing disks. As for the knockouts, there are some similarities to other insects' frizzled functions but there were major differences observed, such as the effect of fz4 knockout in V. cardui compared to RNAi results observed in Tribolium. To validate these results, further work will be done to study these four genes on different embryonic stages to see if there is a relation between time and where the gene is expressed. We would also like to work with other butterfly species to compare results, and test the effects of knocking out different combinations of the fz genes: what would happen if we knock out a combination of fz1+fz2 at different time-points? The methods used in this study have been proven to be both important and efficient in understanding gene involvement in developing organisms (not only insects), and by further studying genes, such as frizzled, we will achieve a deeper understanding of their functions and importance in other organisms.

# GENETIC ASSOCIATION ANALYSIS OF TYPE 1 DIABETES IN > 60,000 ANCESTRALLY DIVERSE SUBJECTS REFINES ASSOCIATION SIGNALS AND IDENTIFIES NOVEL RISK LOCI

Catherine C Robertson*[1], Jamie R Inshaw*[2], Suna Onengut-Gumuscu[1], Wei-Min Chen[1], John A Todd*[2], Stephen S Rich*[1]

[1]University of Virginia, Center for Public Health Genomics, Charlottesville, VA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Oxford, United Kingdom

*Authors contributed equally

**Background:** Type 1 diabetes (T1D) is a complex disease driven by diverse genetic and environmental factors. Over 60 chromosome regions are implicated in T1D, yet ~20% of genetic risk in Europeans remains unexplained by known T1D-associated regions and little is known about genetic risk in non-European populations. Causal variants and genes are not established at most T1D-associated loci. To address these gaps, we analyzed an ancestrally diverse cohort twice the size of previous studies with ImmunoChip genotyping and imputation using a large multi-ethnic reference panel.
**Methods:** DNA samples were genotyped on the Illumina ImmunoChip. Genotypes were imputed with the NHLBI Trans-Omics for Precision Medicine (TOPMed) reference panel. Imputed variants were filtered for ancestry-specific imputation quality and Mendelian inconsistencies. Accuracy and coverage of imputation in 184 regions densely covered by the ImmunoChip was assessed using whole genome sequencing available in a subset of participants. Association with T1D was assessed in affected trios by the transmission disequilibrium test (TDT). Case-control analysis was performed in unrelated individuals by logistic regression, adjusting for principal components. Summary statistics were combined using inverse variance-weighted fixed effects meta-analysis. A Bayesian stochastic search method (GUESSFM) was applied to the European case-control data to define 95% credible sets within each region.
**Results:** After quality filtering, 61,104 participants and 164,603 genotyped variants were included in analyses. Genotype imputation provided 2- to 3-fold increase in coverage in regions of interest. In fixed effects meta-analysis, 70 loci were associated with T1D at genome-wide significance ($p<5x10-8$), twenty-one (21) regions were novel for association with T1D including 1q21.3 (*IL6R,nearest gene and/or plausible candidate causal gene*), 5p13.1 (*TTC33*), and 1p31.3 (*PGM1*). We defined 95% credible variant sets for 52 associated regions with sufficiently dense genotype coverage. Seven credible sets contained a single variant. An additional sixteen sets had fewer than five variants.
**Conclusion:** Increased sample size and improved mapping methods have significantly improved our genetic understanding of T1D, which will improve our identification of the genes, molecules and cells involved in the development of T1D.

# THE IMPACT OF MODERN HUMAN SPECIFIC SITES ON HUMAN PHENOTYPES

Christopher R Robles[1], Sriram Sankararaman[1,2]

[1]UCLA, Human Genetics, Los Angeles, CA, [2]UCLA, Computer Science, Los Angeles, CA

A major question in human evolution centers around understanding genetic changes that make anatomically modern humans "unique", i.e. genetic mutations that have enabled cultural and technological breakthroughs over the past 200,000 years. The comparison of modern human genomes with genome sequences from our closest evolutionary relatives, Neanderthals and Denisovans, can be used to study the function of these genetic changes. We identified a list of 745 SNPs in the coding sequences of genes or regulatory elements in which the derived mutation is nearly fixed in the African individuals in the 1000 Genomes Project but not present in at least some of the sequenced high-coverage archaic genomes. These Fixed Derived Mutations (FDMs) are mutations that rose to high frequency in modern humans since the split from the archaic humans and might provide clues to the biology that causes modern humans to differ from our closest relatives.

One approach to study the functional consequences of the FDMs uses the idea that even though these mutations are nearly fixed for the derived allele in African populations, a small fraction of European individuals (~2%) likely carry the ancestral variant at several of these sites due to Neanderthal introgression around 50,000 years ago. As a result, several of the FDMs are polymorphic in Europeans and their impact can be studied by genotyping these mutations in large numbers of individuals with phenotypic information.

We added the FDMs to the UK Biobank Axiom array to study the impact of these mutations, and analyzed about 120 FDMs across 107 distinct phenotypes measured in up to 495,000 people of European ancestry in the UK Biobank dataset. We discovered 102 independent associations of FDMs with 37 phenotypes with p-values that pass a threshold accounting for both number of FDMs and phenotypes tested. These hits are in phenotypes including standing height (13 hits), bone heel density (5), and thyroid-related diseases (2). After appropriately controlling for frequencies and the linkage disequilibrium patterns at these variants, we find that the contribution of FDMs to phenotypic variation at FDMs is significantly depleted in traits including whole body fat mass (Z-score of -4.21), trunk fat mass (Z=-4.91), heel bone mineral density (Z=-6.64), and heel quantitative ultrasound index (Z=-7.18). Our results provide insights into genetic and phenotypic changes that were important for modern human biology since their split from archaic humans.

# POPULATION-SPECIFIC VARIANTS OF INTEREST FOR CARRIER SCREENING IN DIVERSE US HISPANICS

Juan L Rodriguez-Flores[1], Juan C Martinez-Cruzado[2], Robert A Paulino-Ramirez[3], Philip L Leopold[1], Ronald G Crystal[1]

[1]Weill Cornell Medical College, Department of Genetic Medicine, New York, NY, [2]University of Puerto Rico at Mayguez, Department of Biology, Mayaguez, PR, [3]Universidad Iberoamericana, Instituto de Medicina Tropical & Salud Global, Santo Domingo, Dominican Republic

Carrier screening provides couples planning on having children valuable information on the potential risk of rare genetic disorders (RGDs) in future children, focusing on recessive disorders where neither parent is affected, but may be heterozygous for risk alleles. Genomic technology has the potential for enabling expanded carrier screening of thousands of variants and disorders. A major obstacle to screening in the US is the lack of a complete catalog of risk variants in the diverse population, including 52% Whites, 23% Hispanics, 15% Blacks, 6% Asians, < 1% American Indian, Alaska Native, Native Hawaiian or Pacific Islander, and >3% unknown among 3.8 million annual newborns in the US. Based on our prior studies discovering population-specific RGD risk alleles in Qatar that with marked differences in risk allele frequency compared to Europeans, this study aims to test the hypothesis that RGD risk variant (risk var) allele frequencies differ significantly across ethnic sub-populations in the US, with an initial focus on Hispanics, the largest non-white minority among US newborns. In order to test this hypothesis, potential RGD risk variants (risk vars, defined as minor allele frequency < 0.05, computationally predicted among the 1% most deleterious SNVs in the genome, and no homozygotes for minor allele) in n=649 genes on the Affymetrix CarrierScan (CS) array were compared between male/female couples in two major Hispanic groups in the US from the 1000 Genomes Project Phase 3, Puerto Ricans in Puerto Rico (PUR, n=92, n=46 couples) and Mexicans in Los Angeles (MXL, n=58, n=29 couples). Analysis of variant calls for these n=150 genomes identified n=1,941 risk vars in n=473 CS genes, of which n=1,619 (84%) were population specific to PUR (n=1,010) or MXL (n=610), n=637 (33%) were observed in two or more genomes, and n=1,218 (63%) were present in ClinVar (including n=82 pathogenic and n=93 likely pathogenic). While the clinical significance of most of the risk vars is currently unknown, this study embarks towards the long-term goal of expanded carrier screening in the diverse US Hispanic population.

# TISSUE-SPECIFIC EVOLUTION OF PROMOTERS AND ENHANCERS IN MAMMALS IS ASSOCIATED WITH TRANSPOSABLE ELEMENTS

Maša Roller*[1], Ericca Stamper*[1,2], Diego Villar [2], Aisling Redmond[2], Louise Harewood[2], Duncan T Odom[2,3], Paul Flicek[1,2,4]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, United Kingdom, [2]University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, United Kingdom, [3]German Cancer Research Center (DKFZ), Division Signaling and Functional Genomics, Im Neuenheimer Feld 280, Heidelberg 69120, Germany, [4]Wellcome Sanger Institute, Wellcome Genome Campus, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

*Authors contributed equally

Almost a half of a mammalian genome is composed of a large number of transposable elements (TEs), which have produced both lineage specific and mammalian shared insertions. The majority of TEs are thought to be silenced in adult somatic tissue to ensure proper genome function. However, some TEs are expressed in a tissue-specific manner and some have been shown to carry functional binding sites for transcription factors in certain mammalian lineages. The extent to which TEs shape the tissue-specific regulatory landscape of mammals is still unclear. To study the regulatory potential of TEs in mammals, we mapped active promoters, enhancers and gene expression across ten mammals and four tissues. Specifically, we performed ChIP-seq of H3K27ac, H3K4me3, and H3K4me1 to map active promoters and enhancers. We also performed RNA-seq of matched total RNA to quantify gene expression.

We first characterised the extent of tissue-specific regulatory activity and gene expression within species. The vast majority of enhancers are active in only one profiled tissue. The largest group of promoters are active across all profiled tissues, while a quarter of promoters have tissue-specific activity. By integrating the RNA-seq data, we found that tissue-shared expression is characterised by a large number of tissue-shared genes with little isoform diversity. This is in contrast to tissue-specific gene expression, which is characterised by substantial differential isoform usage. Next, we compared the evolutionary rates of tissue-specific and tissue-shared regulatory regions. We found that tissue-specific regulatory regions evolve more rapidly than their tissue-shared counterparts. To explore the mechanisms that might be driving this difference, we examined the TE content of recently evolved regulatory regions. We found that tissue-specific and tissue-shared regulatory regions have consistent preferences for subclasses of long interspersed elements (LINEs) across all mammals. However, the use of short interspersed nuclear elements (SINEs) seems to be lineage specific. Next, we explored tissue-specific regulatory elements with orthologous activity across mammals.Our results provide important insights into the evolution of the tissue-specific mammalian pan-regulome.

# PREDISPOSITION TOWARDS MALIGNANT TRANSFORMATION IN THE HUMAN DISTAL FALLOPIAN TUBE

Ian Rose[1], Mallikarjun Bidarimath[1], Heiko Schultz[2], Blaine Harlan[1], Andrew K Godwin[3], Andrea Flesken-Nikitin[1], Alexander Y Nikitin[1]

[1]Cornell University, Biomedical Sciences, Ithaca, NY, [2] Ludwig Maximilians University, Institute of Pathology, Munich, Germany, [3]Kansas University, Molecular Oncology, Pathology and Laboratory Medicine, Kansas City, KS

An increasingly large body of evidence indicates that Fallopian tube epithelium (TE) is the origin for a substantial fraction of high-grade serous ovarian carcinomas (HGSOCs). Interestingly, serous tubal intraepithelial carcinomas (STICs), the putative precursor lesions to HGSOC tend to occur more frequently in the distal region (closer to the ovary) than in the proximal region (farther from the ovary) of the TE. Histologic and functional assays of human and mouse TE suggest the distal TE may have more adult tissue stem cells (SCs). Since it is generally held that SCs tend to be more amenable to malignant transformation, a higher concentration of SCs in the distal TE could make the distal TE more cancer prone. Alternatively, cyclical exposure to follicular fluid during ovulation could also contribute to the TE's tendency towards malignant transformation. Follicular fluid is rich in inflammatory cytokines, including TNFα, which are thought to play a role in cancer initiation. Exposure to follicular fluid might therefore promote malignant transformation of the distal TE by virtue of the distal TE's proximity to the ovary. However, efforts to understand the basis for the distal TE's tendency towards malignant transformation are hampered by a lack of information about normal human TE biology. To address this problem, we developed a fluorescence activated cell sorting (FACS) strategy to purify populations of stem and differentiated cell enriched populations from both the proximal and distal human TE. To validate our approach, we conducted organoid formation assays. Consistent with our expectations, we find that the stem cell enriched populations show a significantly higher rate of organoid formation. Next, we created mRNA-seq libraries from human stem and differentiated cell enriched populations from both the proximal and distal regions of the TE. Our analysis finds up-regulation of inflammation and immune associated pathways in the distal TE, consistent with a role for follicular fluid in influencing the distal TE's tendency towards malignant transformation.

# AN EMPIRICAL P-VALUE METHOD FOR DETECTING CANCER CELL OF ORIGIN FROM SINGLE CELL RNA-SEQ WITH AN APPLICATION TO LUNG CANCER

Mahboubeh Rostami, Ronald G Crystal, Jason G Mezey

Weill Cornell Meidcal College, Genetic Medicine, New York, NY

The malignant phenotype in many cancers is sustained by cancer stem cells which drives tumor initiation, progression, metastasis, recurrence and drug resistance. For most cancers, experimental work has not definitively demonstrated the existence of cancer stem cells in nonmalignant tissue or identified specific cell types of origin. Motivated by the hypothesis that the genome-wide gene expression profile of a cancer will retain the greatest overall similarity to the cell type were the cancer stem cell originated, we suggest a novel and straightforward approach for identifying cell type of origin using single cell RNA-Seq data. The method is implemented by fitting ANOVA models individually for each gene for cancer tissue paired with each available cell type, where the same number of cells in each gene / pair comparison are preserved by re-sampling. Different p value cutoffs starting from the least significant are then applied to the genome-wide distribution of p values to identify the number of similar genes between the cancer and each cell type. The number of similar genes at the cutoff is then used to assess overall similarity, where statistical significance is assessed with an empirical p-value produced from the re-sampling within each cancer and cell type pair. To demonstrate the method, we re-analyzed the single-cell RNA-Seq lung cancer data of (Lambrechts D et al, Nature Med 2018; 24:1277), which includes data for n=5 patients with lung cancer tumors: two lung squamous cell carcinoma (LUSC), two lung adenocarcinoma (LUAD) and one non-small cell lung cancer (NSCLC). We re-identified cell types within each individual patient with a t-SNE dimension reduction and clustering / batch correction and identified 8 main cell types in each sample: alveolar epithelium, airway epithelium, endothelial, B-cell, fibroblast, myeloid, T-cell and tumor cells. We found that, regardless of the cut-off used, in each patient, the patient's cancer was statistically most similar to one of the cell types. As an example, the NSCLC cancers were each most similar to alveolar epithelium, and the LUAD were each most similar to airway epithelium. While there are clearly caveats to the cell of origin conclusions derived from this method (e.g., if the genome-wide expression profile of cancer cells is significantly altered / impacted by local cell environment or the true cell of origin type is not sampled), our approach represents a new methodology for generating strong hypotheses about cancer stem cell origin when analyzing single cell cancer RNA-Seq data.

# EXCERPT: A COMPREHENSIVE ANALYTIC PLATFORM FOR EXTRACELLULAR RNA PROFILING

Joel Rozowsky[1,2], Robert Kitchen[1,2], Timur Galeev[1,2], Jonathan Park[1,2], James Diao[1,2], Jonathan Warrell[1,2], William Thistlethwaite[3], Oscar Murillo[3], Sai Subramanian[3], Matt Roth[3], Aleksander Milosavljevic[3], Mark Gerstein[1,2]

[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]Baylor College of Medicine, Bioinformatics Research Laboratory, Molecular and Human Genetics Department, Houston, TX

Extracellular RNAs (exRNA) are gaining considerable interest for their potential use as biomarkers and their role in intercellular communication. Small RNA-sequencing has been widely adopted to shed light on the diversity of exRNA in biofluids, but the field faces unique technical challenges. In particular, exRNA samples are more vulnerable to contamination and artifacts from different isolation techniques, they are present in lower concentrations than cellular RNA, and they are occasionally of exogenous origin. To address these challenges, we present the extra cellular RNA processing toolkit (exceRpt), the standardized extracellular small RNA-seq analysis pipeline of the NIH Extracellular RNA Communication Consortium (ERCC). exceRpt is structured as a cascade of filters and quantifications prioritized based on the confidence of a given set of annotated RNA. This involves:

**Pre-processing**: Supports random-barcoded libraries, spike-in sequences for calibration or titration, and explicit removal of common laboratory contaminants and ribosomal RNAs.

**Endogenous alignment**: Alignment to the genome and full set of annotated, potentially spliced, endogenous RNA transcripts including all known miRNAs, tRNAs, piRNAs, snoRNAs, lincRNAs, mRNAs, retrotransposons, and circular RNAs.

**Exogenous alignment**: Alignment to all annotated exogenous miRNAs in miRBase and all exogenous rRNA sequences in the RDP. Finally, alignments to the full genomes of all sequenced bacteria, viruses, plants, fungi, protists, metazoa, and select 'food' vertebrates.

The pipeline generates quality control reports and abundance estimates for RNA biotypes, and is also capable of characterizing mappings to exogenous rRNAs and genomes, which can be used to generate phylogenetic trees. The exceRpt pipeline that has been used to uniformly process all ~2500 exRNA-seq datasets in the exRNA atlas (available at genboree.org and github.org/gersteinlab/exceRpt).

These exRNA profiles are available from the exRNA atlas which is accessible via the consortium website at exRNA.org. The current release of the atlas includes ~6,000 exRNA profiles derived from over 25 billion reads processed uniformly using the exceRpt pipeline. Faceted filtering and data navigation tools are enabled by rich metadata standards developed by the consortium and metadata annotations contributed by the data producers. Users can browse, select, filter, and download processed results and raw data files of open access datasets available in the exRNA Atlas.

# FLEA-CHIP ENABLES REPRODUCIBLE IMMUNOPRECIPITATION FROM 100 CELLS

Marina Ruiz-Romero[1], Cecilia C Klein[1,3], Roderic Guigó[1,2], Sílvia Pérez-Lluch[1]

[1]Centre for Genomic Regulation (CRG), Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Universitat de Barcelona, Departament de Genètica, Microbiologia i Estadística, Barcelona, Spain

Chromatin immunoprecipitation (ChIP) is the preferred technique to uncover in-vivo interactions between proteins and DNA regions. The emergence of the Next-Generation Sequencing technologies and their application to the detection of the immunoprecipitated DNA has allowed for a dramatic increase of the throughput, the resolution and the sensitivity of the ChIPs. The increase in the sensitivity of detection has prompted researchers to challenge the input material to perform the assay. The identification of cell-specific interactions will allow for untangling the epigenetic differences existing within cell populations, as well as to uncover co-occurrence of chromatin modifications. Although some approaches have been reported in this direction, the low amount of immunoprecipitated DNA still represents a challenge when performing libraries for next-generation sequencing.

To fill this technological gap we implemented a fast, low-input, easy and affordable ChIP (FLEA-ChIP). This assay is an evolution of the previously reported iChIP (indexed-first) assay. In the latter, chromatin from ultra-low-input samples is immobilized in proteinA-magnetic beads and indexed prior to immunoprecipitation. However, we performed several improvements to support further reduction of the input material and a deeper-sequencing of immunoprecipitated DNA, allowing for ChIP-Seq sample generation from up to 100 cells for a fair cost (140$ including sequencing). Our approach offers advantages to the currently available protocols: (i) It is highly reproducible; (ii) it can work with frozen samples; (iii) it does not require specific and sophisticated machinery; and (iv) it permits, both, pooling of several samples to reduce technical variability, as well as performing individual assays for those cases in which the sample amount is an important constrain. We applied FLEA-ChIP to produce ChIP-Seq data for the study of chromatin dynamics of Drosophila imaginal tissues during differentiation using a wide collection of antibodies. Our results confirm that FLEA-ChIP is an excellent and versatile approach when working with very-low input samples, and a more reliable and reproducible assay to find out differences between several conditions.

In conclusion, FLEA-ChIP protocol is an adaptable method to any low-input condition as low as 100 human cells. It is affordable for all laboratories thanks to a non-sophisticated procedure and its economic budget. Its implementation will favor the production of high quality, replicable, trustworthy results and it allows chromatin immunoprecipitation assays to get a bit closer to the single-cell era.

# IDENTIFICATION OF LOCAL COPY NUMBER ABERRATIONS IN CANCER GENOMES USING NANOPORE SEQUENCER

Yoshitaka Sakamoto, Ayako Suzuki, Yutaka Suzuki

The University of Tokyo, Computational Biology and Medical Sciences, Chiba, Japan

It is known that a number of structural aberrations exist in cancer genomes in addition to single base mutations, such as large indels, inversions, duplications, and translocations. Some aberrations have very important roles in tumorigenesis, cancer progression, and clinical outcome of the patient, such as the ALK- and RET-fusions in non-small-cell lung cancer. Long-read sequencing technology is now available for identification and characterization of these aberrations, many of which have been overlooked in previous short-read sequencing analyses. Recently, nanopore sequencing technology, namely, PromethION has increased the number of the pores on each flowcell and further parallelizing the flowcells, thereby enabled the production of 100 Gb sequencing per flow cell.
In this study, we conducted the long-read sequencing of four human lung cancer cell lines and ten clinical samples using PromethION. We generated data at the depth of 30x for LC2/ad cells and 10x for the other cases. We characterized the advantages and the disadvantages of the PromethION sequencing using the data from LC2/ad cells. Then, we attempted to detect large-scale structural variations. We identified a unique pattern of local copy number aberrations, which we named Cancerous Local Copy-number Lesions (CLCLs). These CLCLs consisted of a series of complex structural aberrations; the combination of local inversions, duplications, and micro deletions. We found the CLCLs occur even in the most important cancer related genes such as *STK11*, *NF1*, *SMARCA4*, and *PTEN*, where the short-read sequencers could not have confidently detected the aberrant structures. The CLCLs may contribute to elucidate the molecular etiology of the cancer cases where causative cancerous mutations remaining elusive.

# WHOLE EXOME SEQUENCING OF CANINE OSTEOSARCOMA AS A MODEL FOR THE CORRESPONDING HUMAN DISEASE

S Sakthikumar[1,2], I Elvers[1,2], J Kim[2], M L Arendt[3], R Thomas[4], J Johnson[2], R Swofford[2], J Alföldi[2], E Axelsson[1], G Couto[5], M E Pettersson[2], J Meadows[2], J F Modiano[6], M Breen[4], M Kierczak[7], K Forsberg-Nilsson[8], V D Marinescu[2], K Lindblad-Toh[1,2]

[1]Uppsala University, Science for Life Laboratory, IMBIM, Uppsala, Sweden, [2]Broad Institute, Vertebrate Genomics, Cambridge, MA, [3]University of Copenhagen, Section for Internal Medicine, Oncology and Clinical Pathology, Dyrlægevej, Denmark, [4]North Carolina State University, College of Veterinary Medicine, Raleigh, NC, [5]Ohio State University, Department of Veterinary Clinical Sciences and Veterinary Medical Center, Columbus, OH, [6]University of Minnesota, College of Veterinary Medicine, Department of Veterinary Clinical Sciences, St.Paul, MN, [7]Uppsala University, Science for Life Laboratory, WABI, Uppsala, Sweden, [8]Uppsala University, Science for Life Laboratory, IGP, Uppsala, Sweden

Osteosarcoma (OSA) is a rare, nevertheless debilitating, bone cancer that affects humans, especially children and adolescents. A homologous form of OSA spontaneously occurs in dogs, sharing much of the clinical course and attributes. It is also far more prevalent in dogs, especially in larger breeds. Its differential incidence observed across breeds allows for the investigation of tumor mutations in the context of multiple genetic backgrounds. Using canine whole-exome sequencing from three susceptible breeds (22 Golden Retrievers, 21 Rottweilers, and 23 Greyhounds) we compared the genomic landscape both among these breeds and to the human data.

In all three breeds, the OSA tumors showed a high frequency of somatic copy number alterations (SCNA) in key oncogenes and tumor suppressor genes which have been implicated in the human disease. We found three mutational signatures (patterns of somatic mutations) including the so-called aging signature and a novel pattern. A total of eleven significantly mutated genes were identified. TP53 was the most frequently altered gene (83% of samples have either mutations or SCNA in TP53), recapitulating observations in human OSA. The second most frequently mutated gene (21% across all three breeds), histone methyltransferase SETD2, has known roles in multiple cancers, but has not previously been strongly implicated in OSA. This study points to the likely importance of histone modifications in OSA and overall highlights the strong genetic similarities between human and dog OSA. Despite the different genetic backgrounds, the somatic alterations across breeds were uniform, suggesting that canine OSA may serve as an excellent model for developing treatment strategies in both species.

# DETERMINING GROWTH PATTERNS AND DRIVER EFFECTS FROM MUTATIONAL FREQUENCIES IN 994 INDIVIDUAL TUMORS

Leonidas Salichos[1,2], William Meyerson[1,2], Jonathan Warrell[1,2], Mark Gerstein[1,2,3]

[1]Yale University, Computational Biology & Bioinformatics Program, New Haven, CT, [2]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT, [3]Yale University, Computer Science, New Haven, CT

Tumors evolve by accumulating hundreds of mutations. With current advances, we have thousands of whole genome tumors sequenced from large cohorts. This has led to the development of many methods that aim to detect cancer drivers through mutational recurrence and/or deviation from genomic metrics. However, these methods require large sample size and might underperform when recurrence is low. Here, we present a framework for calculating tumor growth and driver effects based on the frequency of hitchhiking mutations for individual tumor samples. We have optimized growth models to define periods of positive or negative growth and detect the presence and effect of putative drivers. To validate our method, we first used simulation models. Then, we applied our method on 993 lower-coverage linear (one child per parent subclone) tumors where the periods of positive growth agree with the presence of drivers identified by the PCAWG consortium through recurrence. Finally, we applied our method to a deep-sequenced AML model tumor where our predicted effect peaks aligned with known cancer genes. Moreover, our results also suggested the presence of additional driver candidates. Overall, our model can be used to shed light on the dynamics of tumor progression, unravel the effect of mutations or genomic regions and highlight affected biological processes (e.g multicellularity, developmental process etc.). Towards an era of personalized diagnosis, our method contributes to the modeling of tumor progression using deep sequenced whole genome data.

# DEVELOPMENTAL DYNAMICS OF lncRNAs ACROSS MAMMALIAN ORGANS AND SPECIES

Ioannis Sarropoulos[1], Ray Marin[2], Margarida Cardoso-Moreira[1,2], Henrik Kaessmann[1]

[1]Center for Molecular Biology (ZMBH), University of Heidelberg, Heidelberg, Germany, [2]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

Only few of the numerous mammalian long noncoding RNAs (lncRNAs) have been functionally characterized and their contribution to organ development remains mostly elusive. Cross-species comparisons provide a powerful framework for the large-scale assessment of the functional relevance of lncRNAs; however, owing to transcriptome data limitations, previous evolutionary studies of mammalian lncRNAs were confined to adult organs.

To investigate the expression patterns, conservation and functional relevance of lncRNAs during organ development, we generated and analyzed an extensive RNA sequencing dataset covering the development of seven major organs (forebrain/cerebrum, cerebellum, heart, kidney, liver, ovary, and testis) from early organogenesis to adulthood across seven amniote species (human, rhesus macaque, mouse, rat, rabbit, opossum, and chicken). We identified thousands of novel lncRNAs in each species, including sizeable sets of loci with dynamic expression during development. These “developmentally dynamic” lncRNAs are enriched for older, longer, more complex and functionally characterized transcripts, and show conserved expression and regulation by organ-specific transcription factors, suggestive of an overall enrichment for functionally relevant loci. Dynamic lncRNAs show the greatest differential expression during peaks of protein-coding gene developmental change, predominantly associated with the establishment of organ identity during early development and the transition to mature organ-specific functions around birth. By integrating the evolutionary and developmental dimensions of our dataset, we identified a striking shift from broadly expressed and conserved lncRNAs in early organ development towards the expression of an increasing number of lineage- and organ-specific lncRNAs during adulthood. Spatiotemporal co-expression with protein-coding genes associated lncRNAs with roles in the development and adult physiology of different organs and identified candidates for the *cis*-regulation of genomic neighbors, including many developmental regulators. Altogether, our study unveils repertoires, expression trajectories, conservation patterns and putative functions of lncRNAs during mammalian organ development.

# INVESTIGATING ATHEROSCLEROSIS PROGRESSION THROUGH SINGLE-CELL TRANSCRIPTIONAL PROFILING OF IMMUNE CELLS OF THE ATHEROSCLEROTIC PLAQUE

A Sartori[1], K Thanopoulou[2], C Borel[1], M Manioudaki[2], I Galani[2], E Andreakos[2], E T Dermitzakis[1,3]

[1]University of Geneva, Genetic Medicine and Development, Geneva, Switzerland, [2]Academy of Athens, Biomedical Research Foundation, Athens, Greece, [3]University of Geneva, iGE3 Institute of Genetics and Genomics of Geneva, Geneva, Switzerland

Atherosclerosis is a chronic and progressive vascular disease of large- and medium-sized arteries. It is characterized by the accumulation and oxidation of low-density lipoproteins in the arteria intima, leading to an innate inflammatory response which results in the formation of an atherosclerotic plaque. The pathogenesis of the plaque is a dynamic process that occurs over decades and its cellular composition is a major determining factor in the disease development. In this study, we propose a single-cell RNA-seq approach, using drop-seq technology, to characterize the immune cell type composition of the plaque at different disease stages from CD45+ sorted immune cells from carotid tissue. Our ultimate goal is to reveal the degree of heterogeneity and variability of cell composition in plaques among individuals and the implications in the plaque development and cardiovascular risk. We also aim to investigate transcriptional differences due to plaque cell composition in symptomatic individuals (i.e. who had plaque fissuring) compared to asymptomatic individuals, to better understand the atherosclerosis progression process. We have sequenced 16225 single-cells from 3 symptomatic and 7 asymptomatic patients, detecting a median of 1152 genes and 360k reads per cell. The top highly expressed genes including *MALAT1*, *CD74*, *B2M* and *TMSB4X* are associated with atherosclerotic lesions, inflammation and adverse cardiovascular outcome. We identified 15 clusters of co-expressed genes differentiating between immune cell subtypes at the plaque level composition with a major proportion of T cells and macrophages, suggesting that the plaque exhibits the characteristics of a classically chronic inflamed site. We identified 4976 differentially expressed genes among clusters suggesting the presence of a high degree of cell heterogeneity in the plaque. Moreover, Gene Ontology revealed enrichment for annotation of pathways involved in pathogenesis of inflammatory diseases, apoptosis, immune response, such as NF-kB and TNF. Further work is aiming to investigate the presence of an expression signature between symptomatic and asymptomatic patients to better characterize the transition among disease stages.

# LARGE, THREE-GENERATION CEPH FAMILIES REVEAL POST-ZYGOTIC MOSAICISM AND VARIABILITY IN GERMLINE MUTATION ACCUMULATION

Thomas A Sasani[1], Brent S Pedersen[1], Ziyue Gao[4], Lisa Baird[1], Molly Przeworski[5,6], Lynn B Jorde[1,3], Aaron R Quinlan[1,2,3]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]University of Utah, Department of Biomedical Informatics, Salt Lake City, UT, [3]University of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [4]Stanford University, HHMI & Department of Genetics, Stanford, CA, [5]Columbia University, Department of Biological Sciences, New York, NY, [6]Columbia University, Department of Systems Biology, New York, NY

In humans, the number of de novo mutations (DNMs) found in an offspring's genome is known to increase with both paternal and maternal age. But does the rate of mutation accumulation in parental gametes differ across families? To answer this question, we analyzed DNMs in 33 large, three-generation pedigrees collected in Utah by the Centre d'Etude du Polymorphisme Humain (CEPH) consortium. We first determined the parent-of-origin for nearly 98% of de novo mutations in the F1 generation, and confirmed significant effects of both paternal (1.57 additional DNMs per year) and maternal age (0.48 additional DNMs per year) on de novo mutation counts in children. Then, using between 4 and 14 F2s in each of the CEPH/Utah pedigrees, we observed significant variability in parental age effects on DNM counts across families, ranging from 0.24 to 3.33 additional DNMs per year. The large numbers of siblings in each CEPH family further revealed that 3% of DNMs originated following primordial germ cell specification (PGCS) in a parent, and differed from non-mosaic germline DNMs in their mutational spectra. Finally, we identified a median of 3 gonosomal mutations per sample in the F1 generation, which, along with post-PGCS DNMs, occurred at equivalent frequencies on the paternal and maternal haplotypes. These results demonstrate that the rate of germline mutation accumulation varies among families with similar ancestry, and confirm that parental mosaicism is a substantial source of de novo mutations in children.

# RETROTRANSPOSON ACTIVITY IN FRONTOTEMPORAL DEMENTIA

Natalia Savytska, Kevin Menden, Tenzin Nyima, Margherita Francescatto, Patrizia Rizzu, Peter Heutink

German Centre for Neurodegenerative Diseases (DZNE), Genome Biology of neurodegenerative diseases, Tuebingen, Germany

Transposable elements (TE) play an important role in brain development and disorders (e.g. Rett syndrome)[1]. We previously demonstrated an abundant genomic activity of TEs in healthy human brains[2]. In the current study we assess the transcriptional activity and potential role TEs may play in fronto-temporal dementia (FTD) patients.

FTD is a progressive presenile neurodegenerative disease. In 10% of cases the phenotype is combined with amyotrophic lateral sclerosis (ALS). FTD has a strong genetic component with as many as 60% of patients, caused by a mutation in MAPT, GRN or C9orf72 and manifested by tau or TDP-43 proteinopathy. GWAS studies on sporadic FTD and ALS cases also detected associated loci awaiting replication. The molecular mechanisms of FTD pathogenesis are subject to intense research efforts.

In a drosophila model of TDP-43-related ALS, TE overexpression in nervous system with relation to motor deficits has been described[3]. Recent analysis of ALS patients brain transcriptome confirmed the overexpression of TEs in humans, but the relation between the disease and this event are yet to be discovered[4].

We aim at assessing the transcriptional activity of TEs in the human frontal lobe, affected by FTD. We hypothesize that in FTD TEs expression is significantly elevated, affecting gene splicing and expression, and contributing to FTD pathogenesis. TEs overexpression leads to transposition rate increase, that manifests in gene disruption and DNA damage response (DDR) activation, detectable in transcriptome.

For this purpose, CAGEseq and total RNAseq was performed for 4 groups – healthy controls, and FTD patients with MAPT, GRN or C9orf72 mutations. 50 samples from the frontal lobes were obtained -16, 12, 8 and 14 for the described groups respectively. CAGEseq data analysis identified 371 TEs actively transcribed in all samples with ERV and Alu TEs enriched (Fisher`s test; p<e-10-p

The study results will identify TEs, co-opted for cis-regulation of gene expression in the frontal cortex and potential effects TEs have on DDR in FTD.

PMID:
1. 23057747
2. 22037309
3. 28301478
4. 28637276</e-10-p

# POPULATION GENETICS OF *IN VITRO* NEURAL DIFFERENTIATION

Daniel Seaton[1], Julie Jerber[2,3], Daniel Pearce[3], Florian Merkle[4], Oliver Stegle[1], Daniel Gaffney[3]

[1]EMBL-EBI, Statistical Genomics, Hinxton, United Kingdom, [2]Open Targets, Neurodegeneration Focus, Hinxton, United Kingdom, [3]Wellcome Sanger Institute, Cellular Genetics, Hinxton, United Kingdom, [4]University of Cambridge, Wellcome Trust-Medical Research Council Institute of Metabolic Science, Cambridge, United Kingdom

*In vivo*, the early stages of human development are highly robust. However, *in vitro* models of development such as iPSC differentiation are highly variable, with cell lines varying in their differentiation capacity. What causes this variation? One hypothesis is that common genetic variation between individuals leads to differences in the capacity of iPSCs derived from these individuals to differentiate.

We have developed a combined experimental and computational strategy for testing this using a population-scale human induced pluripotent stem cell (hIPSC) bank. By combining single-cell RNA sequencing (scRNA-seq) with a multiplexed experimental design of iPSC lines from 145 individuals, we assayed gene expression and cell type composition at three timepoints in neural development, from early neural progenitors (day 11), to functional neurons (days 30, 52).

In order to identify common genetic variants involved in differentiation, we developed a novel computational pipeline. First, we identified genetic variants with strong effects on gene expression during differentiation (i.e. eQTL variants) in each cell type. A significant challenge in discovering eQTL using scRNA-seq is the extreme variation in the number of cells sampled for each individual, spanning 3 orders of magnitude. We developed a simple statistical model to account for this variation, increasing the number eQTL discoveries by up to 2-fold in each cell type, up to a maximum of 1,700 eQTL in any one cell type (10% FDR). Variation in the strength of these eQTL across individual cells was then quantified by using allele-specific expression. This provided an internally controlled assessment of eQTL effect sizes, robust to the high technical noise present in scRNA-seq datasets.

Finally, we tested for association of our discovered eQTL variants with the developmental trajectory of each iPS cell line. This identified a small subset of common variants that appear to act from early developmental stages onwards to disrupt *in vitro* development. Future work will seek to establish whether these variants are also important during *in vivo* development, and ultimately impact disease risk and other traits.

# DISSEMINATION OF BIOINFORMATICS TOOLS AND WORKFLOWS FOR REUSIBLITY, FINDABILITY, UNDERSTANDABILITY THROUGH TRUWL.COM

Karl B Sebby, Tim Thompson

xD Bio Inc., NA, Whitefish, MT

Truwl.com (truwl) is a web based application specifically designed for broad dissemination of bioinformatics methods in a findable, understandable, and usable way. The ability to find, vet, and implement proper computational methods remains exceedingly complex, slow, and costly and is hampering advances in genomics research. Lack of solutions to find, understand, and apply established methods has led to a lack of reproducibility, underuse of open source tools, and continued dependence on outdated tools. Efforts to address these problems include placing code and resources in publicly available repositories, handling software installation and dependencies through package managers and containerization, and assembling calls to multiple tools into workflows which can be run much like a single program with a workflow management system. Commercial and open source bioinformatic platforms exist to provide users with graphical interfaces for using methods without needing skills to work on the command line, but the content in these platforms is siloed, difficult to find, and often inaccessible behind paywalls and account barriers. Bioinformatics tools and workflows published on truwl are exposed to crawlers through server-side rendering and provide context through tags, commenting, and voting making methods easily findable through search engine queries. The proper use of tools can be understood through collections of use cases that show examples of commands line inputs and parameter settings. File inputs link to pages that describe the them and link to any know locations where the files can be obtained. Workflows are displayed as interactive graphs that can be navigated to the tools that they use and are similarly linked to sets of use cases that enable users to see how the workflows are used in practice. Further, the tools and workflows displayed on truwl are described by files written in the Common Workflow Language and Workflow Description Language that are standards to make methods executable by a variety of workflow management systems. The combination of a user friendly web interface, focus on findability, platform agnostic executable descriptions, and community curation tools promises to allow users to share, find, and implement the most appropriate and up to date bioinformatics methods.

# APPLICATION OF CNVNATOR TO ANALYZE COPY NUMBER ALTERATIONS IN CANCER

Shobana Sekar, Milovan Suvakov, Alexej Abyzov

Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN

Changes in the number of copies of a genomic region, also known as copy number variations and alterations (CNVs/CNAs) have been linked to several human diseases including cancer. Around 90% of tumors exhibit CNAs as well as gain or loss of a part or whole chromosome (i.e., aneuploidies). CNAs and aneuploidies are similarly frequent in cancer subclones and metastasis. Discovery and analysis of CNAs and aneuploidies is important for understanding cancer evolution, cancer progression and invasiveness, as well as individual response to treatment. CNVnator is a read depth based approach to perform copy number analysis and genotyping on whole genome sequencing (WGS) data. Initially developed for detection of CNV in the normal genome, CNVnator is being expanded for the analysis of somatic CNAs in cancer. Recently added features include functionality to import single nucleotide polymorphism data from variant call format (VCF) files. Using this data, users can generate B-allele frequency (BAF) likelihood plots along with RD signal to facilitate analysis of subclonal CNAs and copy-number neutral losses of heterozygosity. Further, in order to address the pressing need for the analysis of large number of WGS data for cancer, we have introduced an option to reduce the footprint of intermediate files by about 10 times. We demonstrate the application of these new functionalities along with the analysis of subclonal CNAs on colorectal cancer samples with residual polyp from 13 patients.

# A MACHINE-LEARNING METHOD TO DETECT NON-CODING REGULATORY VARIANTS IN CANCER GENOMES USING ATAC-SEQ AND CHIP-SEQ

Arko Sen, Selene Tyndale, Graham McVicker

Salk Institute of Biological Studies, Integrative Biology Laboratory, La Jolla, CA

Most known cancer driver mutations are within protein coding regions of the genome, however, there are now several established and emerging examples of non-coding regulatory mutations that drive cancer progression. Despite the importance of non-coding mutations, most cancer studies have focused on the protein-coding genome. In part, this this because most cancer studies have relied upon on exome sequencing, but it is also difficult to predict the function of non-coding mutations without additional information from functional genomics experiments such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) or the assay for transposase-accessible chromatin (ATAC-seq). To address these limitations, we have developed a scalable machine-learning based method to identify insertions and deletions in regulatory regions of the genome using aligned ChIP-seq and ATAC-seq reads. We call our method BreakCA for Breaks in Chromatin Accessible regions. Under most conditions, BreakCA out-performs popular variant callers such as the GATK-HaplotypeCaller and VarScan2 and we verify that it detects known oncogenic regulatory mutations using ATAC-seq reads from T-cell acute lymphoblastic leukemia cell lines. We applied BreakCA to identify indels in H3K27ac ChIP-seq peaks in 23 neuroblastoma cell lines and filtered out common germline variants using allele frequencies from the genome aggregation database (gnomAD) and the genotype tissue expression project (GTEx). We identified a small number of rare or somatic indels that occur in multiple neuroblastoma tumors and that are associated with the expression of nearby genes. Of particular interest is a 7bp CGCCTCGG/C rare germline deletion (MAF =1e-4) which replaces a TFAP2B binding motif with a ZNF263 motif within the first intron of the histone demethylase KDM5B. The two neuroblastoma cell lines with this indel have very high KDM5B expression, which is a known prognostic indicator in gliomas. Further characterization of this indel as a potential oncogenic risk factor is therefore warranted.

# MULTI-OMICS MODULES CONSTRUCTION BY NON-PARAMETRIC DETECTION OF CONSERVATION IN FEATURES' RANKS CORRELATIONS IN MULTIPLE CANCERS.

Sarun Sereewattanawoot, Ayako Suzuki, Masahide Seki, Yutaka Suzuki

The University of Tokyo, Department of Computational Biology and Medical Sciences, Chiba, Japan

**Backgrounds:** Malignancies are generally classified according to their tissue of origin and morphology, recent successes in the field of targeted therapy showed that molecular or genetic level classification could provide a more comprehensive treatment. Many studies also revealed that the same driver genes could be over-expressed or mutated across wide varieties of malignancy, hinting that many kind of malignancies could arise from the same roots. Moreover, with the roles of epigenetic features becoming clearer and many specific promoters or enhancer's action have been verified, it should be natural that epigenetics regulation could played a big role in carcinogenesis.
**Methods:** To elucidate which omics-features served as the fundamental building blocks across various cancers, we integratively analyzed genomics, epigenetics and transcriptomics features in laboratory cell line and clinical sample of various origins in public data deposit. For clinical dataset, we retrieved TCGA samples with Array-based CNV and Methylation levels and sequencing based expression profiles. For cell line, we utilized 23 Lung adenocarcinoma cell lines with matching WGS, 8 Chip-Seq (Pol-II and 7 H3's Lysine modifications) and RNA-seq. We based our approach on feature-to-feature Spearman's rank correlation to facilitate the integration of multiple datasets and to eliminate the need to account for how each feature are influencing each other. By looking for conservation of ranks' correlations across multiple cancer samples and types, we were able to integrally grouped features that mutually behave in the same direction, that is, consistently move up and down in rank in the same samples, suggesting associations between them and by observing how these correlations waned off in different groups across different samples, inter-groups interactions could be implied. From these features groups and interactions, we re-applied the results back into the inputs with the intention to determine the features-modules compositions of each sample.
**Results:** We were able to group multi-omics features with strong associations, hinting functional interaction between them and bridging the current enigmatic epigenetics features to the transcriptomes, which could provide a systemic approach in elucidating their functions. By observing how each group behave in different samples, many unknown features could be evaluated and genomics, epigenetics and transcriptomics compositions of each sample could be resolved.

# ENZYMATIC METHYL-SEQ UNCOVERS THE REGULATORY BACKBONE OF THE CANNABIS SATIVA GENOME

Brittany Sexton[1], Louise Williams[1], V.K. Chaithanya Ponnaluri[1], Keerthana Krishnan[1], Luo Sun[1], Yvonne Helbert[2], Lei Zhang[2], Christine Sumner[1], Fiona J Stewart[1], Bradley W Langhorst[1], Timothy Harkins[2], Kevin J McKernan[2], Eileen T Dimalanta[1], Theodore B Davis[1]

[1]New England Biolabs, Inc., Applications and Development, Ipswich, MA, [2]Medicinal Genomics Corp., Woburn, MA

*Cannabis sativa* is an industrial crop producing more economic value than the top five crops in California combined. The FDA has recently approved two cannabinoids for use to treat and alleviate symptoms for a wide range of diseases, including nausea in cancer patients undergoing chemotherapy, while studies are examining their use for reduction of anxiety, inflammation, and pain relief. The female plant produces cannabinoids, and two factors known to reduce cannabinoid production are pollination and stress. Sex determination in cannabis is therefore an important economic factor and in order to express high levels of cannabinoids female plants are grown in isolation from male plants. Unfortunately, female plants are capable of stress induced hermaphroditism that can lead to acres of pollination and low cannabinoid yields.

DNA methylation is an important epigenetic mechanism that regulates many cellular processes, including sex determination and seed production. Currently, whole genome bisulfite sequencing (WGBS) is the gold standard for determining 5-methylcytosine (5mC); however, WGBS introduces DNA damage, results in high GC-bias leading to skewed methylation profiles, and often requires large amounts of starting material. To further complicate matters, the cannabis genome is 67% AT-rich and 83% AT-rich after bisulfite conversion making methylome analysis exceedingly difficult and to date no methylome exists for cannabis.

Enzymatic Methyl-seq (EM-seq™) is a novel enzymatic method used to detect 5mC and 5hmC. Here, we used EM-seq to better understand sex determination and the hermaphroditic process in cannabis. For technical comparative purposes, female leaf genomic DNA was used to generate EM-seq and WGBS libraries. To gain insight into methylation based regulation of the hermaphroditic process, EM-seq libraries were prepared using genomic DNA from female flowers, female seeded flowers and male flowers. EM-seq successfully identified 5mCs and 5hmCs across the cannabis genome. The EM-seq libraries were superior to WGBS, having higher library yields, more accurate methylation information, reduced DNA damage, increased sequencing length and decreased GC-bias. Additionally, RNA-Seq data was generated to establish functionality between DNA methylation and the cannabis transcriptome. EM-seq enabled the generation of the first cannabis methylome and opens up the possibility to discover critical and exciting regulatory functions for the hermaphroditic process and cannabinoid pathway regulation.

# GERMLINE AND SOMATIC VARIANT CALLING WITH NOVASEQ™ 6000 2X250BP READS

Minita Shah, Molly Johnson, Kanika Arora, Rashesh Sanghvi, Marta Byrska-Bishop, Wayne E Clarke, Uday Evani, Kshithija Nagulapalli, Michael C Zody, Soren Germer, Jade Carter, Giuseppe Narzisi, Nicolas Robine

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

Longer reads can substantially improve bioinformatics analysis. Despite the recent advances in ultra long read technologies (e.g. PacBio and Oxford Nanopore), Illumina's short reads have an unmatched throughput and error rate. The recently released NovaSeq™ 6000 system now supports a 500 cycle kit (so far only available on the SP flow cell) that allows the generation of 2x250bp paired reads, effectively increasing by 66% the read length over their previous 2x150bp kit.

To evaluate the improvements provided by the longer Illumina reads, we sequenced 7 samples on the new NovaSeq™ 6000 SP 500 cycle kit (2x250bp) and evaluated the performance of the run, in comparison with the same samples sequenced on the standard S4 2x150bp kit. We employed the well-characterized CEU HapMap trio (NA12878, NA12891, and NA12892) and two breast cancer cell lines (HCC-1143 and HCC-1187) with matched normal cell lines from lymphoblasts isolated in peripheral blood.

We ran our standard germline and cancer pipelines, including alignment with BWA-MEM, variant calling with a variety of algorithms, filtering, and annotation. We compared concordance of small and large variants between the two kits. Overall the 250bp reads show comparable accuracy to the 150bp reads under a variety of different variant calling experiments, while also boosting additional power in specific applications. For example, comparison of the SNP and indel calls to the "Genome In A Bottle" (GIAB) truth set for NA12878 shows increased recall for indels using the 2x250bp reads while preserving the same high standards of precision and recall for SNPs achievable with the 2x150bp data. Specifically, there were approximately half as many false negative indel calls in the 2x250bp as compared to the 2x150bp data. For the cancer cell lines, the somatic SNVs, indels and structural variants from the 2x250bp data show good concordance with the 2x150bp data, at similar coverage, additionally capturing low-VAF variants only seen in super-deep (~275x) 2x150bp data. Finally, we also explore the improvements provided by 2x250bp for HLA typing and alignment to pharmacologically-relevant duplicated genes.

# IDENTIFYING SURVIVAL-ASSOCIATED BIOMARKERS FROM >17,000 CANCER GENOMES

Jason M Sheltzer

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

Successful treatment decisions in cancer depend on the accurate assessment of patient risk. To improve our understanding of the molecular alterations that underlie deadly malignancies, we collected and analyzed the genomic profiles of 17,879 solid tumors from patients with known outcomes. To our knowledge, this is the largest such study that has ever been conducted, and the first study to perform a pan-cancer, exome-wide analysis of prognostic mutations, copy number alterations, CpG methylation sites, and more.

Our analysis finds that mutations in cancer driver genes are almost never associated with patient survival time. At the same time, copy number changes in these same genes exhibit significant prognostic power, independent of their mutation status. Focal copy number alterations are associated with worse outcomes than broad alterations, and driver gene CNAs remain strongly prognostic when controlling for tumor stage, grade, TP53 status, and total tumor aneuploidy. Analysis of methylation, microRNA, mRNA, and protein expression in primary tumors define several additional prognostic patterns, including signatures of tumor mitotic activity and tissue de-differentiation, that remain strongly associated with patient outcome after correcting for common clinical parameters. By combining data across independent patient cohorts, we further identify robust genetic biomarkers of patient outcome, and demonstrate that a subset of these alterations also confer specific therapeutic vulnerabilities. In total, our analysis establishes a comprehensive resource for biomarker identification, underscores the importance of gene copy number profiling in assessing clinical risk, and provides a precision-medicine approach to treating cancer patients with the worst likely outcomes.

# FTX-1821 EFFECTIVELY TREAT FSHD AT ITS ROOT CAUSE, AND SPECIFICALLY REPRESS EXPRESSION OF DUX4 PROGRAM GENES

Ning Shen, Alejandro Rojas, Erin Valentine, Peter Rahl, Anthony Accorsi, Joseph Maglio, Alan Robertson, Richard Barnes, Steven Kazmirski, Lorin A Thompson, Owen Wallace, Diego Cadavid, Lucienne Renco, Angela Cacace, Aaron Chang

Fulcrum Therapeutics, Cambridge, MA

Facioscapulohumeral muscular dystrophy (FSHD) is a common muscular dystrophy, affecting approximately 4 people per 100,000 across EU and US. This genetic disease is caused by the loss of repression of the homeobox transcription factor DUX4, leading to its expression in skeletal muscle. This event results in the activation of its downstream transcriptional program causing apoptosis, loss of skeletal muscle and progressive motor disability. Using optimized in vitro myotube culture systems and robust assays to screen a small molecule probe library, we identified and validated a drug target that, when inhibited, reduces or prevents expression of DUX4, its downstream transcriptional program and muscle cell death. FTX-1821, a potent and highly selective small molecule inhibitor of this drug target, prevented activation of DUX4 without impacting myogenesis. Transcriptomic profiling of FTX-1821 treated FSHD myotubes showed concentration-dependent reduction of DUX4 and its downstream program, and FTX-1821 is highly specific for inhibiting expression of DUX4 program genes. To understand how FTX-1821 specifically impacted DUX4 program genes without inducing much gene expression change for genes not associated with DUX4, we carried out H3K27ac ChIP-seq experiment. We found that compared to the DUX4 program specific gene expression change, FTX-1821 treatment has moderate but broader impact genome-wide at chromatin activity level. Interestingly, genes associated with differential H3K27ac peaks do not show difference in neither promoter activity nor RNA expression. In contrast, the DUX4 program genes that responded to FTX-1821 treatment showed decreased H3K27ac signal at promoters. We identified several transcriptional regulators that could account for the global chromatin activity change and DUX4 expression change. We are continuing to explore the regulation of DUX4 expression and mechanisms of action of FTX-1821 in FSHD myotube cells. This work not only brings new insights in understanding FSHD disease and DUX4 regulation, but also supports our efforts to bring transformative therapies that restore the balance in gene expression for genetically defined diseases.

# IDENTIFYING TRANSCRIPTION FACTOR BINDING SITE MOTIFS FROM ENCODE EPIGENOMIC DATASETS BY MAPPING POSITION WEIGHT MATRICES TO GAPPED K-MERS

Dustin Shigaki[1], Ayoti Patra[1], Michael A Beer[1,2]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD,
[2]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

Regulatory variants have to been shown to cause disease by disrupting or creating transcription factor binding sites (TFBS's). Gapped k-mers are effective features for predicting the impacts of these variants, but building improved models with higher order configurational constraints on TFBS grammar requires reduced binding site descriptions with more statistical support for training. A natural choice is the position weight matrix (PWM), which is the most popular mathematical structure to model a TFBS. It is easy to construct and interpret, but it lacks the predictive power of gapped k-mers (Ghandi et al 2014). gkm-SVM is a support vector machine which outputs a large weight vector (length ~10^6) that ranks the relative importance of each gapped k-mer in predicting enhancer activity. Due to the size of this weight vector, it is taxing to search for TFBSs by eye. In order to create more compact representations of gapped k-mer weights, we developed a function that maps PWMs to gapped k-mers, which we call the PWM generated gapped k-mer counts. We show that the weight vector from gkm-SVM is a linear combination of the PWM generated gapped k-mer counts of some set of PWMs, which allows us to identify the core set of active transcription factors in chromatin accessibility data by using simple regression methods. Furthermore, we demonstrate how we can use this mapping to construct multiple *de novo* PWMs. Lastly, we present the results of the CAGI5 Regulation Saturation Challenge, where the participants were asked to predict variant effects on gene expression from massively parallel reporter assays in regions associated with known disease. We found that incorporating multiple models that were trained on chromatin accessibility data (DHS) and transcription factor binding data (ChIP-Seq) considerably improves prediction performance.

# NATURAL SELECTION ON MUTATIONS AFFECTING HUMAN COMPLEX TRAITS

Yuval B Simons[1], Guy Sella[2,3,4]

[1]Stanford University, Genetics, Stanford, CA, [2]Columbia University, Biological Sciences, New York, NY, [3]Columbia University, Systems Biology, New York, NY, [4]Columbia University, Program for Mathematical Genomics, New York, NY

Many phenotypes of interest are genetically complex, in that heritable variation in the phenotype arises from numerous segregating loci distributed across the genome. While the evolution of complex traits has been studied for many decades, we still know little about the underlying evolutionary parameters that give rise to heritable variation. Here, we leverage human genome-wide association results to infer the distribution of selection effects and target size of mutations affecting several human morphological, biomedical and life history complex traits, to our knowledge for the first time.

Specifically, we build on our recent theoretical work to show that the population frequency of a trait-affecting variant imposes an upper bound on the strength of selection acting on it, while its effect size on the trait imposes a lower bound. We then construct a smoothed, maximum composite-likelihood estimator that relies on all the genome-wide significant (GWS) associations for a trait to infer the distribution of selection coefficients of the variants that they tag. We use this distribution to estimate the rate of mutations affecting the trait and their distribution of selection coefficients. Simulations suggest that our method performs well for GWAS with at least 100 GWS associations, providing precise estimates of selection coefficients in the range of $10^{-5} < s < 10^{-2}$.

We apply our method to GWAS performed in the UK Biobank, focusing on traits with non-negligible explained heritability ($h_g^2 > 0.05$), including height, BMI, male pattern boldness, age at menarche and systolic blood pressure. With estimates of these evolutionary parameters in hand, we can predict how the number of associations and explained heritability for these traits should increase with GWAS sample sizes.

# IMPROVING AND USING A CHROMOSOME-SCALE ASSEMBLY OF THE ENORMOUS (32 GB) AXOLOTL GENOME

Jeramiah J Smith[1], Nataliya Timoshevskaya[1], Vladimir A Timoshevskiy[1], Melissa C Keinath[2], S. Randal Voss[3,4,5], AGC [1]

[1]University of Kentucky, Biology, Lexington, KY, [2]Carnegie Institution for Science, Department of Embryology, Baltimore, MD, [3]University of Kentucky, Spinal Cord and Brain Injury Research Center, Lexington, KY, [4]University of Kentucky, Department of Neuroscience, Lexington, KY, [5]University of Kentucky, Ambystoma Genetic Stock Center, Lexington, KY

The axolotl (*Ambystoma mexicanum*) has served as an important model for studying diverse areas of inquiry including: evolution, development and regeneration. However, its large genome (~32 gigabases) presents a formidable barrier to genetic analyses. Building on long-term efforts to sequence the axolotl genome, we recently adapted an established mapping approach, using low coverage sequencing form 48 siblings, that allowed us to dramatically improve the genome assembly. This cost-effective approach allowed us to increase the contiguity of the assembly by three orders of magnitude beyond the previous draft assembly and assemble the axolotl genome in 14 scaffolds. Subsequent validation by fluorescence *in situ* hybridization was used to verify the structure of these 14 scaffolds and assign each to its corresponding physical chromosome. This new assembly covers 27.3 gigabases and encompasses 94% of annotated gene models on chromosomal scaffolds. Individual assembled chromosomes range in size from 0.66 to 3.17 gigabases and are each the size of many entire vertebrate genomes, the biggest being approximately as large as the human genome. Ongoing studies have further improved this assembly and leverage the assembly to further our understanding of the axolotl's genome biology, its developmental biology, and key genomic changes that have occurred over its recent evolutionary history in nature and during its 150 years of domestication in the laboratory. This chromosome-scale assembly has also permitted the implementation of a cost-effective and bulked-sergeant RNAseq strategy that permits rapid identification of mutant genes within this large genome. These include the genes underlying two mutant phenotypes that have evaded genomic characterization for several decades: *cardiac* (which impacts heart development) and *short toes* (which impacts limb development/regeneration).

# GENOME REPROGRAMMING OVER EONS AND EMBRYOGENESIS: LESSONS FROM THE SEA LAMPREY (*PETROMYZON MARINUS*)

Jeramiah J Smith, Cody Saraceno, Courtney K Waterbury, Vladimir A Timoshevskiy, Nataliya Timoshevskaya, SLGC

University of Kentucky, Biology, Lexington, KY

The lamprey genome provides unique insights into both the deep evolutionary history of vertebrate genome/developmental biology and the maintenance of genome structure/integrity over development. The lamprey lineage diverged from all other vertebrates approximately 500 million years ago. As such, comparisons between lamprey and other vertebrates permit reconstruction of ancient duplications and rearrangement events that defined the fundamental genome architecture, gene content and developmental biology of all extant vertebrate genomes. Lamprey also undergoes programmatic changes genome structure that result in the physical elimination of ~20% of its genomic DNA (~0.5 Gb from a ~2 Gb genome) from all somatic cell lineages during early embryonic development. Here, we outline recent progress in assembly and analysis of the lamprey germline genome, and progress in the development of methods for characterizing the cellular events that mediate DNA elimination and epigenetic reprogramming. We have integrated information from several sampling approaches and sequencing technologies to achieve a highly contiguous assembly of lamprey genome. This genome assembly has dramatically improved our ability to dissect the molecular basis and genetic outcomes of programmed genome rearrangements (PGRs) and has improved our understanding of the tempo and mode of large-scale duplications and translocations within the ancestral vertebrate lineage. Analysis of the germline genome identifies several genes that are physically eliminated from all somatic tissues. These eliminated genes correspond to several known oncogenes that are somatically silenced by Polycomb Repressive Complex during early embryogenesis and reveal several novel oncogene candidates. Cas9-midiated knockout of Polycomb proteins indicates that these genes play a functional role in DNA elimination during lamprey PGR, and functional analyses in human cells provide more direct support for the hypothesis that lamprey uses PGR to eliminate a specific subset of "germline" genes that act as oncogenes in somatic tissues. These functional analyses are complemented by the development of approaches for in situ analysis of 3D preserved cells, which have revealed that PGR unfolds through a series of dramatic cellular events that involve the programmatic alteration of several fundamental mechanisms of mitosis and epigenetic silencing, including: chromatid cohesion, nuclear envelope formation, and covalent modifications of DNA and histones.

# *DROSOPHILA MELANOGASTER* RNA EXPRESSION PATTERNS IN FLIES SELECTED FOR LONG AND SHORT NIGHT SLEEP DURATION

Caetano Souto-Maior, Yanzhu Lin, Yazmin L Serrano, Susan T Harbison

Laboratory of Systems Genetics, National Heart Lung and Blood Institute, Bethesda, MD

Sleep patterns are complex, polygenic phenotypes. The genes and networks responsible for them are therefore not fully known and understood. While engineered constructs in *Drosophila* genes enable the identification of genetic pathways influencing sleep, they cannot assess the importance of natural variants. Artificial selection, however, can uncover variation found naturally in wild populations. We used transcription profiling of wild-derived populations to investigate the changes in RNA levels associated with increased or decreased sleep. From a previously created outbred population, two replicate populations of short- and long-sleeping flies were selected over 13 generations, with two additional unselected populations as controls. We measured sleep in 100 males and 100 females of each population each generation over that time, choosing the most extreme 25 sleepers of each sex as parents for the next of generation. RNA-seq data was obtained for two pools of 10 randomly-selected flies per sex, selection scheme, replicate population, and generation, a total of 312 libraries. Long-sleeper and short-sleeper lines had significantly different night sleep across all generations ($P = 0.0003$). At generation 13, the long-sleeper populations averaged $642.2 \pm 3.83$ and $667.8 \pm 2.97$ minutes of night sleep for replicates 1 and 2, respectively; short-sleeper counterparts averaged $104.3 \pm 6.71$ and $156.2 \pm 8.76$ minutes of night sleep. In contrast, the two control populations did not have significant differences ($P = 0.7083$). Analysis of the RNA-seq data using generalized linear models (GLMs) revealed significant shifts in the expression of individual genes of both known and unknown functions for each selection scheme. Additionally, the analysis of GLM-significant subsets with non-linear, non-parametric gaussian processes models allow further inference about networks of interacting genes; improving the size of subset analyzed jointly, precision of the method, and validating results experimentally are some of the next steps required for better of time-course artificial selection data.

# TRACING THE EVOLUTION OF TWO TRANSMISSIBLE CANCERS IN TASMANIAN DEVILS

Maximilian R Stammnitz[1], Young Mi Kwon[1], Kevin C Gori[1], Adrian Baez-Ortega[1], Tracy Wang[1], Rodrigo Hamede[2], Hannah V Siddle[3], Mathew J Garnett[4], Fengtang Yang[4], Zemin Ning[4], Elizabeth P Murchison[1]

[1]University of Cambridge, Department of Veterinary Medicine, Cambridge, United Kingdom, [2]University of Tasmania, School of Natural Sciences, Hobart, Australia, [3]University of Southampton, Centre for Biological Sciences, Southampton, United Kingdom, [4]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Transmissible cancers are clonal lineages that spread through populations via contagious cancer cells. Such epidemics have only ever been observed to affect two mammals in nature, yet two distinct clones known as devil facial tumour 1 (DFT1) and devil facial tumour 2 (DFT2) affect the same marsupial carnivores: Tasmanian devils. As a consequence of these aggressive infectious cancers and their rapid spread across the island of Tasmania, the species has been undergoing a substantial population decline over the past two decades.

Using whole genome sequencing, drug screening and cytogenetics, we have recently provided a first genetic and functional comparison of these two transmissible Tasmanian devil cancer lineages [1]. Our results suggest that both DFT1 and DFT2 emerged recently and independently, from similar Schwann cell related progenitors in two unremarkable devils. Analyses of the sex chromosomes show that DFT1 arose in a female devil, whereas DFT2 first arose in a male. This research has also helped us to establish a cancer type specific set of chromosomal rearrangement biomarkers, which allowed us to screen over 500 tumour biopsies from wild animals over a wide spatiotemporal range [2].

By deep sequencing of whole genomes from tumour cohorts of DFT1 and DFT2 infected devils, we are starting to shed light on the evolutionary trajectories of both lineages. Our somatic phylogenies allow for a detailed longitudinal comparison of inter and intra-specific diversification, mutation rates, clonal selection and immunological adaptation events. Additionally, we integrate ultra-long nanopore sequencing reads to improve the existing Tasmanian devil reference and to resolve the two transmissible cancer genomes at unprecedented detail.

[1] Stammnitz, M., Coorens, T., Gori, K., *et al.* (2018). The origins and vulnerabilities of two transmissible cancers in Tasmanian devils. *Cancer Cell 33 (4)*
[2] Kwon, YM., Stammnitz, M., Wang, T., *et al.* (2018). Tasman-PCR: a genetic diagnostic assay for Tasmanian devil facial tumour diseases. *Royal Society Open Science 5:180870*

# MOBILE ELEMENT INSERTIONS AND ASSOCIATED STRUCTURAL VARIANTS OCCUR EARLY IN LONGITUDINAL BREAST CANCER SAMPLES

Cody J Steely, Julie E Feusier, Yi Qiao, Gabor T Marth, Lynn B Jorde

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

BACKGROUND: Mobile elements comprise roughly half of the genome, with *Alu*, SVA, and LINE-1 elements actively inserting and mediating structural variants (SVs). Mobile elements are inactive in most somatic cells but have shown increased activity in cancer. The majority of previous studies have focused on LINE-1, the only active autonomous mobile element. Here we examine each active mobile element in four breast cancer patients throughout cancer progression, characterizing insertions and mobile element-associated SVs to better understand the timing and significance of these variants through cancer development and progression.

METHODS: Longitudinal whole genome sequencing (WGS) data were acquired from malignant tissue from four breast cancer patients. Multiple mobile element detection algorithms were employed to identify insertions or SVs that were mediated by mobile elements in tumor WGS. These potential variants were compared to matched germline WGS data and visualized using IGV to ensure they were unique and not the product of mismapping.

RESULTS: The majority of insertions and variants identified were found at the first tumor sampling time point, and remain present in later tumor samples. A smaller number of events continue to accrue as the tumor develops. Additionally, at least one mobile element-associated structural variant occurred in a tumor suppressor gene and was likely inactivating.

SIGNIFICANCE: This study improves our understanding of the timing of mobile element insertions and SVs in cancer development. With most of the identified insertions and SVs arising early in development, they have the potential to act as driver mutations.

# GRAMENE SUBSITES: PANGENOME BROWSERS FOR CROPS

Joshua C Stein[1], Sharon Wei[1], Kapeel Chougule[1], Marcela Tello-Ruiz[1], Andrew Olson[1], Yinping Jiao[1], Bo Wang[1], Ivar Meijs[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

Continued advances in sequencing/assembly technologies are generating an abundance of high quality reference assemblies within crop species, ushering a transition from single-genome to pan-genome research approaches. With this transition, communities will need ready access to pre-computed comparisons of genome assemblies to identify and characterize common and variable regions. To accommodate this need, the Gramene comparative genomics project is developing Gramene subsites, each dedicated to the study of individual crop groups. We will describe the status of four pangenome subsites that support rice, maize, sorghum and grapevine. A key feature of pan-genome subsites is the application of uniform annotation protocols to minimize methodological artifacts, and the application of Ensembl and Gramene infrastructures for comparative analysis and visualization. Extending Compara gene tree output, we define conserved syntelog sets and assign conservation scores based on the proportion of genomes with membership in each set. We then score individual genomes for presence/absence and copy-number variation, additionally supported by whole genome alignments. Using related approaches in the *Oryza* genus, we showed that, compared to ancient families, recently emerged genes have higher rates of evolution, higher lability, more limited expression, prevalence in pericentromeric regions, reduced coding-length, and enrichment for stress-response functions. We gratefully acknowledge support from grants NSF#1744001, NSF#1127112, and USDA-ARS # 58-8062-7-008.

# THE ENCODE DATA ANALYSIS PIPELINES ARE PORTABLE WORKFLOWS FOR FUNCTIONAL GENOMICS

J Seth Strattan, Ulugbek K Baymuradov, Idan Gabdank, Keenan Graham, Ben C Hitz, Otto A Jolanki, Jin W Lee, Vanessa Sochat, Anshul Kundaje, J Michael Cherry

Stanford University, Department of Genetics, Stanford, CA

Results from nearly 15,000 functional genomics experiments using more than 40 assays in more than 500 different cell types and states are available for download through the ENCODE Portal. The value of this resource for integrative and comparative analysis depends on the coherence of the dataset, which, in turn, depends on the consistency of experimental protocols and analysis pipelines across labs and over time. Without such consistency technical artifacts such as batch effects can be introduced that confound subsequent integrative analyses. Bioinformatics pipelines can be a source of such effects when primary data are analyzed in different labs or at different times using different software versions, parameters, or run platforms. To address this problem, the ENCODE Data Coordinating Center (DCC) have developed a general containerization and deployment architecture that produces portable, internally-validating compute environments that run reproducibly for the analysis of core ENCODE experiments. These analysis pipelines produce the uniform ground-level annotations ENCODE applies to the epigenome and transcriptome from ChIP-seq, RNA-seq, DNase-seq, ATAC-seq, Hi-C, and whole-genome bisulfite experiments. We describe how the ENCODE uniform analysis pipelines can be run by anyone on cloud, HPC, and local compute platforms. By running these analysis pipelines on their own data, individual investigators can produce outputs that are directly comparable to ENCODE. The ENCODE DCC pipelines and architecture are free and open-source. The DCC codebase is at https://github.com/ENCODE-DCC. The pipeline architecture is demonstrated at https://github.com/ENCODE-DCC/demo-pipeline. ENCODE analyses are distributed through the ENCODE Portal at https://www.encodeproject.org/.

# THE IMPACT OF RARE GENETIC VARIATION ON DIVERSE TRANSCRIPTOMIC SIGNATURES

Benjamin J Strober*[1], Jonah Einson*[2], Nicole Ferraro*[3], Xin Li[4], Tuuli Lappalainen[5,6], Pejman Mohammadi[7], Stephen Montgomery[4,8], Alexis Battle[1,9]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Columbia University, Biomedical Informatics, New York, NY, [3]Stanford University, Biomedical Informatics, Stanford, CA, [4]Stanford University, Pathology, Stanford, CA, [5]Columbia University, Systems Biology, New York, NY, [6]New York Genome Center, New York, NY, [7]Scripps Research Institute, Integrative Structural and Computational Biology, La Jolla, CA, [8]Stanford University, Genetics, Stanford, CA, [9]Johns Hopkins University, Computer Science, Baltimore, MD

Rare genetic variation is abundant in the human genome and can contribute to disease risk, yet interpreting and predicting its functional effects remains a significant challenge for personal genomics. Recent work has shown that gene expression measurements can aid in evaluation of rare variants, and that rare variants are enriched near genes where an individual exhibits abnormal expression levels. Here, we significantly expand the use of RNA-sequencing to guide interpretation of rare variation, extracting diverse signals of abnormal transcriptomic patterns including total expression, allele-specific expression, and alternative splicing. We utilized whole genome sequencing and multi-tissue RNA-sequencing data across 714 European-ancestry individuals from the Genotype-Tissue Expression (GTEx) project. We applied probabilistic methods to identify outlier individuals for each of the three transcriptional signals, including a novel approach to detect unusual patterns of splice junction usage. Finally, we developed a general method to identify outliers specific to a subset of tissues. Compared to normal individuals, outlier individuals of all types are more likely to harbor a nearby rare variant. Furthermore, each signal is enriched for unique functional classes of rare variants. For example, alternative splicing outliers show extreme enrichment for variants disrupting or creating consensus sequences defining splice sites. Similarly, tissue-specific outliers are enriched for rare variants in enhancers in corresponding tissues. Finally, we developed Watershed, an unsupervised probabilistic graphical model that scores rare variants from a personal genome by integrating information from genomic annotations along with all three signals from the personal transcriptome, ultimately predicting the probability a rare variant has a functional effect on each transcriptional signal. Our method outperforms methods based on genome sequence alone and improves over models where each transcriptomic signal is treated independently. Overall, we provide a comprehensive analysis of the impact of rare variation on gene expression and describe an integrated framework for personal genome interpretation.

# EVOLUTIONARY ANALYSIS OF SEQUENCE DIVERGENCE OF *CIS*-REGULATORY MODULES ASSOCIATED WITH CHANGES IN CHROMATIN AND GENE EXPRESSION IN BRAIN TISSUE OF PRIMATES.

Abusaid Shaimardanov, Dmitry Svetlichnyy

National Research University Higher School of Economics, Computer Science, Moscow, Russia

Investigating gene expression evolution in brain tissue together with the divergence of genomic regulatory elements in a phylogenetic context is important to decipher genes that might shape phenotypic differences.
In this study, we have investigated evolution of the gene expression level in the context of sequence changes in the cis-regulatory elements. To this end, we performed a combined analysis of ChIP-seq [1] and RNA-seq [2] data for eight brain regions in five primates.
Using RNA-seq data for humans, chimpanzees, gorillas, gibbon, and macaques we identified genes whose expression was evolving through drift, stabilizing selection, or a lineage-specific shift. Next, we analyzed divergence of cis-regulatory signals on sequence level using the strength of the binding site for transcription factors as a quantitative trait and employed Ornstein–Uhlenbeck process. We linked obtained predictions with experimental data reflecting the presence of active chromatin mark (H3K27Ac) and identified both genomic regions and transcription factors (TFs) operating in these loci. Overall, we identified genes and predicted candidate regulatory TFs whose binding demonstrate a lineage-specific shift.

[1] Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S, Bontrop R, Cuppen E, de Laat W, Creygthon MP: Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. Nat Neurosci 2016.

[2] Chuan Xu, Qian Li, Olga Efimova, Liu He, Shoji Tatsumoto, Vita Stepanova, Takao Oishi, Toshifumi dono, Katsushi Yamaguchi, Shuji Shigenobu, Akiyoshi Kakita, Hiroyuki Nawa, Philipp Khaitovich, Yasuhiro Go: Human-specific features of special gene expression and regulation in eight brain regions. Genome Research, 2018

# HUMAN PAPILLOMAVIRUS AND THE LANDSCAPE OF SECONDARY GENETIC ALTERATIONS IN ORAL CANCERS

Maura L Gillison*[1], Keiko Akagi*[1], Weihong Xiao[1], Bo Jiang[1], Robert K Pickard[2], Jingfeng Li[2], Benjamin J Swanson[3], Amit D Agrawal[4], Mark Zucker[5], Birgit Stache-Crain[6], Anne-Katrin Emde[7], Heather M Geiger[7], Nicolas Robine[7], Kevin R Coombes[5], David E Symer*[8]

[1]University of Texas MD Anderson Cancer Center, Thoracic/ Head and Neck Medical Oncology, Houston, TX, [2]Ohio State University, Medical Oncology, Columbus, OH, [3]University of Nebraska Medical Center, Pathology, Microbiology, Omaha, NE, [4]Ohio State University, Otolaryngology- Head and Neck Surgery, Columbus, OH, [5]Ohio State University, Biomedical Informatics, Columbus, OH, [6]Complete Genomics, Mountain View, CA, [7]New York Genome Center, New York, NY, [8]University of Texas MD Anderson Cancer Center, Lymphoma & Myeloma, Houston, TX

Human papillomavirus (HPV) is a necessary but insufficient cause of a subset of oral squamous cell carcinomas (OSCC) that is increasing markedly in frequency. To identify contributory, secondary genetic alterations in these cancers, we used comprehensive genomics methods to compare 149 HPV-positive and 335 HPV-negative OSCC tumor/normal pairs. Different behavioral risk factors underlying the two OSCC types were reflected in distinctive genomic mutational signatures. In HPV-positive OSCC, the signatures of APOBEC cytosine deaminase editing, associated with anti-viral immunity, were strongly linked to overall mutational burden. By contrast, in HPV-negative OSCC, T>C substitutions in the sequence context 5'-ATN-3' correlated with tobacco exposure. Universal expression of HPV E6*1 and E7 oncogenes was a sine qua non of HPV-positive OSCC. Significant enrichment of somatic mutations was confirmed or newly identified in *PIK3CA*, *KMT2D*, *FGFR3*, *FBXW7*, *DDX3X*, *PTEN*, *TRAF3*, *RB1*, *CYLD*, *RIPK4*, *ZNF750*, *EP300*, *CASZ1*, *TAF5*, *RBL1*, *IFNGR1* and *NFKBIA*. Of these, many affect host pathways already targeted by HPV oncoproteins, including the p53 and pRB pathways, or disrupt host defenses against viral infections, including interferon and nuclear factor-κB signaling. Frequent copy number changes were associated with concordant changes in gene expression. Chr. 11q (including *CCND1*) and 14q (including *DICER1* and *AKT1*) were recurrently lost in HPV-positive OSCC, in contrast to their gains in HPV-negative OSCC. High-ranking variant allele fractions implicated *ZNF750*, *PIK3CA* and *EP300* mutations as candidate driver events in HPV-positive cancers. We conclude that virus-host interactions cooperatively shape the unique genetic features of these cancers, distinguishing them from their HPV-negative counterparts.

*These authors contributed equally to this work.

# ASSESSING THE ACCURACY OF PIPELINES FOR VARIANT CALLING, COPY NUMBER CALLING, AND SUBCLONAL DECONVOLUTION OF HETEROGENEOUS TUMOUR GENOMES

Georgette Tanner[1], David R Westhead[2], Alastair Droop[3], Lucy F Stead[1]

[1]University of Leeds, Leeds Institute of Medical Research at St James's, Leeds, United Kingdom, [2]University of Leeds, School of Molecular and Cellular Biology, Leeds, United Kingdom, [3]University of Leeds, Leeds Institute of Data Analytics, Leeds, United Kingdom

The accumulation of genetic variation that occurs in tumours, after transformation, is highly implicated in cancer progression and leads to intratumour heterogeneity, where multiple subclonal cell populations coexist within a single tumour. In some cases, the somatic variation that distinguishes a subclone may confer those cells with a selective advantage, leading to an increased growth rate, or enabling metastasis or treatment resistance. Characterising this subclonal structure is therefore critical for understanding tumour progression and for identifying novel therapeutic approaches. Such analysis is commonly performed through somatic variant and copy number calling from whole exome sequencing (WES) and subsequent subclonal deconvolution to determine the number of subclones and the variation that defines them. Whilst many methods exist for these techniques, their accuracy is confounded by the complexity of tumour genomes as well as a lack of realistic known 'ground truth' mutation profiles with which to assess them. Previous studies designed to benchmark such tools have used either real data that lacks certainty in its ground truth, or simulated data that hasn't captured the noise and complexity seen in real tumour genomes. To overcome this, we recently developed novel tools (HeteroGenesis and w-Wessim) to allow much more complex and realistic datasets to be created. In this study we use these to simulate WES datasets at 60x, 100x and 250x coverage for tumours with a range of variant and copy number frequencies, and with complex subclone architectures. We then use these datasets to benchmark entire analysis pipelines by assessing the performance of tools for calling somatic variants (Lancet, MuTect2, Varscan2) and copy number aberrations (FACETS, Sclust, Sequenza, TITAN), and for performing subclonal deconvolution (Ccube, CHAT, Pyclone, Sclust). This provides researchers with a much needed guide for a complete pipeline with which to analyse the subclonal composition of tumours, in order to avoid the high levels of inaccuracy that likely exist in many such studies.

# ASSESSMENT OF THE IMPACT ON TRANSCRIPTIONAL OUTPUT OF RARE FINE-MAPPED INTRONIC VARIANTS WITH LARGE EFFECT SIZES ON HAEMATOLOGICAL TRAITS IN THE UK BIOBANK COHORT

Manuel Tardaguila, Eve Coomber, Dragana Vuckovic, Nicole Soranzo

Wellcome Sanger Institute, Human Genetics, Cambridge, United Kingdom

Across human traits and diseases, resolving the path connecting genetic variants to molecular and cellular function and, ultimately, disease, remains to date the biggest hurdle towards unlocking the investment of early genomic studies. GWAS loci lie predominantly in non-coding regions of the DNA, where for most traits and diseases we have a limited understanding of the impact of genetic effects on the regulatory code, and of the specific cellular, developmental and environmental contexts in which they occur. An important fraction of this non-coding variation intersects the intronic regions of known genes. There is now increasing awareness of the importance of introns in transcriptional control, beyond their primary role in splicing regulation. DNA methylation of the intron closer to the Transcriptional Start Site (TSS) correlates inversely with gene expression across species and tissues[1] and Copy Number Variation (CNV) within introns significantly affect gene expression[2]. Moreover, intron-CNV eQTLs are enriched in TF binding sites[2].
Here we seek to explore the function of a subset of 74 high-effect, fine-mapped rare intronic variants impacting variation in blood cell phenotypes, and identified across the 500K participants of the UK Biobank study. To assess the effect of the variant on the transcription rate of the host gene we implement a recent procedure, genERA[3], a semi high throughput approach that leverages on the array of mutations induced by CRISPR/Cas9 after DNA damage. In the genERA framework, desired point mutations produced after Homology Directed Repair (HDR) and variable size indels resulting from Non-Homologous End-Joining (NEHJ) are analysed jointly in search of variants that cause major perturbations of their corresponding pre-spliced mRNA output at a given allele frequency. Here we present results of the first results from the genERA screen of candidate intronic causal variants. This framework provides a robust and systematic approach to tackle the trove of high impact non-coding variants with uncertain biological modes of action, and to prioritize intronic variants for subsequent deeper functional exploration.

1. Anastasiadi D. et al. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics and Chromatin (2018).
2. Rigau, M. et al. Intronic CNVs cause gene expression variation in human population. bioRxiv 171165 (2018).
3. Wu, Q. et al. In situ functional dissection of RNA cis-regulatory elements by multiplex CRISPR-Cas9 genome engineering. Nat. Commun. (2017).

# STREAMLINING ACCESSIBILITY AND COMPUTABILITY OF LARGE-SCALE GENOMIC DATASETS WITH THE NHGRI GENOMIC DATA SCIENCE ANALYSIS, VISUALIZATION, AND INFORMATICS LAB-SPACE (ANVIL)

James Taylor[1], on behalf of the AnVIL project team[2]

[1]Johns Hopkins, Biology and Computer Science, Baltimore, MD,
[2]http://anvilproject.org

The traditional model of genomic data sharing – centralized data warehouses such as dbGaP from which researchers download data to analyze locally – is increasingly unsustainable. Not only are transfer/download costs prohibitive, but this approach also leads to redundant siloed compute infrastructure and makes ensuring security and compliance of protected data highly problematic.

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space, or AnVIL, inverts this model, providing a cloud environment for the analysis of large genomic and related datasets. By providing a unified environment for data management and compute, AnVIL eliminates the need for data movement, allows for active threat detection and monitoring, and provides elastic, shared computing resources that can be acquired by researchers as needed. AnVIL will provide access to key NHGRI datasets, such as the CCDG (Centers for Common Disease Genomics), CMG (Centers for Mendelian Genomics), eMERGE (Electronic Medical Records and Genomics), as well as other relevant datasets.

The platform is built on a set of established components that have been used in a number of flagship scientific projects. The Gen3 data commons framework provides data and metadata ingest, querying, and organization. The Terra platform provides a compute environment with secure data and analysis sharing capabilities. Dockstore provides standards based sharing of containerized tools and workflows. Bioconductor and Galaxy provide environments for users at different skill levels to construct and execute analyses.

AnVIL will provide a collaborative environment for creating and sharing data and analysis workflows for both users with limited computational expertise and sophisticated data scientist users. It will provide multiple entry points for data access and analysis, including execution of batch workflows written in CWL or WDL, notebook environments including Jupyter and RStudio, Bioconductor packages for building analysis on top of AnVIL APIs and services, and Galaxy instances for interactive analysis. It will be possible to integrate additional analysis environments through standard APIs.

Long-term, the AnVIL will provide a unified platform for ingestion and organization for a multitude of current and future genomic and genome related datasets. Importantly, it will ease the process of acquiring access to protected datasets for investigators and drastically reduce the burden of performing large-scale integrated analyses across many datasets to fully realize the potential of ongoing data production efforts.

# FAMILIAL LANGUAGE HISTORY LINKS WITH POPULATION STRUCTURE AND ADMIXTURE PATTERNS

Natalie Telis[1], Hamutal Arbel[1], Elena P Sorokin[1], Ahna R Girshick[1], Miao Zhang[1], J M Granka[1], E L Hong[1], K G Chahine[2], C A Ball[1]

[1]Ancestry.com LLC, DNA, San Francisco, CA, [2]Ancestry.com LLC, DNA, Lehi, UT

Language and population structure are intricately connected in human history. Past genetic studies have found that language can roughly mirror population genetics, though the two are not exactly identical. We pilot an approach using linguistic change to approximate recent fine-scale population structure. In a large cohort with rich pedigree information, we examine the relationship between genetics and recent immigration patterns, which we approximate by identifying changes between reported grandparent and reported individual language. We identify subtle differences in ancestry composition, likely determined by both subpopulation structure and admixture patterns. We leverage maternal/paternal language information to further probe these differences in ancestry, identifying likely recent sex-biased admixture through linguistic differences. These sex-biased admixture events may impact ancestry assignment. This suggests linguistic patterns may be a reasonable proxy for substantially more complex population structure affecting different chromosomes disproportionately, and less identifiable by standard association frameworks using principal components.

# COMMUNITY CURATION STRATEGIES FOR IMPROVING GENOMIC ANNOTATIONS AND INFORMING AUTOMATED GENE PREDICTION

Marcela K Tello-Ruiz[1], Cristina F Marco[1], Kapeel Chougule[1], Andrew Olson[1], Liya Wang[1], Doreen Ware[1,2], David Micklos[1]

[1]Cold Spring Harbor Laboratory, CSHL, Cold Spring Harbor, NY, [2]USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Accurate and complete genome annotation is of paramount importance to the effective analysis of genomic and transcriptomic data, as well as their application in synthetic biology research. In spite of extensive improvements in sequencing technologies having dramatically increased the amount and quality of genome assemblies, accurate gene annotation continues to be a challenge. Gene prediction software tools remain error prone, so that for example, exon boundaries, small exons, and non-coding exons are frequently mispredicted. In addition, genes with complex alternative splice variant patterns make gene structure and length difficult to predict. Manual curation offers a solid approach to tackle this problem. Nonetheless, curatorial efforts are laborious, and constitute a bottleneck in genomics. We have evaluated distinct community curation approaches on the B73 *Zea mays* RefGen_V4 gene annotation using widely accepted annotation quality metrics such as annotation edit distance (AED), a gene tree alignment tool purposely developed to expedite flagging of models with potential annotation errors based on orthology (Gramene Tree Triage Tool), and a well-known collaborative genomic annotation editing platform (Apollo). For this colossal task, we envision the need to involve large numbers of researchers, faculty, and students. To date, we have recruited graduate students, postdoctoral researchers, and teaching faculty at Primarily Undergraduate Institutions (PUIs) – who are in turn engaging large numbers of students in authentic research – in the context of undergraduate courses. In view of the fast pace increase in the quantity of published genomes, we propose to implement community curation approaches with sustainable infrastructure and methods to ensure that the accuracy and quality of genome annotations are maintained, and will present a summary of our findings, ongoing activities, and future plans.
Sponsored by the NSF-funded MaizeCODE (IOS-1445025) and Gramene (IOS-1127112) projects.

# DUAL TRIAGE TO IDENTIFY POORLY ANNOTATED TRANSCRIPTS IN MAIZE: THE MISSING LINK IN COMMUNITY CURATION

Marcela K Tello-Ruiz[1], Cristina F Marco[1], Kapeel Chougule[1], Andrew Olson[1], Liya Wang[1], Doreen Ware[1,2], David A Micklos[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]USDA ARS, NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Accurate transcript annotation is important to the downstream analysis of genomic and transcriptomic data. The sophistication of gene prediction algorithms and abundance of RNA-based evidence for the maize genome might lead one to believe that manual curation of transcript models is no longer necessary. However, we used quality metrics generated by the Maker-P annotation pipeline to identify 9,916 of 39,498 (25%) transcript models in the B73 Reference Genome V4 with low concordance to available biological evidence.

Working with nine graduate students and one post-doctoral fellow, we used the Apollo annotation editor to curate a test set of 134 transcripts from five major gene families and a selected group of classical maize genes. Virtually all of the transcripts flagged by quality metrics and a complimentary method using the Gramene gene tree visualizer required new annotations. The majority of flagged transcripts had exons of incorrect length and/or missing exons, including about 20% that used non-canonical splice sites. Extra exons were less common. About half of the curated models matched another transcript predicted for the same gene in the V4 or V3 assembly. Virtually none of the non-flagged genes required additional curation.

The quality metrics and the Gramene gene tree visualizer identified different but partially overlapping sets of genes. Taken together, this dual triage is sensitive and specific in identifying transcripts in need of hand curation. In addition to maize researchers, we have begun to train faculty at primarily undergraduate teaching institutions (PUIs) to use our system for student research experiences. We believe this dual triage is the missing link to motivate widespread community annotation of maize and other eukaryotic genomes.

This work is part of the NSF-funded MaizeCODE (IOS-1445025) and Gramene (IOS-1127112) projects.

# DEVELOPMENT OF SEQUENCE VARIATION GRAPHS AND GRAPH-BASED SOFTWARE FOR GENOMICS STUDIES

Devin Locke, Vladimir Semenyuk, Alexey Dolgoborodov, Amit Jain, John Browning, Serhat Tetikol, Ivan Johnson, Deniz Turgut, Ozem Kalay, Duygu Kabakci

Seven Bridges Genomics, Graph Development Team, Charlestown, MA

As the pan-genome paradigm gains momentum, there are numerous challenges and design decisions concerning how to represent genomic variants in graphs, as well as how to implement tools and workflows in order to process sequencing data efficiently and effectively. We reported our initial efforts to develop an approach based on sequence variation graphs, which utilize a graph-based representation of genomic variation, along with an aligner and variant caller that take advantage of graph information to improve alignment and variant calling with next generation sequencing data (Nature Genetics, v51, pages 354–362 (2019)). Here we present advancements to our methods and performance benchmarks that show significant improvements in indel detection performance compared to non-graph or other graph-based approaches, while maintaining competitive performance in single-nucleotide variant detection, based on head-to-head comparisons using publicly available whole genome sequencing data such as Genome-in-a-Bottle. In terms of computational performance (runtime & cost), our graph-based tools are competitive with current best practice tools for whole genome sequence analysis, such as BWA & GATK4, without requiring specialized hardware due to our memory-optimized implementation. To facilitate access to the tools, we also present a Common Workflow Language whole genome sequence alignment and variant calling pipeline using publicly accessible resources such as Docker Hub, GitHub, and Dockstore.

# A NEW METHOD TO DISTINGUISH FUNCTIONAL EXPRESSION FROM BACKGROUND NOISE IN TRANSCRIPTOMES.

Ammon Thompson[1], Michael R May[1,2], Brian R Moore[1], Artyom Kopp[1]

[1]UC Davis, Ecology and Evolution, Davis, CA, [2]UC Berkeley, Ecology and Evoution, Berkeley, CA

Though all organs and cells in an organism share identical genomes, the regulatory machinery put in place during development activates a unique transcriptome in each cell type. These transcriptomes in turn determine tissue form and function. As populations evolve, genes are periodically added to and removed from transcriptomes, leading to transcriptome divergence and turnover in tissues. Ascertaining the relationship between transcriptome composition and the unique characteristics of organs and cells is an important goal of genome biology. Within a single organism, are the differences between tissues shaped primarily by quantitative differences in transcript abundance, or does qualitative presence/absence of specific transcripts play an essential role? And as organs evolve, do their functions diverge mainly through quantitative changes in gene expression, or through discrete activation and inactivation of a subset of genes? To address these questions, we developed new methods for detecting qualitative (ON/OFF) states of genes in transcriptomes.

RNA sequencing yields high resolution estimates of relative transcript abundance, but the qualitative activity state of each gene remains difficult to determine. Technical and biological noise, as well as insufficient sequencing depth, often result in detection of inactive genes in RNA-seq libraries; or conversely, cause active genes to go undetected. For many low-abundance transcripts, it is especially hard to discern whether their detection in a sample is due to genuine functional expression, or to biological or technical noise. To distinguish signal from noise, we developed a hierarchical Bayesian model to leverage data from replicate libraries to estimate the ON/OFF activity states of genes. Our method calculates the posterior probability that each gene in the genome is actively expressed in the transcriptome, given the expression patterns exhibited among replicate libraries . We demonstrate that our method provides probability estimates that are consistent with other data such as chromatin-state profiles, and is sensitive enough to accurately classify the expression states of genes known to be expressed in a small subset of cells within a complex organ. We use our method to estimate that roughly 60–70% of protein-coding genes are actively expressed in each human tissue. Finally, we use this method to examine transcriptome turnover during primate brain evolution, and show human brain transcriptomes appear to have recently gained and lost the expression of several genes.

By distinguishing between quantitative and qualitative gene expression differences, our method facilitates comparisons between tissues and species. The ability to characterize qualitative transcriptome composition and evolutionary turnover will greatly augment the analytical toolkit of functional and evolutionary genomics.

# SINGLE-CELL ISOFORM RNA SEQUENCING IN MOUSE BRAIN REVEALS ISOFORM EXPRESSION ACROSS CELL TYPES

Hagen U Tilgner

Weill Cornell Medicine, Brain and Mind Research Institute, New York City, NY

Full-length isoform sequencing has been applied to bulk tissue, cell lines and sorted cells to characterize transcriptomes but application of this technology to single cells has proven difficult, with less than 10 single cell transcriptomes analyzed. Although single splicing events have been described for <=200 single cells with statistical confidence, full-length mRNA analyses for 100s of cells has not been reported. Single cell short-read 3' sequencing enables identification of cellular sub-types, but full-length mRNA isoforms for these cell types cannot be profiled. We recently developed our new method of single-cell-isoform RNA sequencing (ScISOr-Seq), which starts with bulk tissue and identifies single-cell types and their full-length RNA isoforms without antibody-fluorescence activated cell sorting. Using single-cell-isoform-RNA-sequencing (ScISOr-Seq) we elucidated RNA isoforms in neurons, astrocytes, microglia and cell subtypes like Purkinje and Granule cells. We identify cell-type specific combination patterns of distant splice sites. We apply ScISOr-Seq to improve genome annotation in mouse Gencode version 10 by adding cell-type specific expression of 18,173 known and 16,872 novel isoforms. Thus for each gene we provide each isoform's expression across 10 cell cerebellar cell types on isoformAtlas.com. Here, I will furthermore report on recent experiments in the developmental brain and isoform expression of precise cell populations.

# DIFCOVER: A COMPUTATIONAL PIPELINE FOR IDENTIFYING STRUCTURAL VARIANTS IN COMPLEX VERTEBRATE GENOMES

Nataliya Timoshevskaya[1], Vladimir Timoshevskiy[1], Melissa Keinath[1,2], Jeramiah Smith[1]

[1]University of Kentucky, Department of Biology, Lexington, KY, [2]Carnegie Institution for Science, Department of Embryology, Baltimore, MD

Largescale copy number variants are an important source of biological variability, both between individuals (e.g. sex chromosomes) and within individuals (programmatic and sporadic deletions/duplications). Several methods exist to identify structural variants using short read datasets, however, some avenues of inquiry require the use of exceedingly fragmented reference assemblies and/or exceedingly repetitive genomes. To resolve issues associated with identifying structural variants in suboptimal reference assemblies, we developed a novel computational pipeline DifCover (https://github.com/timnat/DifCover), that addresses these issues by introducing the notion of window "stretching". Essentially each genomic scaffold is scanned sequentially to form windows of variable size, but with predefined number of bases that have depth of coverage within user-defined limits. These stretched windows allow bridging across over- and under-represented fragments permitting more precise analyses. For highly contiguous genomes, adjacent windows with similar coverage ratios can be merged to generate consensus estimates for larger continuous regions and regions with significant difference in coverage can be extracted for downstream analyses.

DifCover has been successfully tested in several complex vertebrate genomes, including those of sea lamprey, hagfish and salamander. In the highly repetitive sea lamprey and hagfish genomes, the pipeline was used to identify intervals that are eliminated from the genome during early embryonic development (via a process known as programmed genome rearrangement). Flexible tuning of pipeline parameters allowed us to initially focus first on low-copy regions in order to identify several hundred eliminated genes, and subsequently on higher-copy intervals to identify a number of highly abundant germline specific repeats. Additionally, the program has been used to identify a relatively small difference that is diagnostic of the *Ambystoma mexicanum* sex determining locus that represents ~1/100,000$^{th}$ of its 32 Gb genome and required analysis of moderate coverage data mapped to 20 million small and disjunct contigs.

# FUNDAMENTAL BIOLOGICAL PROCESSES ARE ENRICHED IN GENETIC SEX DIFFERENCES.

Michela Traglia, Lauren A Weiss

University of California, San Francisco, Department of Psychiatry and Institute for Human Genetics, San Francisco, CA

The mechanisms underlying sex differences observed in a large set of complex disorders such as autism spectrum disorder (ASD) are understudied. In ASD, evidence for higher genetic burden of rare copy number variants (CNVs) and de novo single nucleotide variants (SNVs) in affected females compared to males has been observed [Jacquemont et al., 2014]. In contrast, we did not observe any disproportionate genetic load of common SNPs from GWAS data in females [Mitra etal., 2016]. However, we found a set of SNPs with sex differential effects in anthropometric traits which also contributed to ASD and several common diseases in case-control studies, but via unknown mechanisms [Mitra et al., 2016; Traglia et al., 2017]. We also identified a significant excess of CNV mutational burden in female controls compared to male controls, particularly CNV categories most likely to be deleterious [Desachy et al., 2015]. However, exome and genome sequencing data has not yet been examined for sex differences in mutational burden. In addition, specific biological pathways or functions underlying genetic sex differences could be informative.

We aimed to characterize (1) key pathways for sex differential SNPs and (2) sex differences in autosomal mutational burden of low frequency variants in control populations.

We found that sex heterogeneous SNPs are significantly enriched in eQTLs (emp $P < 0.01$), suggesting they are functionally important. These SNPs are enriched in crucial signaling pathways and biological processes for early development, involving forebrain and neuronal function as well as mammary gland development and hormone-related pathways relevant for fertility and reproduction (FDR<0.05; emp $P < 0.05$).

Secondly, we observed an excess of autosomal mutational burden of low frequency SNVs predicted to be loss-of-function and missense variants in female controls compared to male controls (Fisher test $P < 10^{-10}$) from gnomAD datasets [Karczewski et al., biorxiv 2019]. These observations will allow us to set a baseline to refine the excess loss-of-function mutations observed in ASD females, to assess sexual dimorphism for constraint at specific loci and in noncoding functional variation in ASD.

Our findings showed an essential link between genetic sex differences, fundamental developmental processes and pathophysiology. We also observed sexually dimorphic genetic variation in rare sites predicted to be deleterious in human populations that can used as expected sex differences in case-control studies design as well as for a better understanding of clinical and research results.

# SEARCHING FOR FERTILITY GENES USING ASSOCIATIONS WITH CARDIOVASCULAR DISEASE

Holly Trochet[1], Jean-Christophe Grenier[1], Marie-Julie Favé[1,2], Philip Awadalla[2], Julie Hussin[1]

[1]Institut de Cardiologie de Montréal, Université de Montréal, Montréal, Canada, [2]Ontario Institute for Cancer Research, University of Toronto, Toronto, Canada

Fertility is a direct target of natural selection, making it challenging to find genes and variants that have direct effects on it. However, fertility traits in women have been correlated with cardiovascular disease (CVD), which comprises disorders whose genetic architectures are much better characterized. Examples include (but are not limited to) the negative correlation between age at natural menopause and risk of heart disease later in life and the positive correlation between multiple miscarriage and risk of ischemic heart disease. However, the literature often lacks biological explanations for these observed relationships. Using data from UK Biobank as well as biobank cohorts gathered in Canada (the Montreal Heart Institute Biobank, the Canadian Partnership for Tomorrow Project, and CARTaGENE), we have replicated these epidemiological associations and are investigating their shared biological bases, with a particular focus on genetics.

Because of the diversity of and heterogeneity in the our datasets, we have unique opportunities to use a variety of tools to conduct our investigation. In particular, we seek to leverage pleiotropy between fertility and CVD variants in order to find associations. For instance, by computing genetic risk scores, we have established the shared genetic bases between the two classes of traits, and conditional analyses allow us to map loci that show cross-trait associations. Additionally, we use a number of statistical models that aim to detect cross-trait associations from GWAS summary statistics. Using these approaches, we show the shared genetic links between coronary artery disease and timing of menarche and menopause as well as ischemic stroke and stillbirth. Our results strongly suggest a shared genetic basis between CVD risk and fertility traits, and demonstrate the promise of approaches that take advantage of pleiotropy and evolutionary relationships in establishing the genetic architecture of fertility traits in humans.

# A NOVEL, POPULATION-ROBUST, SMALL-COHORT RISK STRATIFICATION METHOD IDENTIFIES PATIENTS WITH 13-FOLD RISK FOR CARDIOVASCULAR DISEASE

Alex M Tseng*[1], Yosuke Tanigawa*[2], Cole A Deisseroth[1], Gill Bejerano[1,2,3,4]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Biomedical Data Science, Stanford, CA, [3]Stanford University, Developmental Biology, Stanford, CA, [4]Stanford School of Medicine, Pediatrics, Stanford, CA

*These authors contributed equally

Genetic risk stratification has great potential in early screening of and preventative therapy for common complex traits. Most existing methods, however, rely on summary statistics from very large-scale genome-wide association studies (GWAS) on the same ethnic cohort with hundreds of thousands of people, limiting the transferability of the model across populations. Moreover, most models ignore the contributions of ultra-rare variants, and the limited interpretability of non-coding marker SNPs hinders the identification of therapeutic leads. To address these challenges, we developed a novel risk stratification method based on in-depth analysis of personal regulatory variants derived from n=1 whole genomes and assessed their functional implications. By comparing the mutation burdens across a given case-control cohort with only hundreds of individuals, the method robustly identifies the most discriminating phenotype, as well as individuals with an elevated risk.

We applied our method across cohorts of different ethnicities and diverse traits. Without ever needing cohort- or trait-specific parameters, we identified the most relevant phenotypes and stratified individuals by our novel risk score. For example, in a cardiovascular disease cohort, we found a 13-fold increase in risk in the top 5% of individuals, which is a drastic improvement in predictive ability over polygenic risk scores derived from large-scale GWAS with orders of magnitudes more participants. Across different cohorts, our method is shown to be effective and robust on different populations. We further highlight the ability of our method to provide actionable mechanistic insights into the variants underlying disease.

Combined with the rapidly growing sequencing efforts of multiple populations around the world, our robust and interpretable risk stratification method has an immediate impact on genomic-based precision health.

# "SEMIBULK" RNA-SEQ ANALYSIS FOR REVEALING LOCAL CELLULAR ENVIRONMENTS

Issei Tsuchiya, Satoi Nagasawa, Seki Masahide, Yuta Kuze, Yutaka Suzuki

The University of Tokyo, Department of Computational Biology and Medical Sciences, Chiba, Japan

Recent technical developments have enabled us to obtain a comprehensive catalogue of the genes and proteins which are different between diseased and normal organs at a single-cell level. Single-cell analysis has been widely used in variety of fields, such as cancer, immunology, neurobiology sciences and so on. Characterizing the transcriptome of individual cells is essential for understanding complex biological systems. And this technique has enabled us to dissect transcriptomic heterogeneity which, in the past, when bulk analysis was done, was difficult to analyze.

However, there are disadvantages in the current single-cell analysis. That is, when conducting single-cell analysis, variety of cells which are not linked in terms of local cellular environments are mixed together. Therefore, it is difficult to acquire data of cell groups which are locally close to each other. It is known that surrounding cells play significant roles. For example, cancer-associated fibroblasts (CAFs) lead angiogenesis, and promote an increase in carcinoma cells. Indeed, a number of studies have recently been aware of the importance of analyzing local cellular environment. It is impossible to fully understand the behavior of the cell groups without recognizing where the cells belong in a tissue or which cell group interacts with another cell group.

I implemented an experiment and analysis pipeline for the “Semibulk” RNA-Seq, which would enable us to examine local cellular environments. Characteristic relation between cells that are locally close to each other, which would not be revealed by transcriptome analysis of bulk cells nor single cells, may play essential roles in revealing the network between the cells, the structure of organs, and the cause of diseases.

# MATHEMATICAL MODELING OF ANEUPLOID CONCEPTION RISK

Katarzyna M Tyc[1], Rajiv C McCoy[2], Karen Schindler[1], Jinchuan Xing[1]

[1]Rutgers, The State University of New Jersey, Department of Genetics, Piscataway, NJ, [2]Johns Hopkins University, Department of Biology, Baltimore, MD

Approximately 10% of women world-wide experience difficulties in conceiving and/or staying pregnant. The mother's age is an established contributing factor to this difficulty, and determining genetic risk factors is an active area of research. Aneuploidy is a major determinant of infertility, recurrent pregnancy loss, and *in vitro* fertilization (IVF) failure. To increase the chances of implantation, IVF cycles are often coupled with embryo biopsy to preselect embryos for uterine transfers. Results of IVF cycles coupled with preimplantation genetic screening (PGS) of embryos are often suggested when counseling infertile patients. This counseling, however, is based on expected aneuploidy rates given the mother's age, hormonal profile and the number of eggs harvested. Although most aneuploid conceptions are thought to originate from female meiotic division errors, quantitative studies that would formally link the observed phenotypes to underlying error mechanisms are lacking. The goal of this project is to unravel mechanisms of erroneous chromosome segregations underlying the production of poor quality eggs.

Using mathematical modeling we investigate the rates of chromosomal mis-segregations that ultimately lead to aneuploid conceptions. In the model, the probabilities of all gamete modes are summed to give an overall probability of producing either euploid or aneuploid embryo. We consider the rates of non-disjunction (ND) in both meiotic divisions and premature separation of sister chromatids (PSSC) in Meiosis I (MI). The rates are parameterized based on published cases of IVF-PGS cycles and analysis of blastomeres and trophectoderm cells. In the process, we find best solution that minimizes the distance between model simulations and proportions of various aneuploid embryos. With only five parameters we can model all of the possible conception outcomes that approximates the observed data. Using the simulations of the model we assessed how specific error types contribute to observed chromosomal abnormalities in existing human data sets. Specifically, we observed PSSC to be more prevalent than ND in MI across all the ages–a result that has long been speculated. We used the model to identify the patients that are outliers in the population with regard to the number of aneuploid embryos. Overall, the dynamic nature of the mathematical model makes it a potential tool in predicting the outcomes of assisted reproduction. This in turn could help to improve implantation rates which have lingered at 30% for the past 20 years.

# RARE AND COMMON VARIANTS IN THE GENETIC ARCHITECTURE OF INTELLECTUAL DISABILITY IN NORTHERN FINLAND

Lea Urpa[1], Mitja Kurki[1,2,3], Elmo Saarentaus[1], Olli Pietiläinen[2,4], Eija Hämäläinen[1], Padhraig Gormley[2,3], Minna Torniainen-Holm[1,5], Veikko Salomaa[5], Dennis Lal[2,3], Sini Kerminen[1], Matti Pirinen[1,6], Jaana Suvisaari[5], Elisa Rahikkala[7,8,9], Riikka Keski-Filppula[7,8,9], Merja Rauhala[10], Satu Korpi-Heikkilä[10], Jonna Komulainen-Ebrahim[7,11], Heli Helander[7,11], Päivi Vieira[7,11], Jukka S Moilanen[7,8,9], Jarmo Körkkö[10], Outi Kuismin[1,7,8,9], Mark J Daly[1,2,3,12], Aarno Palotie[1,2,3,12]

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Human Genetics, Helsinki, Finland, [2]Broad Institute of MIT and Harvard, The Stanley Center for Psychiatric Research, Cambridge, MA, [3]Massachusetts General Hospita, Psychiatric and Neurodevelopmental Genetics Unit, Boston, MA, [4]Harvard University, Department of Stem Cell and Regenerative Biology, Cambridge, MA, [5]National Institute for Health and Welfare, Helsinki, Finland, [6]University of Helsinki, Helsinki Institute for Information Technology HIIT,Department of Public Health, Department of Mathematics and Statistics, Helsinki, Finland, [7]University of Oulu, PEDEGO Research Unit, Oulu, Finland, [8]Oulu University Hospital, University of Oulu, Medical Research Center, Oulu, Finland, [9]Oulu University Hospital, University of Oulu, Department of Clinical GeneticsOulu, Finland, [10]Northern Ostrobothnia Hospital District, Center for Intellectual Disability Care, Oulu, Finland, [11]Oulu University Hospital, Oulu University, Department of Children and Adolescents, Oulu, Finland, [12]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Department of Medicine, Department of Neurology, Boston, MA

Genomic information is a key tool in understanding complex neurological disorders such as intellectual disability (ID). Discovering causative variants implicated in ID can facilitate diagnosis and treatment of patients, currently complicated by phenotypic and genetic heterogeneity. The Northern Finland ID (NFID) study aims to identify inherited variants contributing to ID in patients from a population subisolate. Finland has a small founder population that has undergone several bottlenecks and genetic drift, particularly in late-settlement Northern Finland. Simulation studies show such populations have higher frequencies of variants impacted by negative selection than older populations: in Finland, there is an increase in the frequency of loss-of-function (LoF) variants compared to other European populations. In Northern Finland, higher prevalence of ID suggests inherited factors contribute to ID risk there. Previous work by Kurki et al found in NFID an enrichment of protein-truncating variants in LoF intolerant genes. The authors also found a common variant load in ID, where polygenic risk scores for educational attainment, IQ, and schizophrenia were associated with ID, explaining 0.9%, 0.5%, and 0.6% of heritability on the liability scale. We present the extended analysis of NFID with genotype and exome sequence data from 989 patients (57% mild ID) and 941 family members, including 77 multiplex families. Unlike most ID studies, we focus on inherited variants as our cohort of patients from a population subisolate, with predominantly mild ID and multiplex families increases power to identify such variants. Top variants will be further investigated in the FinnGen cohort of 150,000 Finns (www.finngen.fi/en).

# WHOLE EXOME SEQUENCING AND CHARACTERIZATION OF CODING VARIATION IN 49,960 INDIVIDUALS IN THE UK BIOBANK

Cristopher Van Hout[1], Ioanna Tachmazidou[2], Joshua Backman[1], Joshua Hoffman[3], Bin Ye[1], Ashutosh Pandey[3], Claudia Gonzaga-Jauregui[1], Nilanjana Banerjee[1], Suganthi Balasubramanian[1], David Ledbetter[4], Jeffrey Reid[1], Matthew Nelson[3], Jonathan Marchini*[1], John Overton*[1], Robert Scott*[3], Goncalo Abecasis*[1], Laura Yerges-Armstrong*[3], Aris Baras* on behalf of the Regeneron Genetics Center[1]

[1]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, [2]GlaxoSmithKline, Human Genetics, Brentford, United Kingdom, [3]GlaxoSmithKline, Human Genetics, Philadelphia, PA, [4]Geisinger Health System, Genomic Medicine Institute, Danville, PA

*Contributed equally.

The UK Biobank is a prospective study of 502,543 participants, with extensive phenotypic and clinical information and streamlined data access for researchers. Here, we sequenced the exomes of 49,960 participants and describe the first tranche of large-scale exome sequence data in the UK Biobank study. Our work greatly enhances available genetic variation data with 3.8 million variants in protein coding sequences (~98.5% of these have frequency <1%), a nearly 10-fold increase in characterized coding variants compared to imputed sequence.

In the enhanced resource, we observed 235,262 predicted loss of function (LOF) variants, and approximately 69% of autosomal genes with at least 10 LOF heterozygous carriers. In imputed sequence in the same set of ~50k and all ~500k UK Biobank participants, this threshold was met for 23% and 35% of genes, respectively. We found that 2% of sequenced participants carry medically actionable variants in the ACMG59 genes; ~55% of such variants are previously reported pathogenic or likely pathogenic in ClinVar (≥2 star), whereas the remainder are novel predicted LOF variants. We characterize the relationship between cancer risk and *BRCA1/2* variation and show that both reported and novel variants increase risk of multiple cancers. Additionally, we replicate previously known loss of function associations, and demonstrate the role of rare variation in the genetic etiology of varicose veins, corneal resistance, and bone mineral density.

This large widely accessible resource of sequenced exomes linked to de-identified health records and extensive longitudinal measures greatly enhances and extends the current genetic resources, particularly in the ascertainment of rare coding variation. Efforts are underway to sequence the exomes of all UK Biobank participants; these efforts will greatly expand the total amount of rare coding variation that can be studied and will continue to power many important findings and advances.

# TUMOR SHEDDING INTO CELL-FREE DNA (CFDNA) IS ASSOCIATED WITH HIGH-MORTALITY CANCERS

Oliver Venn[1], Earl Hubbell[1], Onur Sakarya[1], Christopher Chang[2], Meredith Miller[3], Kristan Steffen[4], Eric T Fung[5], Anne-Renee Hartman[5], Tara Maddala[6], Alex M Aravanis[4]

[1]GRAIL, Inc., Bioinformatics, Menlo Park, CA, [2]GRAIL, Inc., Software Engineering, Menlo Park, CA, [3]GRAIL, Inc., Clinical Laboratory, Menlo Park, CA, [4]GRAIL, Inc., Research & Development, Menlo Park, CA, [5]GRAIL, Inc., Clinical Development, Menlo Park, CA, [6]GRAIL, Inc., Biostatistics, Menlo Park, CA

Previous studies indicated that circulating tumor DNA (ctDNA) fraction varies among cancer types and individuals (Abbosh et al., 2017). However, these studies have been limited by small sample size. ctDNA fraction distribution across stages in 24 cancer types was evaluated in the Circulating Cell-free Genome Atlas (CCGA, NCT02889978).

Matched tumor biopsies and white blood cells (WBC) from a subset of CCGA cancer participants (pts, 692/1628) were subjected to whole-genome sequencing (WGS, 30x); results were compared to targeted cfDNA sequencing (507 genes; 60,000x) to estimate ctDNA fraction from single-nucleotide variants/indels. 24 tumor types were surveyed (323 breast, 54 colorectal, 53 lung, 47 prostate, 31 uterine, 30 renal, 20 esophageal, 15 lymphoma, and 119 samples with <15 samples per cancer type). Independent of tumor information, cfDNA classifiers of cancer vs non-cancer for whole-genome bisulfite sequencing (WGBS, 30x) and WGS were trained then evaluated on a test set from a pre-specified case-control CCGA substudy (1628 cancer; 1172 non-cancer).

Positive evidence of ctDNA (at least one alternative allele read, typical molecular depth 3000) in targeted cfDNA sequencing was observed in 420 of 692 pts. ctDNA fraction was an important covariate of classification performance, explaining >99% of score variance for WGBS and WGS classifiers. Classifiers' limits of detection (LOD, ctDNA fraction with 50% detection probability at 98% specificity) were 0.7% (±0.07%) for WGBS vs 2% (±0.19%) for WGS.

ctDNA fraction increased with stage across cancer types. High mortality cancers (bottom 10th percentile of incidence weighted Surveillance, Epidemiology, and End-Results [SEER] 5-year survival: esophageal, gastric, hepatobiliary, lung, pancreas) had median ctDNA fractions of 0.07% (bootstrap q5, q95: 0.05%, 0.48%) in stage I, 0.28% (0.06%, 0.6%) in stage II, 4.8% (1.6%, 15%) in stage III, and 11% (2.7%, 21%) in stage IV. By contrast, ctDNA fraction was lower among high survival cancers (top 90th percentile: breast, prostate, thyroid). The ctDNA ratio between low survival and high survival was 1-fold (bootstrap q5, q95: 1, 12) for stage I, 4-fold (1, 9) for II, 14-fold (6, 58) for III, and 5-fold (2, 45) for IV.

ctDNA fraction was strongly associated with cfDNA classifier performance. Tumor shedding combined with LOD of a given classifier and assay drives the observed differences in detection among cancer types; specifically, cancers with high mortality appear to have higher cfDNA fraction. Together, these data suggest that cfDNA-based assays may detect these cancers at earlier stages.

Funding: GRAIL, Inc.

# LARGE SCALE GENOMIC TESTING FOR DISCOVERY IN ADULT MEDICINE

Eric Venner[1], David Murdock[1], Tsung-Jung Wu[1], Sara Kalla[1], Victoria Yi[1], Yunyun Jiang[1], Kimberly Walker[1], Jianhong Hu[1], Viktoriya Korchina[1], Christie Kovar[1], Stacey Pereira[1], Christie Ballantyne[1], Mullai Murugan[1], Ginger Metcalf[1], Eric Boerwinkle[1,2], Donna M Muzny[1], Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]UT Health Science, School of Public Health, Houston, TX

Sequencing genomes at a large-scale, reporting genetic results in a clinical setting, and integrating this data with Electronic Medical Records (EMR) are driving precision medicine. Together with cloud-based research tools that operate on these datasets, these efforts form a virtuous cycle in which data generation drives more discovery, and thus clinical utility, which then drives more data generation.

We have developed a suite of tools and an accompanying support infrastructure that allows a seamless flow of samples and data between multiple genetic testing centers and clinical sites, allowing the delivery of thousands of clinical genomic reports into the EMR, as well as creating a research resource that supports secure data mining. This work provides a case study for a range of precision medicine implementation projects, including the NIH eMERGE and All of Us Programs.

The combined laboratory and analytical methods enable rapid data generation, variant annotation, diagnostic interpretation and generation of clinical reports using capture panels within a CAP/CLIA environment. Reports are returned to participants through genetic counsellors at the clinical sites, and a structured version of reports is available for communication with the EMR. Work is ongoing to adapt this format to the FHIR standard. The clinical reporting process is highly automated to support batch review and approval of large sets of clinical reports by a laboratory director. Reports can include pharmacogenomics, copy-number variants and pathogenic or likely pathogenic variants from a select set of disease genes. Genomic data are aggregated with rich phenotype and other participant metadata by a central Data Resource Center, for use by the wider research community.

Most recently we been working to efficiently integrate these methods into clinical practice. Using a clinical test, called HeartCare, we will screen individuals for risk of cardiovascular disease, the number one cause of adult mortality. This test is implemented through a gene panel of 158 genes related to cardiovascular disease, plus sites for pharmacogenomics analysis and generation of a polygenic risk score. During a pilot study, samples will be collected at clinics throughout the Texas Medical Center. Key outcomes will be the clinical utility of the pathogenic variants, pharmacogenomic findings and polygenic risk score, as well as the impact on the clinical workflow.

# LEVERAGING EVOLUTIONARY DRIVERS OF GENETIC VARIATION REVEALS COMPLEXITY UNDERLYING DISEASES ACROSS eMERGE EPISTASIS STUDY OF 100,000 SAMPLES

Shefali S Verma[1], Pankhuri Singhal[1], Steven J Schrodi[2], Sarah Pendergrass[3], Daniel Schaid[4], Iftikhar Kullo[4], David Fasel[5], Murray Brilliant[2], Ozan Dikilitas[4], Chunhua Weng[5], Hakon Hakonarson[6], Patrick Sleiman[6], Scott Hebbring[2], Alexander T Frase[1], Scott Dudek[1], Anurag Verma[1], Marylyn D Ritchie[1]

[1]University Of Pennsylvania, Department of Genetics, Philadelphia, PA, [2]Marshfield Clinic Research Institute, Center for Precision Medicine Research, Marshfield, WI, [3]Geisinger, Danville, PA, [4]Mayo Clinic, Rochester, MN, [5]Columbia University, Department of Biomedical Informatics, New York, NY, [6]Children's Hospital of Pennsylvania, Center of Applied Genomics, Philadelphia, PA

Evolution is a product of random and non-random associations that help shape the human genome. Linkage disequilibrium (LD) is a driver of non-random associations that occur at a greater frequency than by chance. Distinctive patterns across chromosomes due to LD can be the result of mutation, gene flow, hitchhiking, genetic drift, or epistatic selection. Capitalizing on LD patterns for determining population substructure and discovering additive association signals is the basis of GWAS. While GWA-studies are well powered to detect additive effects, the underlying assumption of an additive model often neglects alternative modes of inheritance such as dominance and epistasis. Epistasis, or the interaction between genes, exists across the genome but it often goes undetected due to lack of statistical power. Furthermore, the adoption of LD pruning as customary excludes detection of linked LD sites that may in fact underlie the genetic architecture of complex traits. We hypothesize that uncovering long-range LD patterns within haplotype blocks that exist due to epistatic selection could help in elucidating genetic mechanisms attributing to common diseases. To test this hypothesis, we identified 132K models in long-range LD (Ohta stats) in the eMERGE dataset with 95,417 samples. Next, to validate Ohta's D approach, we tested for evidence of statistical epistasis using MDR to identify models associated with disease status. Among the top results, we identified variants in STXBP5L gene associated with osteoarthrosis, senile macular degeneration of retina and viral infections. Our results also inform several pleiotropic associations across chromosome 11. Functional analyses of HiC data suggest that the significant models identified by statistical epistasis coincide with regulatory regions of physically-interacting chromatin. Our statistical epistasis models demonstrate that changes in genomic arrangement due to evolutionary mechanisms drive complexity underlying disease which cannot be identified by additive effects alone. Identifying and testing long-range LD regions in genome-wide analyses can bolster epistasis detection and enhance our understanding of disease etiology.

# HOMOZYGOUS *RET* LOSS-OF-FUNCTION LEADS TO PRECOCIOUS DIFFERENTIATION AND CHANGES IN CELL FATE OF ENTERIC NEURAL CREST CELLS

Elizabeth Vincent[1], Sumantra Chatterjee[2], Gabrielle Cannon[3], Dallas Auer[1], Holly Ross[4], Aravinda Chakravarti[2,5], Loyal Goff[1,6]

[1]Johns Hopkins School of Medicine, Institute of Genetic Medicine, Baltimore, MD, [2]NYU Langone, Center for Human Genetics and Genomics, NYC, NY, [3]UNC Chapel Hill, Center for GI Biology and Disease, Chapel Hill, NC, [4]University of Maryland School of Dentistry, Neural and Pain Sciences, Baltimore, MD, [5]NYU Langone, Institute for Systems Genetics, NYC, NY, [6]Johns Hopkins School of Medicine, Neuroscience, Baltimore, MD

The enteric nervous system (ENS) is formed by enteric neural crest cells (ENCCs), which migrate through the gut and produce neurons and glia. In Hirschsprung disease (HSCR), part of the ENS fails to form. ~50% of HSCR cases are attributed to loss of function (LOF) of *RET*, a receptor expressed by all ENCCs; however, the mechanism remains unresolved. We hypothesized that *RET* LOF leads to precocious differentiation of ENCCs, depletion of the progenitor pool, and changes in cell fate. To test our hypothesis, we performed scRNAseq on ENCCs in the gut of $Ret^{+/-}$ and $Ret^{-/-}$ mice at E12.5 and E14.5, the midpoint and endpoint of ENS development. Dimensionality reduction revealed multiple developmental trajectories. Mature cell type markers were insufficient to annotate developmental data; therefore, we applied coordinated gene activity in pattern sets (CoGAPS) to identify patterns of gene co-expression in published data from the adult mouse ENS. Classification using transfer learning identified proliferating progenitors, maturing glia, and neuronal subtypes. Differential gene expression analysis revealed the effects of *Ret* LOF are cell-type-specific. $Ret^{-/-}$ glia cells express mature markers while $Ret^{+/-}$ glia cells express cell cycle genes. Additionally, the proliferating progenitor population contains cells of both genotypes at E12.5, but only $Ret^{+/-}$ cells at E14.5, suggesting depletion of the progenitor pool. In the neuronal population, the nitrergic trajectory is exclusively comprised of $Ret^{+/-}$ cells and is inaccessible to $Ret^{-/-}$ cells. Another neuronal trajectory is accessible to both genotypes and exhibits downregulation of the *Ret* gene regulatory network as well as expression of a non-canonical Ret co-receptor, suggesting the absence of *Ret* restricts competency to fates that use alternative, *Ret*-independent signaling pathways. These data demonstrate a role for *Ret* in regulation of progenitor proliferation, differentiation, and cell fate determination.

# NON-ADDITIVE GENETIC EFFECTS ON GENE EXPRESSION MODULATED BY CELL HETEROGENEITY IN WHOLE BLOOD

A Viñuela[1], A A Brown[1], A Martinez-Perez[2], N I Panousis[1], A Ziyatdinov[2], M Sabater-Lleal[2], A Hamsten[3], J C Souto[2], A Buil[4], J M Soria[2], E T Dermitzaki[1]

[1]University of Geneva, Geneva, Switzerland, [2]Inst. Investigaciones Biomedicas Sant Pau, Barcelona, Spain, [3]Karolinska Institutet, Stockholm, Sweden, [4]Institute of Biological Psychiatry, Roskilde, Denmark

Thousands of genetic variants have been reported to influence blood cell counts with implications for disease risk. To investigate the cell specific nature of gene expression and its additive and non-additive regulation, we quantified cell counts and RNAseq in blood in 915 individuals from 35 pedigrees. We identified 154,558 significant associations with 14 blood phenotypes across 16,726 genes (99.8%), ranging from 910 genes associated with basophils ($\Pi_1$= 0.26) to 16,679 genes with lymphocytes ($\Pi_1$= 0.94). Enrichment analysis showed more than 50% of the genes were associated with multiple cell types ($\Pi_1$ ranges from 0.50 to 0.98). Next, we compared heritability ($h^2$) estimates of expression before and after controlling for cell counts. Since cell type abundance is heritable, we expected cell counts to contribute to the genetic component, decreasing $h^2$. However, we observed a significant increase (from mean $h^2$ = 0.192 to $h^2$ = 0.201, Pvalue = 4.36e-05) suggesting cell composition plays a more important role in the environmental component than the genetics.

To investigate non-additive regulation, we first looked genetic effects dependent of cell composition (cell-type x genotype interactions) and identified 1,311 significant interactions in 636 genes across the 14 cell types. The interactions were often shared, for 53.45% of the genes (N=340) interactions with the same lead SNPs (eSNP) or another in high LD ($R^2$> 0.8) were discovered for multiple cell types. However, for 11 genes multiple interactions involving different loci ($R^2$<0.4) were found. Second, we investigated whether parent-of-origin effects were also influenced by cell-type abundance. Focusing on known imprinting genes (N=195), we look for SNPs that act in a parent-of-origin dependent manner (PO-eQTL) and also interact with cell-type abundance. We detected 30 genes with PO-eQTLs and cell-types interactions. For 16 genes, the same or a nearby eSNP ($R^2$>0.8) were significant for both effects, including *SNURF* (lymphocytes and neutrophils), *NOTCH3* (monocytes), or *CAST* (basophils); while for 14 genes other loci than the PO-eSNPs (mean $R^2$<0.2) were significantly interacting.

In conclusion, the influence of genetic variation on expression is affected by a combination of factors such as cell composition and allelic inheritance. Considering these factors together may be crucial to understand the mediating role of expression between genetics and disease.

# WHY DETAILS MATTER – THE ROLE OF ALTERNATIVE SPLICING IN GLIOBLASTOMA CANCER STEM CELLS

Kristoffer Vitting-Seerup[1,2], Yi C Lim[1], Albin Sandelin[2], Petra Hamerlik[1,3]

[1]The Danish Cancer Society, Brain Tumor Biology Lab, Copenhagen, Denmark, [2]University of Copenhagen, Department of Biology, Copenhagen, Denmark, [3]Copenhagen University, Department of Drug Design and Pharmacology, Copenhagen, Denmark

Glioblastoma (GBM) is one of the deadliest types of cancers with have a median survival rate of approximately 15 months. This poor outcome is in-part explained by the radio-resistant properties of GBM as well as the rapid recurrence of the disease. One of the key factors responsible for both the prompt onset of recurrence and radioresistance is the GBM cancer stem cells (GSCs) population. Unfortunately, current knowledge on GSCs is limited as past studies have focused on using pre-defined cell surface markers from other cancer types to determine GSCs. Here, we used GBM single cell RNA-sequencing data to, in an unbiased manner, identify and characterize GSCs. These results offer new insights in both the development and maintenance of GBM. As we have previously shown alternative splicing plays an important role in solid cancers, we developed a new algorithm capable of identifying differential isoform usage from single cell RNAseq data. This allowed us to identify isoform switches that are predicted to give rise to proteins with different functional properties, due to differences in amongst others protein domains. When applied to the single cell GBM data, we can identify substantial differences in which isoforms are used in GSCs and non-GSCs, indicating alternative splicing plays a vital role in gliogenesis.

# MUTATIONAL SIGNATURES ARE SHAPED BY DNA DAMAGE AND REPAIR

Nadezda Volkova*[1], Bettina Meier*[2], Peter J Campbell[3], Anton Gartner[2], Moritz Gerstung[1]

[1]EMBL-EBI, European Bioinformatics Institute, Cambridge, United Kingdom, [2]University of Dundee, Centre for Gene Regulation and Expression, Dundee, United Kingdom, [3]Wellcome Trust Sanger Institute, Cancer Genome Project, Cambridge, United Kingdom

Mutagenesis occurs when DNA damage or DNA synthesis errors fail to be repaired by the cell's DNA repair pathways. To delineate the contributions of DNA damage and repair we sequenced the genomes of 2,776 nematode lineages subjected to 196 combinations of 12 genotoxins and 63 different knockouts of DNA repair genes.

In about 2/3 of experiments combined genotoxicity and DNA repair deficiency changes mutagenesis, generally increasing the numbers of mutations to superlinear levels with mostly moderate shifts of the mutation types. Knockouts of translesion synthesis polymerases *rev-1*, *rev-3* and *polh-1*, however, reduce the mutation load under UV damage or DNA alkylation, replacing single nucleotide variants with deletions longer than 50bp. Alkylating agents display diverse changes in the mutation spectrum in response to knockouts of alkyl transferases *agt-1* and translesion synthesis polymerase *polk-1*, reflecting alternative base-specific repair mechanisms.

Reanalysing data from 9,946 cancer samples from The Cancer Genome Atlas demonstrates the significance of DNA repair component status, especially translesion synthesis polymerases, to the mutational burden of the tumour and appearance of the mutational signatures including that of APOBEC mutagenesis.

# GENETIC CONTROL OF EXPRESSION AND SPLICING IN DEVELOPING HUMAN BRAIN INFORMS DISEASE MECHANISMS

Rebecca L Walker[1], Gokul Ramaswami[1], Christopher Hartl[1], Nicholas Mancuso[2], Michael J Gandal[1,3,4], Luis de la Torre-Ubieta[1,4], Bogdan Pasaniuc[2,3], Jason L Stein[5], Daniel H Geschwind[1,3]

[1]University of California, Los Angeles, Neurology, Los Angeles, CA, [2]University of California, Los Angeles, Pathology and Laboratory Medicine, Los Angeles, CA, [3]University of California, Los Angeles, Human Genetics, Los Angeles, CA, [4]University of California, Los Angeles, Psychiatry, Los Angeles, CA, [5]University of North Carolina, Chapel Hill, Genetics, Chapel Hill, NC

Most genetic risk for human diseases lies within non-coding regions of the genome, which are predicted to regulate gene expression in a tissue and stage specific manner. This has motivated building of extensive quantitative trait loci (QTL) resources across tissues. Given the importance of regulatory pathways during brain development, we characterized the genetic control of the developing human cerebral cortical transcriptome in 201 mid-gestational human brains, identifying 7962 expression (eQTL) and 4635 splice QTL (sQTL). We leveraged these eQTL and sQTL to identify splicing and transcriptional drivers affected by human genetic variation. By integration with GWAS, we characterized the genes and isoforms contributing to specific neuropsychiatric disorders, including SCZ and ASD, as well as other cognition or behavioral-related phenotypes, such as brain size. SCZ and ASD impact distinct developmental gene co-expression modules, yet in both disorders, common and rare genetic variation converge. These analyses demonstrate the highly distinctive effects of splicing and transcriptional control, as well as divergent age-related contributions to disease. More broadly, these findings demonstrate the genetic mechanisms by which early developmental events have a striking and widespread influence on adult anatomical and behavioral phenotypes.

# HAPLOTYPING USING FULL-LENGTH TRANSCRIPT SEQUENCING IN MAIZE F1 HYBRID REVEALS ALLELE-SPECIFIC EXPRESSION

Bo Wang[1], Elizabeth Tseng[2], Primo Baybayan[2], Kevin Eng[2], Michael Regulski[1], Yinping Jiao[1], Andrew Olson[1], Liya Wang[1], Kapeel Chougule[1], Peter Van Buren[1], Doreen Ware[1,3]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY, [2]Pacific Biosciences, Menlo Park, CA, [3]USDA ARS, Robert W. Holley Center for Agriculture and Health Cornell University, Ithaca, NY

Maize is an important genetic model for elucidating transcriptional networks. Haplotype phasing of genetic variants is important for interpretation of the maize genome, population genetic analysis and functional genomic analysis of allelic activity. Recently, full-length transcript sequencing using long read technology has enabled us to characterize alternative splicing events and improve the maize genome annotation. However, the general Iso-Seq algorithm ignores SNP-level information, focusing instead on identifying alternative splicing differences. Here, we present an algorithm called IsoPhase that post-processes Iso-Seq data for transcript-based haplotyping. For each gene, IsoPhase gathers the associated full-length reads, each representing a single transcribed molecule. It then calls SNPs and is able to infer the haplotype of the reads due to the full-length nature of the sequencing. We applied IsoPhase to a maize Iso-Seq dataset consisting of two homozygous parents (B73 and Ki11) and two F1 reciprocal hybrids (B73xKi11, Ki11xB73). We validated the majority of the SNPs called with IsoPhase against matching short read data and identified cases of allele-specific, gene-level and isoform-level expression. Our results show that maize parental lines and hybrid lines display different splicing activity, and 6,847 genes can be phased through IsoPhase in two reciprocal hybrids using embryo, endosperm and root tissues. Our study identified parental origin isoforms in maize hybrids, different novel isoforms between maize parent and hybrid lines, provides measures of haplotypic expression that increase power and accuracy in studies of allelic expression. It is the first study of phased full-length isoforms in maize, as well as in plants, which provides insights about maize and plant heterosis at allele-specific full-length transcriptional level. The approach used in this study also provide important information for many other phasing studies in different species.

Funding acknowledgement: National Science Foundation (NSF), United States Department of Agriculture (USDA)

# ACCESS TO MAIZECODE DATA

Liya Wang[1], Zhenyuan Lu[1], Xiaofei Wang[1], Doreen Ware[1,2], Thomas Gingeras[1], MaizeCODE Consortium[1]

[1]Cold Spring Harbor Laboratory, Plant Science, Cold Spring Harbor, NY, [2]USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

MaizeCODE, a project for an initial analysis of functional elements in the maize genome, has assayed five tissues of four maize genomes (B73, NC350, W22, TIL11) for RNA-seq, Chip-seq, Rampage, small RNA, and MNase (outside collaboration). MaizeCODE is committed to open accesses and reproducible science based on the FAIR data principles, providing both human and machine access to the data. There are three ways to access MaizeCODE data: First, all raw data are available through CyVerse Data Store, user can bulk download all data sets through iCommands (command line) or CyberDuck (GUI); Second, all raw data and ground-level analysis results of MaizeCODE data are available through SciApps (https://www.SciApps.org), a cloud-based bioinformatics workflow platform, for integrative analysis; Third, peaks and signals from ground level analysis are available on JBrowser. In addition, all raw data has been submitted to NCBI short read archive (SRA) using the SRA submission pipeline in the CyVerse Discovery Environment (DE). Through the submission process, all experimental metadata are stored in the iRODS based Data Store of CyVerse and used for automating the ground level analysis on SciApps. SciApps organizes both replicates (and controls if available) of each assay as one experiment (or a workflow with the unique id), which represents an entity that chains raw data, analysis results, experimental metadata, and computational metadata together. The ground level analysis includes quality control (QC), alignment to the reference genome, filtering, quantification (e.g. for gene expression), and peak calling (if needed). SciApps provides both a Graphical User Interface (GUI) and a RESTful API for users to check QC results, process new data, and reproduce existing analysis on the Texas Advanced Computing Center (TACC) cloud.

MaizeCODE is supported by an NSF grant IOS 1445025; SciApps is supported by an NSF grant DBI-1265383, and USDA-ARS (1907-21000-030-00D).

# A WEB TOOL FOR PHYSICIAN-DRIVEN CLINICAL DIAGNOSTIC GENOMIC ANALYSIS

Tony Di Sera[1], Alistair Ward[1,2], Matt Velinder[1], Aditya Ekawade[1], Gabor Marth[1,2]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]Frameshift Genomics, Boston, MA

The benefits of DNA sequencing to support diagnosis of inherited disease, cancers and other genetic disorders have been clearly demonstrated. Whole genome, exome, or panel sequencing is already a mainstay of disease discovery, and is increasingly deployed for clinical diagnostics. Here we address an emerging need: providing subspecialty clinicians with the ability to carry out analysis in their own patients' genomic data. Existing software tools for genomic analysis require substantial bioinformatics involvement. We are developing ***clin.iobio***, based on our popular IOBIO web apps, to guide physicians through all necessary steps of clinical diagnostic genomic analysis, from data quality review, candidate gene generation, variant interrogation and prioritization, to reporting of all findings, in a visual, compelling fashion. Analysis is carried out in real time, without requiring the help of bioinformaticians.

***Clin.iobio*** brings together multiple IOBIO apps in a single platform, and provides a directed workflow through the analysis process. At each step, a checklist of tasks promotes analysis consistency and repeatability. The workflow begins with data quality examination. Next, the user is prompted to build a set of candidate genes most likely to underlie the patient's disorder. ***Genepanel.iobio*** (http://genepanel.iobio.io) combines NCBI's genetic testing registry, the Phenolyzer tool, and the user's knowledge of likely genes to generate this gene list. These genes are propagated to ***gene.iobio*** (http://gene.iobio.io) for real-time annotation and analysis of the variants in these genes. In reviewing the variants automatically prioritized by ***gene.iobio***, users are encouraged to assign a significance level and attach notes to variants to promote collaboration. Finally, variants drawn from the whole exome or genome, independent of the candidate gene list, and conforming to predetermined filters are presented for review. This step ensures all potentially disease-causing variants in all genes are examined. A final report summarizes the findings of the analysis process. Importantly, ***clin.iobio*** facilitates dynamic analysis and re-analysis. For example, if no candidate variants are discovered, the user can jump back in the workflow to create a more expansive gene list, bringing in additional variants for review. We demonstrate ***clin.iobio*** in a live demonstration, using University of Utah Undiagnosed Clinic cases, for which the clin.iobio tool provided case solutions as part of our practice. This dynamic and collaborative analysis platform will help move genomic analysis into the clinical practice of subspecialty clinicians.

# FUNCTIONAL CHARACTERIZATION OF PROGRAMMATICALLY ELIMINATED GENES IN THE SEA LAMPREY

Courtney KM Waterbury[1], Christine Brainson[2], Jeramiah J Smith[1]

[1]University of Kentucky, Biology, Lexington, KY, [2]University of Kentucky, Toxicology and Cancer Biology, Lexington, KY

Nearly every multicellular organism possesses a genome that is invariant across all cell types, although individual cells only express a subset of genes that are encoded in the genome and different cell lineages may differ dramatically with respect to the genes that they express. Differentiation of cell lineages is often mediated by epigenetic silencing events that restrict the expression of genes, including DNA and histone methylation. Notably though, a few species, are known to employ an additional mode of silencing that involves the physical removal of DNA from the genomes of somatic cell lineages during development, a phenomenon generally known as programmed DNA elimination. One of the better studies examples of DNA elimination in vertebrates occurs in the sea lamprey (*Petromyzon marinus*), which results in the physical elimination of nearly 0.5 Gb of "germline-specific" DNA from all somatic tissues during early embryogenesis. Initial analyses of these germline specific regions have shown that the eliminated DNA consists of both repetitive and single copy DNA, and that the orthologs of these single copy DNA among CHIP experiments correspond to binding sites of polycomb repressive complex proteins as well as binding targets of transcriptional activators in cancer. Approximately half of the genes that are eliminated are homologs of known oncogenes and we hypothesized that the other eliminated genes may represent as-yet unclassified oncogenes. To test this idea, we performed *in vitro* overexpression assays to test for potential oncogenic function. Initial assays were performed in a single semi-transformed cell line (HEK-293T). In these assays, overexpression of 75% of assayed genes resulted in the formation of tumors/organoids that were significantly larger than controls, suggesting that these genes are to promote tumor growth if misexpressed in somatic cell lineages. Further analyses using cell lines that are more karyotypically stable and growth-regulated can provide further perspective on the potential oncogenic roles of these genes and variability associated with cell lineage/tissue of origin.

# AN INVESTIGATION OF DNA METHYLATION CORRELATION STRUCTURE ON THE ILLUMINA 450K METHYLATION ARRAY

Sarah H Watkins, Matthew Suderman, Debbie A Lawlor, Nicholas J Timpson, Josine L Min, Tom R Gaunt

University of Bristol, MRC Integrative Epidemiology Unit, Department of Population Health Sciences, Bristol Medical School, Bristol, United Kingdom

Previous studies have shown that DNA methylation (DNAm) sites in close proximity (1-2kb) often have correlated DNAm states (1). However, we know little about the genome-wide patterns of DNAm correlation, whether this correlation structure is replicable, and whether it changes over time. The correlation structure may have implications for the design and interpretation of EWAS studies, and for analysing DMRs where DNAm sites are analysed jointly.

We are investigating whether intra and interchromosomal correlations in DNAm data are due to genetic influences and/or environmental factors, and whether these DNAm sites are enriched for genomic features. To evaluate the stability and replicability of correlation structure we analysed DNAm measured at multiple timepoints (birth, 7 and 15 years) from ARIES, a subset of over 900 children from the ALSPAC study. We are validating it using Born in Bradford DNAm data (birth). We will present an analysis of 50,000 representative probes.

We find that correlations from absolute 0.8 to 1 are dominated by sites on the X and Y chromosomes, and correlations between these DNAm sites change very little between birth and 15 years. Between birth and 7 years the mean change in correlation is no more than 0.05 (SD 0.04), with the highest absolute correlations showing the smallest mean change. Between birth and 15 years, correlations between absolute 0.3 to 0.7 show an increased mean change of up to 0.085 (SD 0.04), whilst correlations around 0 and absolute 1 show the same changes as from 7 years.

Around 10% of correlations between absolute 0.9 and 1 are cis; for lower correlation levels, over 99% of correlations are trans. The highest absolute correlations have the lowest percentage of cis mQTLs, suggesting that cis genetic effects are not a major cause of high correlations between DNAm sites. Correlations with one or both DNAm sites overlapping a cis mQTL change over time at around the same rate as correlations which do not overlap a cis mQTL. In ARIES we find that this pattern is stable from birth to 15 years.

In summary, our results show that high DNAm correlations are preserved between birth and 15 years, with intermediate correlations showing the greatest rate of change. The majority of high correlations between DNAm probes are in trans, and are less likely to have a cis mQTL, suggesting that DNAm correlation is not predominantly driven by genetics.

1.Saffari, A. et al (2018). Estimation of a significance threshold for epigenome-wide association studies. Genetic epidemiology, 42(1), 20-33.

# ENZYMATIC METHYL-SEQ: NEXT GENERATION METHYLOMES

Louise Williams, V K Chaithanya Ponnaluri, Brittany Sexton, Lana Saleh, Katherine Marks, Mala Samaranayake, Bradley W Langhorst, Zhiyi Sun, Thomas C Evans, Jr, Romualdas Vaisvila, Eileen T Dimalanta, Theodore B Davis

New England Biolabs, Ipswich, MA

DNA methylation is important for gene regulation. The ability to accurately identify 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) gives us greater insight into potential gene regulatory mechanisms. Bisulfite sequencing (BS) is traditionally used to detect methylated Cs, however, BS does have its drawbacks. DNA is commonly damaged and degraded by the chemical bisulfite reaction resulting in libraries that demonstrate high GC bias and are enriched for methylated regions. To overcome these limitations, we developed an enzymatic approach, NEBNext Enzymatic Methyl-seq (EM-seq), for methylation detection that minimizes DNA damage, resulting in longer fragments and minimal GC bias.

Illumina libraries were prepared using bisulfite and EM-seq methods. Libraries generated with NA12878 DNA inputs ranging from 10 ng to 200 ng were sequenced using Illumina's NovaSeq 6000. Reads were adapter trimmed (trimadap) and aligned to GRCh38 using BWAMeth. Aggregate metrics like GC bias and insert size distribution (Picard) were assessed before evaluating methylation status of individual Cs (MethylDackel). MethylKit was used for correlation analysis. EM-seq libraries have longer inserts, lower duplication rates, a higher percentage of mapped reads and less GC bias compared to bisulfite converted libraries. Global methylation levels are similar between EM-seq and whole genome bisulfite libraries (WGBS) indicating overall detection of methylated Cs is similar. However, CpG correlation plots demonstrated higher correlation coefficients indicating that EM-seq libraries are more consistent than WGBS across replicates and input amount. GC Bias and dinucleotide distribution showed that EM-seq has more even dinucleotide representation compared to the AT rich representation observed for WGBS. EM-seq's more even coverage allows for a higher percentage of CpGs to be assessed leading to more consistent evaluation of methylation across key genomic features (TSS, CpG island, etc.). EM-seq is more robust than WGBS, works over a wide range of DNA input amounts, has superior sequencing metrics, and detects more CpGs.

# MAPPING GENOME-WIDE DNA METHYLATION PATTERNS IN GLIOMAS IN CONTEXT OF IDH GENE MUTATION STATUS AND REST TRANSCRIPTION FACTOR BINDING

Bartosz Wojtas[1], Michal J Dabrowski[2], Agata Dziedzic[2], Michal Draminski[2], Rafal Guzik[1], Karolina Stepniak[1], Bartlomiej Gielniewski[1], Tomasz Czernicki[3], Pawel Nauman[4], Bartosz Czapski[5], Wieslawa Grajkowska[6], Katarzyna Kotulska[7], Bozena Kaminska[1]

[1]Nencki Institute of Experimental Biology of Polish Academy of Sciences, Laboratory of Molecular Neurobiology, Warsaw, Poland, [2]Institute of Computer Science, Polish Academy of Sciences, Computational Biology Lab, Warsaw, Poland, [3]Medical University of Warsaw, Department of Neurosurgery, Warsaw, Poland, [4]Institute of Psychiatry and Neurology, Department of Neurosurgery, Warsaw, Poland, [5]Mazovian Brodno Hospital, Department of Neurosurgery, Warsaw, Poland, [6]Children's Memorial Health Institute, Department of Pathology, Warsaw, Poland, [7]Children's Memorial Health Institute, Neurology and Epileptology, Warsaw, Poland

Methylation of DNA regulatory regions influence gene expression. Alterations of methylome play an important role in the glioma pathogenesis. Here we have identified differentially methylated sites in gliomas of different histopathological WHO grades (I,II,III and IV) or IDH gene mutation status (n = 21). We used bisulphite conversion and SeqCap Epi CpGiant Methylation panel with Illumina NGS sequencing. Additionally ChIPseq analysis for REST transcription factor was performed on chromatin isolated from freshly resected glioma tumors as well as from IDH mutated and paired isogenic WT cell line. For the same brain tumor samples RNA-seq was performed. We have detected differential pathways affected in IDH mutant and WT samples. We have also noted differential REST transcription factor binding to differentially methylated promoters. We believe that REST may be a mediator of IDH-related phenotype in human gliomas.

Supported by the NCN grant 2015/16/W/NZ2/00314 and by the Foundation for Polish Science TEAM-TECH Core Facility project „NGS platform for comprehensive diagnostics and personalized therapy in neuro-oncology".

# ASSESSMENT OF BIAS AND miRNA ISOFORM QUANTIFICATIONS OF WIDELY-USED SMALL RNA SEQUENCING METHODS REVEALS MAJOR DIFFERENCES

Carrie Wright[1,2], Anandita Rajpurohit[1], Emily E Burke[1], Courtney Williams[1], Leonardo Collado-Torres[1], Martha Kimos[1], Nicholas J Brandon[3], Alan J Cross[3], Andrew E Jaffe[1,4,5,6], Daniel R Weinberger[1,6,7,8], Joo Heon Shin[1,7]

[1]Lieber Institute for Brain Development, Data Science/ Computational Biology, Baltimore, MD, [2]AstraZeneca, Postdoc Program, Cambridge, MA, [3]AstraZeneca, Neuroscience, Cambridge, MA, [4]Johns Hopkins University (JHU), Center for Computational Biology, Baltimore, MD, [5]JHU, Epidemiology, Baltimore, MD, [6]JHU, Psychiatry and Behavioral Sciences, Baltimore, MD, [7]JHU, Neurology, Baltimore, MD, [8]JHU, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

RNA sequencing (RNAseq) offers advantages for profiling small RNA, yet several important questions remain about miRNA quantifications using this method. In particular, the accuracy of miRNA isoform (isomiRs) quantifications has been understudied. Although previously believed to be artifacts, isomiRs have recently been discovered to be functional and are associated with several diseases. To improve research in this area it is critical to determine that our quantifications of isomiRs are accurate. We compared the quantification of isomiRs using a variety of library preparation methods from both a synthetic sample designed to not contain isomiRs and from RNA derived from homogenate brain tissue. Our results indicate that there are major differences in the detection of isomiRs by the tested methods, and that some widely-used methods appear to falsely detect isomiRs in high abundance. One method detected 400% more unique false isomiRs on average than the other methods. Furthermore, bias in small RNAseq has largely been attributed to adapter ligation bias and method improvements have successfully addressed this, yet we hypothesized that reverse transcription and amplification may also contribute considerable bias. We therefore evaluated if using unique molecular identifiers (UMIs) to collapse duplicate reads in small RNAseq libraries generated from synthetic miRNAs and human brain total RNA samples to only the reads present before reverse transcription and amplification would improve the accuracy of miRNA quantifications. We compared miRNA quantifications obtained with the same library preparation methods with and without the use of UMIs and to quantifications using a random subset of reads as a control. Our results indicate that reverse transcription and amplification may also contribute considerable bias, and that the inclusion of UMIs improves the accuracy of miRNA quantifications. Overall our work provides guidelines for future quantifications of small RNA.

# PREDICTING DIFFERENTIAL EXPRESSION OF DROUGHT-ASSOCIATED GENES IN MAIZE

Travis Wrightsman[1], Jeremy Pardo[1,2], Jacob Washburn[3], Hai Wang[3,4], Edward Buckler[1,3,5]

[1]Cornell University, Section of Plant Breeding and Genetics, Ithaca, NY, [2]Michigan State University, Department of Horticulture, East Lansing, MI, [3]Cornell University, Institute for Genomic Diversity, Ithaca, NY, [4]Chinese Academy of Agricultural Sciences, Biotechnology Research Institute, Beijing, China, [5]United States Department of Agriculture, Agricultural Research Service, Ithaca, NY

Plants allow us to study the mechanisms and behaviors of large, complex genomes in a wide variety of controlled environments at relatively large scales. Maize, sorghum, and their relatives in the Andropogoneae tribe are among the most productive and widely-adapted plants in the world, largely due to their evolution of C4 photosynthesis. Compared to other commonly-studied plants, such as Arabidopsis and rice, maize has a large (2.3 Gbp) genome and long-range (>40 kbp) cis-acting regulatory elements. Drought is already a critical threat to maize production systems across the globe and climate change will potentially worsen its impact. We will present on using machine learning to develop models that can predict the differential expression (DE) in response to drought conditions directly from genomic sequence. We collected RNAseq data from 27 three-week-old maize genotypes under control, 1 day without watering, and 3 days without watering conditions. We previously developed machine learning models that predict on/off expression state as well as continuous expression levels from promoter and terminator sequence. Using the RNAseq data, we have trained three types of machine learning models to predict differential expression given 1500bp flanking sequence around the transcription start site and 1500bp around the transcription termination site. The first two are classification models, extended from the on/off model, that try to predict (1) DE state or (2) DE direction: no change, upregulated, or downregulated. The third model is continuous and predicts log2 fold change. Future work to improve prediction accuracy will extend the input of these models to include more distant cis regulatory sequences.

# CASTE- AND PESTICIDE-SPECIFIC EFFECTS OF NEONICOTINOID PESTICIDE EXPOSURE ON GENE EXPRESSION IN BUMBLEBEES

Thomas J Colgan[1], Andres Arce[2], Richard Gill[2], Yannick Wurm[1]

[1]Queen Mary U London, Biology, London, United Kingdom, [2]Imperial College, Biology, Ascot, United Kingdom

Social bees are important insect pollinators of wildflowers and agricultural crops, making their reported declines a global concern. A major factor implicated in these declines is the widespread use of neonicotinoid pesticides. Indeed, recent research has demonstrated that exposure to low doses of these neurotoxic pesticides impairs bee behaviors important for colony function and survival. However, our understanding of the molecular-genetic pathways that lead to such effects is limited, as is our knowledge of how effects may differ between colony members.
To understand what genes and pathways are affected by exposure of bumblebee workers and queens to neonicotinoid pesticides, we implemented a transcriptome-wide gene expression study. We chronically exposed Bombus terrestris colonies to either clothianidin or imidacloprid at field-realistic concentrations while controlling for factors including colony social environment and worker age. We reveal that genes involved in important biological processes including mitochondrial function are differentially expressed in response to neonicotinoid exposure. Additionally, clothianidin exposure had stronger effects on gene expression amplitude and alternative splicing than imidacloprid. Finally, exposure affected workers more strongly than queens. Our work demonstrates how RNA-Seq transcriptome profiling can provide detailed novel insight on the mechanisms mediating pesticide toxicity to a key insect pollinator.

# BUILDING A COMPREHENSIVE ENCODE LONG READ TRANSCRIPTOME PIPELINE

Dana Wyman[1,2], Gabriela Balderrama-Gutierrez[1,2], Shan Jiang[1,2], Brian Williams[3], Barbara Wold[3], Ali Mortazavi[1,2]

[1]University of California, Irvine, Department of Developmental and Cell Biology, Irvine, CA, [2]University of California, Irvine, Center for Complex Biological Systems, Irvine, CA, [3]California Institute of Technology, Division of Biology, Pasadena, CA

Conventional short-read RNA sequencing has been widely used to quantify gene expression in a variety of applications. However, short reads on their own lack the ability to resolve full-length isoforms, which can be several kilobases in length. Furthermore, computational methods developed to reconstruct isoforms from short read data are plagued by challenges, and results from different algorithms tend to be inconsistent. While long read sequencing technologies such as PacBio cDNA sequencing and Oxford Nanopore direct RNA sequencing have a higher error rate than short reads, they have great potential for isoform discovery and characterization of the 90% of multi-exon human genes that are thought to undergo alternative splicing. We have developed TALON, a computational pipeline to process long reads into cleaned isoforms and to compare them to existing annotations. TALON tracks isoforms and expression levels across different datasets as well as sequencing platforms, allowing novel discoveries to be vetted and quantified. We demonstrate the features of our pipeline on PacBio and Oxford Nanopore datasets from five ENCODE human and mouse cell lines as well as mouse brains, and show that long read technologies are mature enough to produce full-length transcriptomes by comparing the results to existing ENCODE data. We also describe how TALON is being deployed by the ENCODE project.

# ISSUES IN EXPERIMENTAL DESIGN IN ENHANCER–PROMOTER INTERACTION PREDICTION

Wang Xi[1], Michael A Beer[1,2]

[1]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

While quantitative modeling of cell-specific enhancer activity and variant impact is progressing rapidly, predicting the promoter and gene targets of enhancers remains challenging. To date, the best performance was achieved by Whalen, et al,1 who reported a computational model (TargetFinder) using epigenomic features that could predict enhancer-promoter (EP) interactions with high F1 accuracy, implying that most EP pairs could be correctly classified as interacting or not based on the epigenomic state of the enhancer (E), promoter (P), and the intervening genomic interval, or window (W), between the enhancer and the promoter. In their training data, multiple contiguous enhancers often positively interact with the same promoter, but are all labelled as independent positive EP interactions for training, even though they share P and W features. If samples with identical features are not in the same cross-fold validation (CV) test set, the shared features could lead to overfitting. In fact, over 92% of EP pairs share a promoter with another EP pair and are thus subject to significant overfitting through the shared promoter features. We reproduced the published results from TargetFinder, but then we used chromosomal CV test sets, so that no EP pairs with shared features could be in the same test set, and found that F1-score dropped dramatically, to only slightly better random guessing. Thus our current ability to computationally predict EP interactions remains very low. We will discuss what additional features may be necessary to improve predictive power.

Ref

[1]Whalen S, Truty R M, Pollard K S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin[J]. Nature genetics, 2016, 48(5): 488.

# SINGLE CELL PROFILING ON siRNA INDUCED C-MYC SILENCING OF HUMAN GLIOBLASTOMA

Bingqing Xie[1,2], Anindita Basu[1], Susan Olalekan[1], Natalia Maltsev[2], Eric Gauchat[1], Heather Eckart[1], Janine Hensel[3], Olga Volpert[3]

[1]The University of Chicago, Genetic Medicine, Chicago, IL, [2]The University of Chicago, Human genetics, Chicago, IL, [3]The University of Texas, MD Anderson Cancer Center, Houston, TX

Understanding structural, functional and cellular heterogeneity present within solid tumors is necessary to advance current cancer therapies. Although bulk RNA-Seq is informative, it fails to identify respective contributions from individual cell-types, and their complex interplay in cancer progression. The goal of this study is to use a high-throughput, droplet-based single-nucleus mRNA sequencing techniques, DroNc-Seq (Nature Methods, '17), to profile the siRNA induced c-Myc silencing in Glioblastoma(GBM). We delivered two type of siRNA, scrambled control and anti-cMyc respectively, to GBM cellline U87 by exosomes and then inject them into mice. The GBM samples were dissected out from these patient derived xenograft model, snap frozen in liquid nitrogen and analyzed using DroNc-seq. Using canonical cell markers, we identify cancer subtypes and constituent cell-types, including astrocyte, neuron, and precursor astrocyte cell compartments in both conditions. Using unsupervised learning techniques, we cluster samples from different condition and investigate the function from the individual gene expression profiles from multiple clusters. The differential expressed(DE) genes are computed on the aligned cell clusters and overall populations from two conditions. We identify the HIF-1 signaling pathway and PI3k signaling pathway from the overall DE genes. The gene prioritization on Protein-protein interaction network is performed on the DE genes from aligned clusters. We detect Myc at top 0.1% to 1% position out of all candidate genes. We further reconstruct the network for the top prioritized genes which shows the potential molecular mechanisms corresponding to the cMyc protein silencing. Trajectory analysis is performed on cell line, control, and anti-cMyc samples and a pseudotime orders the cells starting from pure cell line sample, branching in control samples, and finally landing in anti-cMyc samples which reveals the transition from these conditions.

# ANALYSIS OF LUNG CANCER SPLICING ABNORMALITY USING NANOPORE-SEQUENCING

Liu Xu, Miho Oka, Masahide Seki, Ayako Suzuki, Yutaka Suzuki

Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, Kashiwa, Japan

In cancer tissues, immune cells recognize cancer cells as non-self by neoantigens derived from somatic mutations. Therefore, recent studies have proposed Tumor Mutation Burden (TMB), which is mainly determined by missense mutation load, as a biomarker of response to immune checkpoint inhibitors. However, there are many other genomic mutations that cancer cells have, such as frameshift mutation and splice site mutation. It is possible that these kinds of mutations alter amino acids significantly. Nanopore Sequencer, such as MinION (Oxford Nanopore Technologies), can sequence ultra-long reads (hundreds of kb). Since MinION doesn't need cut mRNAs to become fragments unlike Illumina sequencer, the entire sequence of transcript can be sequenced as single read by nanopore sequencer.

In this study, we applied full-length cDNA sequencing using nanopore sequencer to detection cancer-specific neoantigens. The innovative approach is able to measure the full length structure of mRNA with frameshift mutation and the number of mRNA with structural abnormality.

Firstly, we performed full-length cDNA sequencing analysis of 23 lung adenocarcinoma cell lines and five clinical samples using MinION (1d^2 sequencing). We mapped the reads to human genome (hg38) using Minimap2 (2.14). The splice junctions detected from high quality were compared with that of RefSeq transcript models. We measured the splicing patterns of mRNA with frameshift mutation and the expression levels of mRNA with abnormal splicing pattern. Next, we predicted the immunogenicity of the identified abnormal mRNA.

Latest results will be presented.

# DETECTION OF CIRCULATING miRNA LEVELS IN LARGE COHORT SCHIZOPHRENIA

Qi Xu

Chinese Academy of Medical Sciences and Peking Union Medical College, Molecular Biochemistry, Beijing, China

Schizophrenia is one of the most common severe mental disorders,with a lifetime risk of 1%in the populationworldwide. Over the years, the diagnosis of schizophrenia has remained symptom based, relying mainly on self-reports from patients, mental state examination, and clinical interviews, and lacking objective laboratory tests. Such a diagnostic strategy can sometimes lead to misdiagnosis and has been criticized widely. To remedy this embarrassing state of affairs, a set of biomarkers has been proposed based on physical and biological tests. In a currently finished study, Global plasma miRNAs were profiled in a test cohort of 850 schizophrenia patients and 963 control subjects, using RNA sequencing, TaqMan Low Density Array, and quantitative reverse transcription polymerase chain reaction (qRT-PCR) assays. The captured miRNAs were then validated by qRT-PCRassays in an independent cohort of 623 schizophrenia patients, 654 control subjects. The global plasma miRNA screening revealed eight miRNAs that were up-regulated in schizophrenia, as revealed by both assay platforms. The qRT-PCR analysis showed the up-regulation of miR-17-5p and miR-193a-3p in schizophrenia but not in nonschizophrenia disorders. Conclusions: The up-regulation of miR-17-5p and miR-193a- 3p is a state-independent biomarker for schizophrenia, and these two miRNAs could be used to develop a diagnostic tool for schizophrenia.

# LIKELIHOOD MODEL TO ESTIMATE STRENGTH OF SELECTION AND GENOMIC ARCHITECTURE UNDERLYING A TRAIT USING GWAS SUMMARY STATISTICS

Alexander T Xue[1], Yi-Fei Huang[2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Pennsylvania State University, Biology, University Park, PA

The growth of genome-wide association studies (GWAS) over the past decade has provided a wealth of resources for elucidating the genetic architecture underlying phenotypic traits. In particular, there has been increasing interest in exploiting GWAS data to detect genome-wide signatures of natural selection acting on a given trait. However, many available methods are not designed to utilize GWAS summary statistics, which include the estimated effect sizes and standard error resulting from a GWAS analysis. GWAS summary statistics are much more easily accessible than individual-level data (i.e. phenotypes and genotypes) due to privacy concerns, hence computational tools that can bypass the latter provide much more utility to potential users. Furthermore, current approaches often do not quantify the strength of selection across a genome-wide distribution acting on a specific trait, instead focusing on testing for the overall presence of selection on a phenotype, or estimating the correlation between effect sizes and selection coefficients. To this end, we introduce a likelihood framework that conditions on genome-wide GWAS summary statistics and allele frequencies to directly infer selection coefficients across sites, i.e. the distribution of fitness effects (DFE). Our likelihood model also concurrently estimates the distribution of true effect sizes (DES) and polygenicity/mutational target size. To assess the accuracy and bias of this method, we performed in silico experiments that tested our inferential framework against pseudo-observed datasets (PODS) of genotypes and corresponding GWAS summary statistics under a given DFE and DES. After demonstrating strong performance in our simulation experiments, we applied our likelihood model against GWAS results from the UK Biobank dataset, allowing comparison between DFEs for different traits in the same population. We extended our comparative analysis by estimating the DFE for several of these traits in other populations, including African-Americans, East Asians, Latin Americans, and South Asians. Our likelihood method offers powerful yet convenient population genomic inference from GWAS summary statistics, addressing a growing need as GWAS results further accumulate, especially in non-European populations.

# COMPARATIVE PAN-CANCER DNA METHYLATION ANALYSIS REVEALS CANCER COMMON AND SPECIFIC PATTERNS

Xiaofei Yang[1], Lin Gao[2], Shihua Zhang[3]

[1]Xi'an Jiaotong University, School of Electronic and Information Engineering, Xi'an, China, [2]Xidian University, School of Computer Science and Technology, Xi'an, China, [3]Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Beijing, China

Abnormal DNA methylation is an important epigenetic regulator involving tumorigenesis. Deciphering cancer common and specific DNA methylation patterns is essential for us to understand the mechanisms of tumor development. The Cancer Genome Atlas (TCGA) project provides a large number of samples of different cancers that enable a pan-cancer study of DNA methylation possible. Here we investigate cancer common and specific DNA methylation patterns among 5480 DNA methylation profiles of 15 cancer types from TCGA. We first define differentially methylated CpG sites (DMCs) in each cancer and then identify 5450 hyper- and 4433 hypomethylated pan-cancer DMCs (PDMCs). Intriguingly, three adjacent hypermethylated PDMC constitute an enhancer region, which potentially regulates two tumor suppressor genes *BVES* and *PRDM1* negatively. Moreover, we identify six distinct motif clusters, which are enriched in hyper- or hypomethylated PDMCs and are associated with several well-known cancer hallmarks. We also observe that PDMCs relate to distinct transcriptional groups. Additionally, 55 hypermethylated and 7 hypomethylated PDMCs are significantly associated with patient survival. Lastly, we find that cancer-specific DMCs are enriched in known cancer genes and cell-type-specific super-enhancers. In summary, this study provides a comprehensive investigation and reveals meaningful cancer common and specific DNA methylation patterns.

# MOTIF-RAPTOR: CELL TYPE SPECIFIC PRIORITIZATION OF TRANSCRIPTION FACTOR AND GENETIC VARIANT BASED ON GWAS SUMMARY STATISTICS

Qiuming Yao[1,4], Paolo Ferragina[2], Yakir Reshef[3], Daniel Bauer[1], Luca Pinello[4]

[1]Boston Children's Hospital, Division of Hematology/Oncology, Boston, MA, [2]University of Pisa, Department of Computer Science, Pisa, Italy, [3]Harvard University, Department of Computer Science, Cambridge, MA, [4]Massachusetts General Hospital, Molecular Pathology Unit, Boston, MA

GWAS have enabled the identification of thousands of common genetic variants associated with normal or disease traits. However, fine mapping and functional interpretation of these variants remain challenging due to linkage disequilibrium. It is well known that transcription factors (TFs) are important in regulating gene expression and that their binding sites can be modulated by genetic variants, however we currently lack a systematic understanding on how this contributes to a phenotypic trait of interest. Although several cell type specific annotations and in particular, transcription factor datasets obtained by chromatin immunoprecipitation followed by sequencing (ChIP-seq) have been helpful to narrow down the search on functional variants and their mechanisms, they are not available for the majority of cell lines. On the other hand, thanks to the ENCODE project, we have chromatin accessibility profiles and gene expression data for hundred of cell types and also models to predict transcription factors binding based solely on DNA sequences.
In this work we address the following question: how can we optimally integrate this information to best predict for a given phenotype which TF is influenced by genetic variants and in which cell type? To this end, we developed a computational approach called Motif-Raptor. This tool integrates TF predictive models, cell type specific and matched DHS and gene expression datasets with GWAS summary statistics to make systematic discoveries of critical functional TF-SNP binding events. Given a phenotypic trait, Motif-Raptor recovers relevant cell and tissue types, TFs for which binding sites may be disrupted or enhanced by genetic variants and ranks them based on their functional potential quantified by both TF specific expression and SNP modulation patterns.
Using Motif-Raptor on publicly available summary statistics for common diseases and traits we uncovered important TFs, cell types and SNPs. In rheumatoid arthritis, we identified two previously characterized factors for which the binding sites are significantly modulated by associated genetic variants in accessible peaks for CD8+ T cells: a significant disruption for Gfi1 and a significant enhancement for NFKB binding sites. Similarly, for red blood cell traits we uncovered that the two critical regulators of erythropoiesis i.e. Tal1 and Gata1 whose binding sites are significantly disrupted by associated genetic variants in K562 erythroid cell line specific accessible chromatin.

Taken together Motif-Raptor can identify relevant cell types, critical TFs and trait-associated variants to drive hypotheses regarding the mechanism of action of non-coding trait associations. In addition, we provide precomputed analysis for several traits, which itself could serve as a valuable resource for the functional genomics community. Motif-Raptor is available as an open-source software at: https://github.com/pinellolab/MotifRaptor.

# MOMI-G: MODULAR MULTI-SCALE INTEGRATED GENOME GRAPH BROWSER

Toshiyuki T Yokoyama, Sakamoto Yoshitaka, Masahide Seki, Yutaka Suzuki, Masahiro Kasahara

The University of Tokyo, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, Chiba, Japan

Long-read sequencing allows more sensitive and accurate discovery of structural variants (SVs) which are associated with various human diseases. As the number of SVs that can be identified by sequencing grows, understanding SVs and the relationships between them in a given genome becomes more difficult, especially when there are large (over a megabase) or nested SV. SV visualization is a very important step for identifying SVs and interpreting their potential impacts because it enables the manual inspection of SVs for achieving two goals. The first is to better understand the relationships between SVs and other genomic features. The second is to ensure a smaller number of false positives. However, large or nested SVs often obscure the relationship between genomic regions that are distant in the reference genome but are actually close in the target genome. Thus, SV visualization tools should be able to simultaneously display multiple intervals along with their relationships, even when the breakpoints are distant or when SVs are nested.
Here we present Modular Multi-Scale Integrated Genome Graph Browser, MoMI-G, a web-based genome browser to visualize SVs, genes, repeats, and other annotations as a variation graph with paths. MoMI-G (1) allows visualization of (possibly distant) multiple intervals; (2) displays SVs that span multiple intervals; (3) displays SVs at varying scales, i.e., chromosome, gene, and nucleotide scales; (4a) the chromosome scale view can show the distribution of SVs on one or more chromosomes; (4b) the gene scale view can show annotations such as exon/intron structures and repeats; (4c) the nucleotide scale view can show nucleotide-level alignments, in particular, read alignments that correspond to both alleles of heterozygous SVs are shown simultaneously; and (5) allows users to manually inspect hundreds of SVs. Therefore, MoMI-G helps in understanding the entire picture of SVs, even those that are nested or large, regardless of their size. Researchers are able to obtain novel biological knowledge by comparing a reference genome with an individual genome by using MoMI-G. MoMI-G is freely available at https://github.com/MoMI-G/MoMI-G under the MIT license.

# EPIGENETICS OF TYPE 2 DIABETES IN THE QATARI POPULATION

Noha A Yousri[1], Omar Albagha[2], Steven C Hunt[1]

[1]Weill Cornell Medicine-Qatar, Genetic Medicine, Doha, Qatar, [2]Qatar Biomedical Research Institute, Hamad Bin Khalifa University, Genetics, Doha, Qatar

**Background & Objectives**: Type 2 Diabetes is highly prevalent in the Middle East with a prevalence of more than 15% in Qatar. Past studies have revealed several DNA methylation sites associated with T2D, yet Middle Eastern populations are barely addressed in this direction. Our aim is to identify methylation sites associated with T2D and investigate pathways linked with the identified sites using a fairly large number of T2D patients (50% of 384 studied samples) and using the latest EPIC array technology to capture more than 850K sites.

**Materials and Methods**: 384 DNA samples (50% T2D) were extracted from blood of Qatari samples provided by Qatar Bio-Bank. All samples were profiled for more than 850K DNA methylation sites using Illumina EPIC arrays. All analyses were done using minfi package in R. Three samples were removed after Quality Control. A multicollinearity of Age and T2D lead to investigating different methods for analysis. Linear regression model was used in R, while correcting for gender, BMI, actual cell counts, batch effect, well position, as well as three principal components computed from genotype array to reflect population stratification.

**Results**: We identified 40 methylation sites which were significantly associated with T2D ($p < 0.05/830K = 6.02 \times 10^{-8}$). 20 of the identified sites are inside a gene body and the rest are outside genes. We also replicated 17 previously published CpG associations with T2D including also few sites which are known to be associated with glycemic traits and age.

**Conclusion**: This study highlights new methylation sites associated with T2D using EPIC arrays, and replicated many of the known sites. It adds results from a different ethnicity with a different lifestyle/environment for comparison with other ethnicities and to further understand the mechanisms underlying the disease in an understudied population.

Acknowledgement: This publication was made possible by PPM2 grant #:0226-170020 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors. The authors would also like to thank Qatar Bio-Bank (QBB), a member of Qatar Foundation for its support.

# GERMLINE GENETIC VARIATION DRIVES THE SOMATIC LANDSCAPE OF TUMORS.

Noah Zaitlen[1], Serghei Mangul[2], Alexander Gusev[3]

[1]UCLA, Neurology, Los Angeles, CA, [2]UCLA, Computer Science, Los Angeles, CA, [3]DFCI, Genetics, Boston, MA

Hundreds of thousands of cancer patients have had their tumors sequenced to identify clinically actionable mutations. In addition to saving lives, this activity has produced valuable research data leading to significant discoveries in basic and translational domains. Unfortunately, the targeted nature of clinical tumor sequencing has precluded powerful interrogation of germline-somatic interactions and their concomitant effects on cancer outcomes. In particular, the extent to which germline alleles impact the fitness, architecture, and clinical impact of somatic mutations is unknown. To address this shortcoming we developed a computational approach to infer research level genome-wide germline variation, including HLA, TCR/BCR, and microbial sequences, from off-target sequencing reads. We applied our method to the targeted tumor sequences of the Dana Farber Cancer Institute PROFILE cohort (n>20,000), thereby producing the world's largest germline-somatic data set by over an order of magnitude.

Levering this unique resource, we test for associations between fundamental components of germline and somatic variation in known oncogenes. Specifically, for the non-silent mutational burden in each of the PROFILE target genes as well for total tumor mutational burden (TMB), we conduct genome-wide association, cis/trans heritability, population prevalence, TWAS, and polygenic risk score association tests. Our completed analysis of the PROFILE lung adenocarcinoma (LUAD) sub-cohort (n=2429; >4x larger than LUAD in TCGA) has already produced remarkable and strongly statistically significant associations between germline genetics and somatic events. We identify over forty genome-wide significant associations (p-value < 5e-8) of germline variants to non-silent mutations in important oncogenes including BRCA1, ALK, DNMT3A, FANCA, GNAS, EPHA3/5, GLI2, as well the total mutational burden. We estimate common variant heritability to be over 60% (p-value < 1e^-4) for multiple genes including BRCA1, EPHA3, and EHPA5. We show that asian ancestry, identified from PCA, is associated with increased EGFR mutation and decreased KRAS mutation, and is driven by genetic ancestry rather than self-identified race and ethnicity. Finally, we show that germline polygenic risk for several lung cancer subtypes and smoking is significantly associated to mutations in specific genes as well as an individual's total mutational burden. Together, these results demonstrate that germline variation is a major contributor to the dynamics of somatic mutations and cancer outcomes.

# THE ENSEMBL POST-GWAS ANALYSIS PIPELINE

Daniel R Zerbino[1,2], Verena Zuber[1,2,3], Seongwon Hwang[1], Gareth Peat[1,2] Michael Nuhn[1,2], José C Marugán[1,2], William Jones[1,2,4], William Newell[2,5], Oliver Stegle[1,2], Ian Dunham[1,2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom, [2]Open Targets, Hinxton, United Kingdom, [3]Imperial College London, Department of Epidemiology and Biostatistics, London, United Kingdom, [4]Wellcome Sanger Institute, Hinxton, United Kingdom, [5]GSK, Medicines Research Center, Stevenage, United Kingdom

We present here a pipeline that compares GWAS results to a large collection of publicly available cis-regulatory datasets to help prioritise candidate causal genes.

Despite an ever increasing number of trait-associated regions, GWAS have not succeeded in translating these associations into many therapeutic success stories. The main bottleneck is inferring truly causal genes from the GWAS results that can then be used as drug targets and consequently new therapies. This gap between genetics research and translational applications is largely explained by the difficulty in functionally interpreting non-coding variants. Although annotating and prioritizing coding variants is a well-studied problem, determining the regulatory effect of non-coding variants is still difficult. In effect, many of the drug targets tested by the pharmaceutical industry fail in clinical trials for lack of efficacy.

To overcome this gap, we have developed the Ensembl Post-GWAS Analysis Pipeline (PostGAP) that analyses GWAS summary statistics in light of a collection of functional datasets such as expression Quantitative Trait Loci (QTL), covariance analyses in chromatin state between distant regions of the genome, and sequencing-based assays such as Promoter-Capture Hi-C. We have run PostGAP across all GWAS Catalog studies and present the results in an interactive web interface which can used to examine the evidence supporting the candidate causal genes. PostGAP can also easily be installed and run locally, providing a fast and convenient pre-processing of GWAS summary statistics in view of candidate gene prioritisation.

Further, we have developed a probabilistic approach for estimating the posterior probability of colocalisation between phenotypic association summary statistics and gene expression QTLs. This method expands on existing colocalisation algorithms by efficiently testing for multiple SNPs in one locus and by taking into account as many available epigenomic or cis-regulatory datasets as possible to adjust a SNP's prior probability of being regulatory.

# DEEP LEARNING-AUGMENTED RNA-SEQ ANALYSIS OF TRANSCRIPT SPLICING

Zijun Zhang[1], Zhicheng Pan[1], Yi Ying[2], Zhijie Xie[2], Samir Adhikari[2,3], John Phillips[4], Russ P Carstens[5], Douglas L Black[2], Yingnian Wu[6], Yi Xing[1,2,3,7]

[1]University of California, Los Angeles, Bioinformatics Interdepartmental Graduate Program, Los Angeles, CA, [2]University of California, Los Angeles, Department of Microbiology, Immunology & Molecular Genetics, Los Angeles, CA, [3]The Children's Hospital of Philadelphia, Center for Computational and Genomic Medicine, Philadelphia, PA, [4]University of California, Los Angeles, Department of Molecular and Medical Pharmacology, Los Angeles, CA, [5]University of Pennsylvania, Department of Medicine, Philadelphia, PA, [6]University of California, Los Angeles, Department of Statistics, Los Angeles, CA, [7]University of Pennsylvania, Department of Pathology and Laboratory Medicine, Philadelphia, PA

The rapid accumulation of RNA-seq data across diverse cell types and conditions provides an unprecedented resource for characterizing transcriptome complexity. However, the use of these large-scale data in routine RNA-seq studies to detect patterns of expression and thereby discover new regulatory events has been limited. Here we report DARTS, Deep-learning Augmented RNA-seq analysis of Transcript Splicing, a computational framework that integrates deep learning-based predictions with empirical evidence in specific RNA-seq datasets to infer differential alternative splicing between conditions. A core component of DARTS is a deep neural network (DNN) that predicts differential alternative splicing using *cis* RNA sequence features and *trans* RNA binding protein levels. DARTS DNN trained on public RNA-seq datasets (ENCODE, Roadmap Epigenomics) displays a high prediction accuracy and generalizability. Incorporating DARTS DNN prediction as an informative prior significantly improves the inference of differential alternative splicing, especially from low-coverage RNA-seq datasets. In cellular models of the epithelial-mesenchymal transition, DARTS reliably predicted alternative splicing changes in lowly expressed genes, that were inaccessible by a conventional RNA-seq analysis even at a high sequencing depth. Thus, DARTS capitalizes on large-scale public RNA-seq resources to discover differential alternative splicing across diverse transcriptomes.

# GENOME-WIDE INTERROGATION OF HUMAN POSITIVE REGULATORY ELEMENTS

Nanxiang (Samuel) Zhao[1], Kayvan Najarian[1], Harm Derksen[2], Alan P Boyle[1,3]

[1]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Department of Mathematics, Ann Arbor, MI, [3]University of Michigan, Department of Human Genetics, Ann Arbor, MI

The intricate spatiotemporal control of gene expression is achieved by the interplay between transcription factors (TFs) and non-coding regulatory elements (REs). REs are classified as positive regulatory elements (PREs) or negative regulatory elements (NREs) based on their impact on gene expression. Recently, a new assay, ATAC-STARR-seq, has been developed to directly measure regulatory activities of positive regulatory regions genome-wide by coupling accessible chromatin extraction with self-transcribing episomal reporters. Even though ATAC-STARR-seq effectively assays potential positive regions, the critical gap is that there lacks a subsequent algorithm to accurately identify the PREs, and to estimate their regulatory activity scores due to noisy background and complex signals the assay returns. The overall objective of this project is to systematically interrogate human PREs in a high-throughput manner. To address this objective, we aim to develop a computational method using signal processing and machine learning techniques to interrogate human PREs genome-wide with higher accuracy and resolution. The rationale that underlies the research is that signal processing techniques, an engineering methodology, have long been used for effective noise cancellation and have played an important role in extracting, processing, and interpreting the information contained in biomedical data, which will offer a new way to interpret the whole genome sequencing results by viewing them as "continuous" signals on chromosomes. However, signal processing techniques alone might be insufficient in classifying regulatory elements in our proposed method. 1-D convolutional neural network in conjunction with the preprocessing step by our newly developed Markov model may enable us to solve the challenging inference problem; this is because we are aware that the prominence of preprocessing step in machine learning application and that convolutional neural network can effectively detect conserved shapes of massive putative signals which may significantly improve our inference accuracy. The completion of this project will delineate a comprehensive genome-wide PRE mapping in the human genome. It will have a positive impact on our field by guiding the discovery and prioritization of disease-associated REs, such as by overlapping the RE mapping with disease profiles like GWAS SNPs.

# COMPREHENSIVE IDENTIFICATION AND FUNCTIONAL ANALYSIS OF CIS REGULATORY ELEMENTS AND 3D ARCHITECTURE OF THE PIG GENOME

Yunxia Zhao[1], Ye Hou[1], Mingyang Hu[1], Yueyuan Xu[1], Yu Luan[1], Tingting Liu[2], Hongbo Yang[2], Feng Yue[2], Xinyun Li[1], Shuhong Zhao[1]

[1]Key Lab of Agricultural Animal Genetics, Breeding, and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan, China, [2]Department of Biochemistry and Molecular Biology, College of Medicine, The Pennsylvania State University, Hershey, PA

Despite the pig is an important species in the livestock industry, it is also increasingly used as human disease model and resource of xenotransplantation research. Thus, understanding the genome of pig is crucial for pig biology and human health. Here, we took a similar approach adopted by the ENCODE and Roadmap Epigenomics projects to address the cis Regulatory Elements (CRE) of pig genome. We performed RNAseq, ATAC-seq and ChIP-seq for H3K27ac and H3K4me3 histone markers to generate a comprehensive map of transcriptomes, open chromatin regions, regulatory elements in a variety of pig tissues including skeletal muscle, spleen, heart, kidney, liver, backfat, lung, small intestine (duodenum), thymus, cerebrum and cerebellum from lean type breeds Large White, Duroc, and fatty type breeds Meishan and Enshi Black pigs. We obtained over 15,000 CRE in the pig genome, which is the most comprehensive functional annotation effort have been made so far in the pig. By comparing the data from the ENCODE and Roadmap Epigenomics projects, we found over 53% of pig CREs were functionally conserved between pig and human which is obviously higher than that between mouse and human (~45%). Moreover, we also found ~68% of human RegulomeDB type 1a SNPs were covered by pig conserved CREs and ~77% genotype of them were same between human and pig genome. Hi-C experiment was also performed to further explore the three dimension (3D) structure of pig genome by using skeletal muscle tissue in this study. Our HiC matrix showed that over 79% of boundaries of topologically-associating domains (TAD) were also functional conserved among pig and human which indicated the TAD were highly conserved between pig and human. However, conservation analysis showed there are 14 human TAD are rearranged by pig inter TAD boundaries, genes and enhancers located in these TADs are significantly associated with human phonotypes of head, such as, abnormality of the central nervous system, abnormality of the face (including eye, nose, mouth etc.) and Schinzel-giedion Midface Retraction Syndrome human disease. Overall, our study provided the first comprehensive map of cis regulatory elements and the 3D architechture of the pig genome. Our comparative genimoics analysis also fills gaps between human and pig genome, further validated pig as valuable model as human disease and biology study model.

# PROFILING OF CHROMATIN ACCESSIBILITY AND IDENTIFICATION OF GENERAL CIS-REGULATORY MECHANISMS THAT CONTROL TWO OCULAR LENS DIFFERENTIATION PATHWAYS

Yilin Zhao, Deyou Zheng, Ales Cvekl

Albert Einstein College of Medicine, Genetics, Bronx, NY

Cell differentiation is driven by elaborate chromatin structure and gene transcription dynamics. At the DNA level, transcription is regulated by cis-regulatory sequences called the promoters and enhancers that interact with transcription factors (TFs) to control gene expression. At the level of chromatin, those regions display reduced nucleosome density (open chromatin). To "open" chromatin, it has been proposed that "pioneer" TFs can bind to closed chromatin, recruit chromatin remodeling enzymes and additional TFs to facilitate transcription. The spatial-temporal chromatin dynamics of in vivo cell differentiation is not well understood due to the complexity of cell types in most tissues. Here we use lens as our research model due to its simplicity of two cell types coupled with ATAC-seq to investigate the dynamic chromatin profiles during mouse lens epithelial and fiber cell differentiation from embryonic (E) day 14.5 to postnatal (P) day 0.5. Comparing to embryonic stem (ES) cell, forebrain and liver chromatin profiles, our analysis revealed thousands of lens specific open chromatin regions, including potential enhancers of important lens genes such as Prox1, Foxe3 and Mip. Combining with gene expression profiling by RNA-seq, we identified 26 and 49 TFs with binding motifs enriched at open chromatin regions that were expressed significantly higher in fiber and epithelial cells respectively. Among them, Pax6 is a key TF that regulates differentiation of both lens epithelium and fiber through recruitment of p300, CBP, Mll1, Mll2, SWI/SNF/Brg1, and ISWI/Snf2h enzymes. A cross examination with Pax6 binding data from ChIP-seq analysis demonstrated that Pax6 bound to both open and closed chromatin regions in the lens and forebrain. Also, Pax6-bound regions showed predominantly DNA hypomethylation in forebrain. In summary, our data provides in depth analysis of chromatin profile and gene expression changes for lens epithelium and fiber differentiation and maturation. The present study also uncovers novel molecular functions of Pax6, namely binding at open and closed chromatin regions. Together, our data and models provide new avenues to conduct mechanistic studies of transcription control regions, reconstruction of gene regulatory network for lens morphogenesis and identification of cataract-causing mutations in functional non-coding regions. Future studies will be aimed to explore the spatial temporal activities of the predicted enhancers at Prox1, Foxe3 and Mip and potential roles of Pax6 as a pioneer TF.

# NOVEL MOLECULAR SIGNATURES FOR HEART FAILURE: A COHORT STUDY OF mRNA, lncRNA, AND microRNA EXPRESSION

Yin-Ying Wang[1], Xiumeng Hua[2], Qing Xiong[1], Peilin Jia[1], Jiangping Song[2], Zhongming Zhao[1]

[1]The University of Texas Health Science Center at Houston, Center for Precision Health, School of Biomedical Informatics, Houston, TX, [2]Chinese Academy of Medical Sciences and Peking Union Medical College, National Center for Cardiovascular Diseases, Fuwai Hospital, Beijing, China

Heart failure (HF) is a chronic condition characterized by structural and functional impairment of the heart. HF has been a major cause of hospitalization, morbidity and mortality. Although numerous advances have been made, the molecular mechanisms and the molecular signatures underlying HF remain largely unknown, especially the role of fibrotic genes in HF progression. In this study, we aim to discover molecular markers in the cellular condition during heart failure in the forms of protein-coding genes, microRNAs, and long noncoding RNAs (lncRNAs). We conducted whole transcriptome sequencing (mRNA and lncRNA) as well as microRNA sequencing of heart tissues from a cohort of 21 HF patients and 9 healthy donors. We developed an in-house transcriptomic data analysis pipeline to obtain the gene expression levels for each molecular type. As a result, we identified 126 protein-coding genes, 42 microRNAs, and 16 lncRNAs that were differentially expressed in the HF patients versus the controls. These differentially expressed genes were further found to be enriched in the regulation of Insulin-like Growth Factor (IGF) transport and regulation of blood pressure, confirming their potential roles in HF. Interestingly, some of them (e.g., KLF4, CDKN1 and MOCS1) showed significant correlation with fibrosis percentage, which is the most important feature of HF. These results supported these genes to be involved in the HF pathogenesis. We further developed an integrative method to investigate the co-regulation between microRNA, lncRNA and mRNA. In the resultant co-regulatory network, we identified several differentially expressed microRNAs (e.g., miR-190-5p, miR-129-5p and mir-1-3p), lncRNAs (e.g., LOC101926975, PDZRN3-AS1 and BANCR) and genes (e.g., JAK2, PROS1 and BDNF) that likely played significant co-functional roles in HF condition. Finally, some of the results were validated by a proteomic dataset (4 matched HF samples and one mixed control of 5 matched samples) based on TMT quantitation. In summary, our results revealed the potential novel molecular signatures and regulatory roles involved in the pathogenesis of HF, providing a better molecular understanding of it.

# HYBRID *DE NOVO* AND REFERENCE ASSISTED GENOME ASSEMBLY WITH MaSuRCA

Aleksey V Zimin

Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

The relatively low cost of third generation long high error read DNA sequencing data from Pacific Biosciences and Oxford Nanopore platforms has made them accessible to a wide community of researchers studying plant and animal genomes. De novo genome assemblies that use these data are usually very contiguous with N50 contig sizes frequently in megabase-range. However, due to fairly high error rates of the sequencing reads, 10-15%, sufficient contig consensus quality for these assemblies can now be only achieved by either using very high (>60-80x) coverage by Pacbio data, which makes the assembly project expensive, or by augmenting the long read data with low-cost Illumina reads. The MaSuRCA genome assembler had been developed from the ground up to utilize the hybrid approach to assembly, which uses Illumina reads early in the assembly process to error-correct the long high error reads. This enables one to produce good quality assemblies even from the low coverage long read data sets (10-20x) combined with 60-100x Illumina data. These low coverage data sets produce poor quality assembly results when assembled by themselves, without Illumina data. MaSuRCA assemblies achieve over 99.995% quality in consensus sequence, measured by alignment of the original Illumina data to the assembly. In addition, when a high quality reference genome of one or more related species is available, MaSuRCA can use these references as "long reads" and produce highly contiguous assemblies from Illumina data alone.
I will present the latest capabilities of MaSuRCA assembler and results both from assembly of hybrid long read+Illumina short read data and reference assisted assembly. MaSuRCA is an open source software and it is freely available from github https://github.com/alekseyzimin/masurca.

# SHOULD “IMPORTANT GENES” ACCOUNT FOR GREATER HERITABILITY IN COMPLEX TRAITS?

Guy Sella[1,2], Yuval B Simons[3], Jonathan K Pritchard[3,4]

[1]Columbia University, Biological Sciences, New York, NY, [2]Columbia University, Systems Biology, New York, NY, [3]Stanford University, Genetics, Stanford, CA, [4]Stanford University, Biology, Stanford, CA

The premise for genome-wide association studies (GWAS) was that heritable variation affecting complex traits would cluster in pathways or around “core” genes that directly affect these traits. Instead, statistical analyses of GWAS indicate that heritable variation is often widely distributed across the genome, with surprisingly little enrichments for such core genes or pathways (Loh et al. Nat. Genet. 2015; Shi et al. AJHG 2016; Boyle et al. Cell 2017). Pritchard and colleagues (Boyle et al. Cell 2017; Liu et al. bioRxiv 2019) recently suggested that these observations can be explained if heritable variation in the expression of core genes arises primarily from many eQTLs that are cis to other genes, but have weak trans effects on the expression of core genes. There is, nonetheless, an evolutionary piece missing in this picture. The contribution of an eQTL to heritable variation depends on its effect on the trait but also on its frequency, which is largely determined by the selection acting on it. The strength of selection reflects the effect of the eQTL on the trait in question and its pleiotropic effects on other selected traits. Notably, because of selection, “important genes", whose expression levels directly affect the value of a trait, may harbor rarer variants and thus contribute less to heritability than would be expected in the absence of selection (Boyle et al. Cell 2017; O’Connor et al. bioRxiv 2018; Liu et al. bioRxiv 2019). To examine this possibility, we consider how the effect of a gene’s expression on a given trait is expected to influence the contribution of its cis eQTLs to heritable variation in the trait, assuming stabilizing selection on the focal and on other traits. Our results make testable predictions about the way in which the heritability of complex traits should be distributed across the genome.

# POOR PORTABILITY OF POLYGENIC SCORES EVEN WITHIN AN ANCESTRY GROUP

Hakhamanesh Mostafavi*[1], Arbel Harpak*[1], Dalton Conley[2,3], Jonathan K Pritchard[4,5,6], Molly Przeworski[1,7]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]Princeton University, Department of Sociology, Princeton, NJ, [3]Princeton University, Office of Population Research, Princeton, NJ, [4]Stanford University, Department of Genetics, Stanford, CA, [5]Stanford University, Department of Biology, Stanford, CA, [6]Howard Hughes Medical Institute, Stanford, CA, [7]Columbia University, Department of Systems Biology, New York, NY

* These authors contributed equally

Fields as diverse as human genetics and sociology are increasingly using polygenic scores obtained from genome-wide association studies (GWAS) for phenotypic prediction. For these applications, polygenic scores need to remain reliable across groups other than the one in which the GWAS was conducted. Recent discussion about this challenge has focused primarily on the portability of polygenic scores across genetic ancestries and in particular on the impact of differences in linkage disequilibrium patterns and allele frequencies. Here, we show that prediction accuracy can differ markedly even across groups with highly similar ancestry. First, using BMI, blood pressure and years of schooling in the UK Biobank as examples, we illustrate that prediction accuracy varies across groups that differ by common enrollment criteria of GWAS (such as age, sex, and socio-economic status), even when they share similar ancestry. We further demonstrate that the prediction accuracy of a polygenic score depends on whether the GWAS is conducted among unrelated individuals or within sibling pairs in the UK Biobank, even when allele frequencies and LD patterns are the same in the two approaches and power and sampling error are matched. We derive analytic results that clarify how indirect parental effects and assortative mating can lead to such differences in the prediction accuracies of polygenic scores. By providing evidence that the reliability of polygenic scores can vary substantially within an ancestry, our findings highlight underappreciated obstacles to their broad use.

# QUANTIFYING THE HERITABILITY OF THE GUT MICROBIOTA USING 13,780 LONGITUDINAL MICROBIOME PROFILES

Laura Grieneisen[1], Mauna Dasari[2], Trevor Gould[1], Johannes Bjork[2], Jean-Christophe Grenier[3], Vania Yotova[3], David Jansen[2], Neil Gottel[4], Jack Gilbert[4], Luis Barreiro[3], Susan C Alberts[5], Jeanne Altmann[6], Jenny Tung[5], Elizabeth Archie[2], Ran Blekhman[1,7]

[1]University of Minnesota, Genetics, Cell Biology, and Development, Minneapolis, MN, [2]University of Notre Dame, Biology, Notre Dame, IN, [3]University of Montreal, CHU Sainte Justine Research Center, Montreal, Canada, [4]University of Chicago, Surgery, Chicago, IL, [5]Duke University, Biology and Evolutionary Anthropology, Durham, NC, [6]Princeton University, Ecology and Evolutionary Biology, Princeton, NJ, [7]University of Minnesota, Ecology, Evolution, and Behavior, St. Paul, MN

Variation in gut microbiome composition—defined as the identity and abundance of different bacterial taxa—is linked to key health outcomes in humans and other vertebrates. Variation in the gut microbiome is affected by many environmental factors, but host genetics may also play an important role. Understanding the relative roles of genetic and environmental factors is a central goal in human disease research. However, it is often difficult to disentangle genetic from environmental effects because relatives often use the same resources and have similar diets. Moreover, it is difficult to account for long-term changes in microbiome composition since longitudinal profiles are usually not available.

Here, we quantified the heritability of the microbiome using an unprecedented dataset of 13,780 microbiome profiles collected over 14 years from a longitudinally studied population of wild baboons (Papio cynocephalus) with available host genetic relatedness data and rich environmental metadata. We identified components of the gut microbiome that are associated with host genetics—i.e. heritable microbes—while controlling for non-genetic effects by including granular metadata on host diet, social group, and weather. We found that >95% (146/153) of the analyzed taxa are significantly heritable, although heritability is generally low, and host genetics accounts for 9-16% of the variation in the abundances of the 15 most heritable taxa. These taxa included Christensenellaceae and Rikenellaceae, bacterial families that are also heritable in humans, suggesting that common host genetic mechanisms may control microbiome composition across primates. Finally, the relative abundance of heritable taxa changes over time within individuals, leading to temporal variation in estimated heritability. This emphasizes the importance of longitudinal studies to account for temporal and environmental effects in shaping gut microbial composition.

# ACCURATE ESTIMATION OF SNP-HERITABILITY FROM BIOBANK-SCALE DATA IRRESPECTIVE OF GENETIC ARCHITECTURE.

Bogdan Pasaniuc[1], Kathryn S Burch[1], Kangcheng Hou[1], Arunabha Majumdar[1], Huwenbo Shi[2], Nicholas Mancuso[1], Yue Wu[1], Sriram Sankararaman[1]

[1]UCLA, Computational Medicine, Los Angeles, CA, [2]Harvard School of Public Health, Biostatistics, Boston, MA

The proportion of phenotypic variance attributable to the additive effects of a given set of genotyped SNPs (i.e. SNP-heritability) is a fundamental quantity in the study of complex traits. All existing methods for genome-wide SNP-heritability inference make various assumptions on the underlying genetic architecture of the trait (e.g., the number of causal effects and the coupling of effect sizes with MAF/LD, or other functional genomic annotations). Since the true genetic architecture of any given trait is unknown, existing methods are susceptible to bias and often yield vastly different estimates of SNP-heritability making unclear which biobank SNP-heritability estimates are reliable.

In this work, we investigate SNP-heritability estimation under a generalized random effects (GRE) model that makes minimal assumptions on the genetic architecture of complex traits: every causal effect can have an arbitrary SNP-specific variance. To the best of our knowledge, all existing methods make additional assumptions on top of the GRE model. For example, the infinitesimal model assumed by GREML imposes an inverse relationship between MAF and effect size whereas the LDAK model assumes that each SNP-specific variance is inversely proportional to both MAF and the LD neighborhood. We derive a closed-form estimator for SNP-heritability under the GRE and show that this estimator is unbiased and does not depend on the underlying genetic architecture of the trait. We provide theoretical derivations and extensive simulations across a wide range of MAF- and LD-dependent architectures in the UK Biobank to find that the GRE estimator provides nearly unbiased estimates across all architectures whereas existing methods are sensitive to model misspecification. Across 126 distinct architectures, the maximum bias we observe for the GRE is 2% of the SNP-heritability whereas state-of-the-art methods such as stratified LD score regression (S-LDSC) and SumHer yield biases between -84% and 28% of the simulated SNP-heritability. In empirical data we find that S-LDSC (controlling for the baseline-LD model) and SumHer differ from the GRE estimates by a median of -9% and 11%, respectively. E.g., for height, S-LDSC (0.56) and SumHer (0.63) are 7% lower and 5% higher, than the GRE estimate (0.60). Taken together SNP-heritability can be accurately estimated from biobank-scale data without prior knowledge of the genetic architecture the trait, motivating the development of new biobank methods with fewer modeling assumptions.

# MICROBIOME-HOST INTERACTION STUDIED DIRECTLY IN HUMANS

Joao Xavier

Memorial Sloan Kettering Cancer Center, Systems Biology, New York, NY

The gut microbiome protects its human host against invasion by pathogens. Antibiotics, however, can compromise this protection and lead to resistant infections. We study this phenomenon in patients hospitalized to receive bone marrow transplantation, where due to their immune-compromised state the microbiome protection is even more important. This talk will describe lessons learned from analyses of thousands of patients, and the results of a pilot trial to restore microbiome composition with autologous microbiota transplantation. The unique patient set allows us to dissect mechanisms of host-microbiome interaction directly in humans with implications for new microbiome-based therapies that improve patient outcomes.

# UPSTREAM STIMULATORY FACTORS ARE PLEIOTROPIC META REGULATORS

Hanna M Ollila[1,2], Nasa Sinnott-Armstrong[2], Kirsi-Marja Zitting[3], Tiina Paunio[4], Tarja Stenberg[5], Richa Saxena[1], Jonathan Pritchard[2]

[1]Massachusetts General Hospital, Center for Genomic Medicine, Boston, MA, [2]Stanford University, Department of Genetics, Palo Alto, CA, [3]Harvard Medical School, Division of Sleep and Circadian Disorders, Boston, MA, [4]University of Helsinki, Department of Psychiatry, Helsinki, Finland, [5]University of Helsinki, Department of Physiology, Helsinki, Finland

Pervasive pleiotropy is a consistent phenomenon across nearly all human traits. Similarly, master regulators are thought to have specific roles in different diseases, characteristic of the organ systems in which they function. This poses a paradox, which we aimed to dissect by looking at sleep, which has substantial genetic correlation with psychiatric, metabolic and immune traits. Timing and amount of sleep are controlled by a set of well-characterized master regulators. At the same time, other factors including USF1 have been shown to modify both sleep and metabolic traits in mammals. Here we examine the pleiotropic and sleep specific effects of USF transcription factors in circadian rhythms, metabolic traits and immune cell fate. We show that in addition to metabolic effects Usf1 knock-out mice have fragmented sleep ($p<0.01$), disrupted circadian rhythms ($p<0.01$) and defective sleep consolidation ($p<0.01$). In addition, neuron specific knock down in flies recapitulates this phenotype. Applying these results to human traits, we show a strong enrichment of USF1 and USF2 binding sites in numerous human sleep traits. USF binding sites explained 10-20% of SNP-based heritability for morningness-eveningness, insomnia, excessive daytime sleepiness and napping ($p<0.001$). Furthermore, pleiotropic effects were seen with immune cell counts ($p<0.001$) height ($p<0.001$) and BMI ($p<0.001$). Indeed, the effect of USFs on sleep were modulated by both specific and pleiotropic variants. This was confirmed with allele specific activity at canonical circadian regulators such as PER1 and PER3, as well as metabolic targets including FTO and FADS1. Our findings provide evidence that pleiotropic effects of sleep and other functions of USFs separately contribute to the heritability of diverse complex traits.

# THE GENETIC ARCHITECTURE AND PHENOTYPIC PREDICTION OF BLOOD CELL PRODUCTION

Dragana Vuckovic

on behalf of BCX Consortium; Wellcome Sanger Institute, Human Genetics, Cambridge, United Kingdom

Blood cells play an essential role in immunity and in basic physiological processes, such as oxygen transport. The blood cell composition has been associated with severe disorders such as anaemias, coagulation defects and cardiovascular disease. Hence,a greater understanding of the genetic regulation of blood indices impacts the understanding of predisposition to diseases with a large healthcare burden. Here we exploit data from UK Biobank and a large-scale international collaborative effort, the Blood Cell Consortium (BCX, N=563,085 European Ancestry) to discover 6,230 new genetic variants independently associated to 29 blood cell phenotypes, including white, red cells and platelets, and to characterise the genetic architecture and the prediction potential for blood cell variation. We fine-map 1,423 signals to a single genetic variant with high posterior probability of causality (>95%). We integrate the fine-mapping results with clinical annotations from ClinVar and HGMD and observe an average 2.4x difference in absolute effect sizes for high confidence pathogenic variants compared to benign. We exploit this shift to re-assign pathogenicity for several variants with uncertain annotation. Using ATAC-seq from 18 primary hematopoietic populations and integrating transcription factor motif and gene annotations we nominate novel functional regulatory variants. We find significant and lineage-specific enrichment of fine-mapped variants in ATAC-seq peaks. We further annotate results with cis and trans cell-type regulatory data (eQTLs, promoter-interacting regions and co-expression networks), in order to partition variants into core and peripheral, as per the proposed omnigenic model. Using 342 known Mendelian genes as putative core, we find significant enrichment for GWAS associations in both core (p=1.7E-4) and first and second “neighbour” genes in the co-expression network (p=7E-11). We then explore how much phenotypic variance is captured by genome-wide variation compared to subsets of SNPs selected ad-hoc based on high posterior probability or annotations. We use discovery estimates derived from the statistically powered UK Biobank to build genetic scores (GS) in a validation cohort (INTERVAL, N=43,059). We find that a GS computed on putative causative variants explains up to 30% of phenotypic variance and is always more predictive than the genome-wide LD pruned approach.
In conclusion, we present the largest set of new GWAS associations for blood traits and show their predictive potential. We identify putative disease-causing variants and regulatory elements and combine the two sets in one biological model, at the interface of complex and Mendelian genetics.

# LONGITUDINAL ANALYSIS OF THE GUT MICROBIOME REVEALS DYNAMIC CHANGES IN RELATION TO MEDICATIONS AND PHENOTYPES

Lianmin Chen, Sanzhima Garmaeva, Alexander Kurilshikov, Ranko Gacesa, Arnau Vich Vila, Rinse Weersma, Cisca Wijmenga, Alexandra Zhernakova, Jingyuan Fu

University Medical Center Groningen, Department of Genetics & Pediatrics & Gastroenterology, Groningen, Netherlands

Interaction between the human genome and gut microbiome is vital to human health. While the human genome is set at birth, the gut microbiome can undergo dynamic changes over the course of an individual's life. However, we still know little about long-term temporal shifts in the human gut microbiome, nor about what drives of these changes or their potential influence on host health and disease.

We have now followed 341 participants of the Dutch population-based Lifelines-DEEP cohort for ~5 years. Each individual has been deeply phenotyped for various lifestyle, clinical and physiological factors at two time points. Metagenomics sequencing of faecal samples has allowed us to assess the changes in microbial taxonomic and functional composition, microbial genetic variability at strain level and the abundance of antibiotics resistance and virulence factors.

We find that temporal changes in the gut microbiome are dynamic. Significant intra-abundance shifts were observed in 40% of the microbial taxonomies and 60% of the functional pathways at FDR<0.05. These changes can be linked to an individual's lifestyle and phenotypic changes. For instance, we found a positive correlation between changes in the bacterial sugar derivatives degradation pathway and BMI, suggesting a causal relationship. We also observe a significant increase in the abundance of 40% of antibiotic resistance genes (FDR<0.05), even though antibiotic usage in the Netherlands is the lowest in Europe. In contrast, the abundance of virulence factors was stable.

We also assessed the genetic stability of dominant strains for more than 40 species, and observed significant difference between species. For example, Ruminococcus torques strains were characterized by a high mutation rate (4.12%), while Methanobrevibacter smithii strains were genetically stable (99.89%). Interestingly, we also noticed that the mutation rate of strains can vary between individuals, which may due to specific lifestyle and phenotypic changes. For instance, in individuals with larger changes in BMI, the genetic variation of Coprococcus sp ART55/1 strains was higher. In proton pump inhibitor and non-steroidal anti-inflammatory drug users, Streptococcus thermophilus and Bifidobacterium longum showed higher mutation rates than in non-users.

This study is the first to examine 5-year temporal changes in the human gut microbiome. These findings yield novel insights into the gut microbiome's impact on the development of complex disease and traits over time.

# TECHNICAL ABSTRACTS
# FOR WORKSHOPS

WORKSHOP ON THE USE OF HIGHLY ACCURATE LONG READS FOR COMPREHENSIVE GENOMIC ANALYSIS

Pacific Biosciences, Menlo Park, CA

Presented by Jason Underwood Ph.D.

In this workshop, we will highlight the latest enhancements to PacBio® Single Molecule, Real-Time (SMRT®) Sequencing technology, including:

- Overview of the new Sequel® II System which increases data output by approximately 8-fold with the new SMRT Cell 8M, alongside with improvements in library preparation and data analysis protocols.

- Highly accurate long reads (HiFi reads) which allow for comprehensive variation detection in human, plant, and animal genomes, from single nucleotide polymorphisms to structural variants, using a single data type.

- Other applications leveraging highly accurate sequence data, including targeted and shotgun metagenomics, massively parallel amplicon sequencing for full-length, phased gene sequencing, and full-length RNA sequencing using the Iso-Seq® method.

- New advances in single cell RNA sequencing which allows for the characterization of full-length mRNA isoforms identified from individual cells.

# THE REDWOOD GENOME PROJECT: SEQUENCING AND ASSEMBLING THE LARGEST TREES IN THE WORLD

Oxford Nanopore Workshop

Featuring: Steven Salzberg, John Hopkins University

Abstract: The Redwood Genome Project is a five-year effort that will sequence the coast redwood and giant sequoia genomes and develop tools to assess genetic diversity. Redwoods (Sequoia sempervirens) and sequoias (Sequoiadendron giganteum) are the two largest trees in the world, and among the longest-lived, with life spans exceeding 3000 years. To re-establish the genetic diversity of the remaining redwood forests, we have undertaken an effort to sequence the genomes of both species. Sequoia has an unusually large genome, estimated at 9 gigabases, and redwood is three times larger, 27 gigabases, due to its hexaploid genome. We have undertaken a strategy that combines Illumina and Oxford Nanopore data for assembly, following by the use of long-range linking information from Hi-C libraries and scaffolding by Dovetail Genomics. This talk will describe the just-finished sequoia assembly and the ongoing effort to assemble the redwood tree.

**NOTES**

**NOTES**

**NOTES**

**NOTES**

**NOTES**

## Participant List

Dr. Haley Abel
Washington University in St. Louis
abelhj@gmail.com

Dr. Ryan Abo
Ribon Therapeutics
rabo@ribontx.com

Dr. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Dr. Aaron Adamson
City of Hope
aadamson@coh.org

Dr. Vikram Agarwal
University of Washington
vagar@uw.edu

Dr. Saumya Agrawal
RIKEN
saumya.agrawal@riken.jp

Dr. Frank Albert
University of Minnesota
falbert@umn.edu

Ms. Maris Alver
Institute of Genomics, University of Tartu
maris.alver@ut.ee

Dr. Carl Anderson
Wellcome Sanger Institute
ca3@sanger.ac.uk

Dr. Lara Rubio Arauna
Institut Pasteur
lara.rubio-arauna@pasteur.fr

Dr. Elena Arciero
Wellcome Sanger Institute
ea6@sanger.ac.uk

Mr. Ricard Argelaguet
European Bioinformatics Institute (EMBL-EBI)
ricard@ebi.ac.uk

Dr. Alexander Arguello
National Instittutes of Health/ NIMH
alexander.arguello@nih.gov

Dr. Taejeong Bae
Mayo Clinic
bae.taejeong@mayo.edu

Dr. Orli Bahcall
Nature
o.bahcall@us.nature.com

Dr. Matthew Bailey
University of Utah
matthew.bailey@genetics.utah.edu

Dr. Sara Ballouz
Cold Spring Harbor Laboratory
sballouz@cshl.edu

Dr. Nicholas Banovich
Translational Genomics Research Institute
nbanovich@tgen.org

Dr. Luis Barreiro
University of Chicago
lbarreiro@uchicago.edu

Mr. Floris Barthel
The Jackson Laboratory
floris.barthel@jax.org

Dr. Elizabeth Thomas Bartom
Northwestern University
ebartom@northwestern.edu

Dr. Anastasia Baryshnikova
Calico Life Sciences
abaryshnikova@calicolabs.com

Mr. Sanjit Batra
University of California at Berkeley
sanjitsbatra@berkeley.edu

Dr. Alexis Battle
Johns Hopkins University
alexisjane@gmail.com

Dr. Serafim Batzoglou
Illumina, Inc.
sbatzoglou@illumina.com

Dr. Brittany Baur
University of Wisconsin-Madison
babaur@wisc.edu

Dr. Michael Beer
Johns Hopkins University
mbeer@jhu.edu

Mr. Ivan Berest
EMBL
berest@embl.de

Dr. Nathan Berkowitz
ancestry.com
nathan.berkowitz.gcb@gmail.com

Dr. Søren Besenbacher
Aarhus University
besenbacher@clin.au.dk

Dr. Preetida Bhetariya
University of Utah
preetiaahir@gmail.com

Dr. Minou Bina
Purdue University
bina@purdue.edu

Dr. Rebecca Birnbaum
Icahn School of Medicine at Mount Sinai
rebecca.birnbaum@mssm.edu

Dr. Ewan Birney
EMBL-EBI
birney@ebi.ac.uk

Dr. Ran Blekhman
University of Minnesota
blekhman@umn.edu

Mr. Daniel Blibaum
Cold Spring Harbor Lab
blibaum@cshl.edu

Mr. Amit Blumberg
Cold Spring Harbor Laboratory
ablumber@cshl.edu

Dr. Jamie Blundell
University of Cambridge
jrb75@cam.ac.uk

Dr. Christoph Bock
CeMM Research Center for Molecular Medicine
cbock@cemm.oeaw.ac.at

Prof. Hidemasa Bono
Research Organization of Information and Systems
bono@dbcls.rois.ac.jp

Dr. Leandros Boukas
Johns Hopkins University
leandrosboukas@gmail.com

Dr. Alan Boyle
University of Michigan
apboyle@umich.edu

Mr. James Brayer
Oxford Nanopore Technologies Ltd
James.Brayer@nanoporetech.com

Dr. Martin Breuss
University of California, San Diego
mbreuss@ucsd.edu

Dr. Christian Brion
University of Minnesota
chbrion@umn.edu

Ms. Sarah Brotman
University of North Carolina-Chapel Hill
sbrotman@email.unc.edu

Dr. Christopher Brown
University of Pennsylvania
casey6r0wn@gmail.com

Dr. David Burkhardt
GRAIL
dburkhardt@grailbio.com

Dr. Mario Caceres
Universitat Autonoma de Barcelona
mcaceres@icrea.cat

Dr. Na Cai
Wellcome Sanger Institute
nc10@sanger.ac.uk

Dr. Thomas Callaghan
FBI
tfcallaghan@fbi.gov

Ms. Fiona Calvert
Wellcome Sanger Institute
fc5@sanger.ac.uk

Dr. Jedidiah Carlson
University of Washington
jed.e.carlson@gmail.com

Dr. Piero Carninci
RIKEN
carninci@riken.jp

Mr. Christopher Castro
University of Michigan
castrocp@umich.edu

Ms. Maria Izabel Cavassim
Aarhus University
mica@birc.au.dk

Dr. Ines Cebola
Imperial College London
idossant@imperial.ac.uk

Dr. Lisa Chadwick
National Human Genome Research Institute
Lisa.Chadwick@nih.gov

Dr. Vered Chalifa-Caspi
Ben-Gurion University
veredcc@bgu.ac.il

Dr. Frank Chan
Friedrich Miescher Laboratory
frank.chan@tue.mpg.de

Ms. Christina Chang
National Human Genome Research Institute
christine.chang2@nih.gov

Dr. Sumantra Chatterjee
NYU School of Medicine
sumantra.chatterjee@nyulangone.org

Dr. Aikaterini Chatzipli
Wellcome Sanger Centre
ac33@sanger.ac.uk

Dr. Junjie Chen
University of North Carolina at Charlotte
jchen67@uncc.edu

Dr. Minhui Chen
University of Southern California
minhuic@usc.edu

Ms. Lei Chen
Washington University in St Louis
leichen@wustl.edu

Mr. Lianmin Chen
University of Groningen
lianminchen@yeah.net

Ms. Xiaoheng Cheng
Penn State University
xh.cheng@psu.edu

Dr. Charleston Chiang
University of Southern California
charleston.chiang@med.usc.edu

Mr. Jeremy Choin
Institut Pasteur
jeremy.choin@pasteur.fr

Mr. Saket Choudhary
University Of Southern California
skchoudh@usc.edu

Mr. Kapeel Chougule
Cold Spring Harbor Laboratory
kchougul@cshl.edu

Dr. Ryan Christ
Washington University School of Medicine
rchrist@wustl.edu

Mr. V. Kartik Chundru
University of Queensland
v.chundru@uq.edu.au

Dr. Ryan Clark
University of Wisconsin-Madison
rlclark2@wisc.edu

Ms. Tal Cohen
Ben Gurion University of the Negev
cotal@post.bgu.ac.il

Dr. Leonardo Collado Torres
Lieber Institute for Brain Development
leo.collado@libd.org

Mr. Ryan Collins
Harvard Medical School
rlcollins@g.harvard.edu

Dr. Laureen Connell
Molecular Case Studies, Senior Editor
connell@cshl.edu

Dr. Montserrat Corominas
Universitat de Barcelona
mcorominas@ub.edu

Dr. Christine Couldrey
LIC
christine.couldrey@lic.co.nz

Dr. Alexssander Couto Alves
Imperial College London
alexessander.alves@gmail.com

Mr. Hongzhu Cui
Worcester Polytechnic Institute
hcui2@wpi.edu

Ms. Anna Cuomo
EMBL-EBI, University of Cambridge
acuomo@ebi.ac.uk

Ms. Hannah Currant
European Bioinformatics Institute (EMBL-EBI)
currant@ebi.ac.uk

Ms. Ciara Curtin
GenomeWeb
ccurtin@genomeweb.com

Dr. Michal Dabrowski
Institute of Computer Science PAS
m.dabrowski@ipipan.waw.pl

Ms. Anupriya Dalmia
German Centre for Neurodegenerative Diseases
anupriya.dalmia@dzne.de

Dr. Matteo D'Antonio
UCSD
madantonio@ucsd.edu

Ms. Julianne David
Oregon Health and Science University
davidju@ohsu.edu

Dr. Benjamin Decato
Bristol-Myers Squibb
benjamin.decato@bms.com

Dr. Jacob Degner
Abbvie
jfdegner@gmail.com

Dr. Ricardo del Rosario
Broad Institute
rcdelros@broadinstitute.org

Dr. Jennifer DeLeon
Genome Research
deleon@cshl.edu

Mr. Deniz Demircioglu
Genome Institute of Singapore
ddemircioglu@gis.a-star.edu.sg

Dr. Ahmet Denli
Genome Research, Associate Editor
denli@cshl.edu

Prof. Emmanouil Dermitzakis
University of Geneva and Health 2030 Genome Center
emmanouil.dermitzakis@unige.ch

Dr. Alexander Dobin
Cold Spring Harbor Laboratory
dobin@cshl.edu

Mr. Ricardo D'Oliveira Albanus
University of Michigan
albanus@umich.edu

Ms. Julia Domingo
Center of Genomic Regulation - CRG
julia.domingo.espinos@gmail.com

Dr. Vera Domingues
Springer Nature
vera.domingues@nature.com

Prof. Peter Donnelly
University of Oxford
donnellypa@well.ox.ac.uk

Dr. Noah Dukler
Cold Spring Harbor Laboratory
ndukler@cshl.edu

Dr. Harly Durbin
University of Missouri
hjdurbin@mail.missouri.edu

Dr. Richard Durbin
University of Cambridge
rd109@cam.ac.uk

Mr. Aditya Ekawade
University of Utah School of Medicine
aditya.ekawade@gmail.com

Dr. Christine Eng
Genome Institute of Singapore
christine_eng@gis.a-star.edu.sg

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Ms. Melissa Englund
University of Michigan
melyssae@umich.edu

Dr. Michael Erdos
National Human Genome Research Institute
mikee@mail.nih.gov

Dr. Eleazar Eskin
UCLA
eeskin@cs.ucla.edu

Ms. Susan Fairley
EMBL EBI
fairley@ebi.ac.uk

Dr. Kai-How Farh
Illumina
kfarh@illumina.com

Dr. Andrew Farrell
University of Utah
jandrewrfarrell@gmail.com

Dr. Elise Feingold
National Human Human Reseach Institute
feingole@mail.nih.gov

Ms. Joanna Ficek
ETH Zurich
joanna.ficek@inf.ethz.ch

Mr. Anthony Findley
Wayne State University
afindley@med.wayne.edu

Prof. Liliana Florea
Johns Hopkins University
florea@jhu.edu

Mr. Andre Flores-Bello
Institute of Evolutionary Biology (CSIC-UPF)
andre.flores@upf.edu

Ms. Elise Flynn
Columbia University
eflynn@nygenome.org

Dr. Adam Frankish
EMBL EBI
frankish@ebi.ac.uk

Ms. Karin Fredrikson
Oxford Nanopore Technologies Ltd
karin.fredrikson@nanoporetech.com

Dr. Robin Friedman
Ohana Biosciences
rfriedman@ohanabio.com

Dr. Rebecca Furlong
SpringerNature
rebecca.furlong@nature.com

Dr. Idan Gabdank
Stanford
gabdank@stanford.edu

Dr. Pedro A Galante
Hospital Sírio-Libanês
pgalante@mochsl.org.br

Ms. Isabel Gamache
Montreal Heart Institute
Isabel.Gamache@mhi-rc.org

Dr. Manuel Garber
University of Massachusetts Medical School
manuel.garber@umassmed.edu

Dr. Eugene Gardner
Wellcome Sanger Institute
eg15@sanger.ac.uk

Ms. Manik Garg
EMBL-EBI
manikg@ebi.ac.uk

Dr. Audrey Gasch
University of Wisconsin-Madison
agasch@wisc.edu

Dr. Laufey Geirsdottir
Weizmann Institute of Science
laufeygeirsd@me.com

Dr. Sarah Geisler
Cell Press
sgeisler@cell.com

Mr. Evan Geller
Yale University
evan.geller@yale.edu

Dr. Giulio Genovese
Broad Institute
giulio.genovese@gmail.com

Ms. Stephanie Georges
University of Utah
stephanie.georges@genetics.utah.edu

Dr. Michel Georges
University of Liege / GIGA Institute
michel.georges@uliege.be

Dr. Mark Gerstein
Yale University
pi@gersteinlab.org

Dr. Sulagna Ghosh
The Broad Institute
ghoshs@broadinstitute.org

Dr. Maya Ghoussaini
Wellcome Sanger Institute
mg29@sanger.ac.uk

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Dr. David Gifford
Massachusetts Institute of Technology
gifford@csail.mit.edu

Prof. Jesse Gillis
Cold Spring Harbor Laboratory
jesse.gillis@gmail.com

Dr. Andreas Gnirke
Broad Institute
gnirke@broadinstitute.org

Dr. Ellen Greytak
Parabon NanoLabs, Inc.
ellen@parabon.com

Dr. Roderic Guigo
Centre for Genomic Regulati9on (CRG)
roderic.guigo@crg.cat

Dr. Rodrigo Gularte Merida
Memorial Sloan-Kettering Cancer Center
gularter@mskcc.org

Dr. Karl Guo
GlaxoSmithKline
karl.x.guo@gsk.com

Dr. Ira Hall
Washington Univerisity St. Louis
s.cody@wustl.edu

Mr. Robert Handsaker
Broad Institute
handsake@broadinstitute.org

Ms. Juliet Handsaker
Wellcome Trust Sanger Institute
jh34@sanger.ac.uk

Mr. Adam Hansen
Baylor College of Medicine
adam.hansen@bcm.edu

Dr. Tim Harkins
Medicinal Genomics, Inc.
timothy.harkins@medicinalgenomics.com

Dr. Chad Harland
LIC NZ
chad.harland@lic.co.nz

Dr. Arbel Harpak
Columbia University
ah3586@columbia.edu

Dr. Christopher Hart
Ionis Pharmaceuticals
chart@ionisph.com

Mr. Joe Harvey
Oxford Nanopore Technologies Ltd
Joe.Harvey@nanoporetech.com

Dr. Shinichi Hashimoto
Kanazawa University
hashimoto@med.kanazawa-u.ac.jp

Ms. Laura Hayward
Columbia University
lauhayward@gmail.com

Dr. James Hazel
Vanderbilt University Medical Center
james.w.hazel.1@vumc.org

Ms. Yuan He
Johns Hopkins University
yuanhe716@gmail.com

Mr. Meng Xiao He
Harvard University
mxhe@broadinstitute.org

Dr. Xin He
University of Chicago
xinhe@uchicago.edu

Dr. Hussein Hejase
Cold Spring Harbor Laboratory
hijazi@cshl.edu

Dr. Javier Herrero
UCL Cancer Institute
javier.herrero@ucl.ac.uk

Mr. Michael Heskett
Oregon Health and Science University
heskett@ohsu.edu

Dr. Peter Heutink
DZNE
peter.heutink@dzne.de

Dr. Naoki Hirose
RIKEN
naoki.hirose@riken.jp

Dr. Eurie Hong
Ancestry
ehong@ancestry.com

Ms. Katie Houlahan
Ontario Institute for Cancer Research
Katie.Houlahan@oicr.on.ca

Dr. Xiaomeng Huang
University of Utah
xm01.huang@gmail.com

Dr. Matthew Hurles
Wellcome Trust Genome Campus
meh@sanger.ac.uk

Dr. Julie Hussin
Universite de Montreal
julie.hussin@umontreal.ca

Dr. Carolyn Hutter
National Human Genomic Research Institute
carolyn.hutter@nih.gov

Ms. Elizabeth Hutton
Cold Spring Harbor Laboratory
ehutton@cshl.edu

Dr. Oleg Iartchouk
Novartis
oleg.iartchouk@novartis.com

Dr. Kazuki Ichikawa
The University of Tokyo
ichikawa@cb.k.u-tokyo.ac.jp

Ms. Anastasia Illarionova
German center for neurodegenerative diseases DZNE
Anastasia.Illarionova@dzne.de

Mr. Daniel Irizarry
Stony Brook University
daniel.irizarry@stonybrook.edu

Dr. Farren Isaacs
Yale University
farren.isaacs@yale.edu

Dr. Sadahiro Iwabuchi
Kanazawa University
s_iwabuchi@staff.kanazawa-u.ac.jp

Dr. Lakshmanan Iyer
LabCorp Inc
iyerl@labcorp.com

Ms. Shruti Iyer
Cold Spring Harbor Lab / Stony Brook University
iyer@cshl.edu

Dr. Andrew Jaffe
Lieber Institute for Brain Development
andrew.jaffe@libd.org

Dr. Julie Jerber
Wellcome Sanger Institute
jj11@sanger.ac.uk

Dr. Peilin Jia
UTHealth
peilin.jia@uth.tmc.edu

Dr. Shan Jiang
The University of Texas Health Science Center at Houston
shan.jiang@uth.tmc.edu

Ms. Ruth Johnson
University of California, Los Angeles
ruthjohnson@g.ucla.edu

Ms. Molly Johnson
New York Genome Center
mjohnson@nygenome.org

Mr. Otto Jolanki
Stanford University
ojolanki@stanford.edu

Dr. Luke Jostins
Oxford University
Luke.jostins@kennedy.ox.ac.uk

Ms. Jenna Kalleberg
University of Missouri
jakth2@mail.missouri.edu

Dr. Fiona Kaper
Illumina
fkaper@illumina.com

Dr. Sekar Kathiresan
Massachusetts General Hospital
sekar@broad.mit.edu

Prof. Sara Katsanis
Duke University
sara.katsanis@duke.edu

Dr. Thomas Keane
EMBL-EBI
tk2@ebi.ac.uk

Mr. Benjamin Keith
The University of North Carolina at Chapel Hill
bpk@unc.edu

Dr. Manolis Kellis
MIT
manoli@mit.edu

Mr. Matthew Kerin
University of Oxford
kerin@stats.ox.ac.uk

Mr. Daniel Kim
Stanford School of Medicine
danielskim@stanford.edu

Ms. Young Kim
Stony Brook University
yschonkim5@gmail.com

Dr. Sergey Koren
NIH
sergey.koren@nih.gov

Dr. Anat Kreimer
University of California, Berkeley and UCSF
anat.kreimer@gmail.com

Dr. Anna Kuchina
University of Washington
akuchina@uw.edu

Mr. Lukas Kuderna
Institut de Biologia Evolutiva (UPF-CSIC)
lukas.kuderna@upf.edu

Dr. Vivek Kumar
Cold Spring Harbor Laboratory
vkumar@cshl.edu

Dr. Sofia Kyriazopoulou Panagiotopoulou
Illumina
spanagiotopoulou@illumina.com

Dr. Kasper Lage
Broad Institute / Mass General / Harvard Med School
lage.kasper@mgh.harvard.edu

Dr. Marina Laplana
Autonomous University of Barcelona
marinalaplana@gmail.com

Dr. Matthew Larson
GRAIL
mlarson@grail.com

Dr. Ryan Layer
University of Colorado, Boulder
ryan.layer@gmail.com

Dr. Sangwon Lee
Agilent Technologies
sangwon.lee@agilent.com

Ms. Joyce Lee
BioNano Genomics
jlee@bionanogenomics.com

Mr. Jon Lerga
Universitat Autonoma de Barcelona
jlergajaso@gmail.com

Dr. Ruijuan Li
inari Agriculture
rli@inari.com

Mr. Jun Li
University of Michigan
junzli@med.umich.edu

Ms. Khine Zin Lin
Stanford University School of Medicine
khzlin@stanford.edu

Dr. Stephen Lincoln
Invitae
steve.lincoln@me.com

Dr. Nicole Lockhart
NIH/NHGRI
lockhani@mail.nih.gov

Dr. Kirk Lohmueller
UCLA
klohmueller@ucla.edu

Mr. Matthew Lollar
University of Wisconsin-Madison
mlollar@wisc.edu

Dr. Quan Long
University of Calgary
quan.long@ucalgary.ca

Mr. Zhenyuan Lu
Cold Spring Harbor Laboratory
luj@cshl.edu

Dr. Francesca Luca
Wayne State University
fluca@wayne.edu

Mr. Nikolaos Lykoskoufis
University of Geneva
Nikolaos.Lykoskoufis@etu.unige.ch

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Dr. Heather Machado
Welcome Sanger Institiute
hm8@sanger.ac.uk

Dr. Elisa Majounie
Illumina
emajounie@illumina.com

Ms. Allison Mandich
National Human Genome Research Institute
allison.mandich@nih.gov

Dr. Serghei Mangul
UCLA
smangul@ucla.edu

Prof. Elaine Mardis
Nationwide Children's Hospital
elaine.mardis@nationwidechildrens.org

Dr. Lasse Maretty
Aarhus University
lasse.maretty@clin.au.dk

Dr. Georgi Marinov
Stanford University
marinovg@stanford.edu

Prof. Gabor Marth
University of Utah
gabor.marth@gmail.com

Dr. Hilary Martin
Wellcome Sanger Institute
hcm@sanger.ac.uk

Mr. Carl-Adam Mattsson
University of British Columbia
cmattsson@bccrc.ca

Dr. Steven McCarroll
Harvard Medical School, Broad Institute
mccarroll@genetics.med.harvard.edu

Dr. Richard McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Ms. Laura McCulloch
NYU Langone Health
laura.mcculloch@nyulangone.org

Mr. Torrin McDonald
University of Michigan
torrin@umich.edu

Dr. David McKinnon
Stony Brook University
david.mckinnon@stonybrook.edu

Dr. Graham McVicker
Salk Institute for Biological Studies
gmcvicker@salk.edu

Dr. Wenbin Mei
Inari Agriculture
wmei@inari.com

Dr. Maria Mejia Guerra
Cornell University
mm2842@cornell.edu

Dr. Amanda Melillo
HHS/NIH
amanda.melillo@nih.gov

Mr. Dan Melodia
Oxford Nanopore Technologies Ltd

Dr. Javier Mendoza Revilla
Institut Pasteur
javier.mendoza-revilla@pasteur.fr

Dr. Mait Metspalu
Institute of Genomics, University of Tartu
mait.metspalu@ut.ee

Dr. Michael Metzger
Pacific Northwest Research Institute
metzgerm@pnri.org

Dr. Wynn Meyer
University of Pittsburgh
wmeyer@pitt.edu

Dr. Stephen Meyn
University of Wisconsin - Madison
stephen.meyn@wisc.edu

Dr. David Micklos
Cold Spring Harbor Laboratory
micklos@chl.edu

Dr. Ryan Mills
University of Michigan
remills@umich.edu

Dr. Rob Mitra
Washington University School of Medicine
rmitra@wustl.edu

Ms. Briana Mittleman
University of Chicago
brimittleman@uchicago.edu

Mr. Taku Monjo
Takeda Pharmaceutical Company Limited
taku.monjo@takeda.com

Dr. Michael Morgan
Wellcome Sanger Institute
mm33@sanger.ac.uk

Dr. Shinichi Morishita
University ot Tokyo
moris@edu.k.u-tokyo.ac.jp

Prof. Leonid Moroz
Univeristy of Florida
moroz@whitney.ufl.edu

Mr. Arnav Moudgil
Washington University in St. Louis
amoudgil@wustl.edu

Ms. Rachel Murphy
UMASS Medical School
rachel.murphy@umassmed.edu

Dr. Tendai Mutangadura
University of Missouri
tendai@missouri.edu

Dr. Luis A Nagai
University of Tokyo
nagailae@hgc.jp

Dr. Fabio Navarro
Yale University
fabio.navarro@yale.edu

Dr. Stanley Neufeld
Max Planck Society
stanley.neufeld@tuebingen.mpg.de

Dr. Luong Nguyen
AncestryDNA LLC
lnguyen@ancestry.com

Ms. Sara Nilson
University of Missouri
smnzr7@mail.missouri.edu

Mr. Jin Woo Oh
Johns Hopkins University
joh27@jhmi.edu

Dr. Hanna Ollila
MGH, Broad institute
hollila@broadinstitute.org

Mr. Andrew Olson
Cold Spring Harbor Laboratory
olson@cshl.edu

Dr. Nathan Olson
National Institute of Standards and Technology
nolson@nist.gov

Dr. Halit Ongen
University of Geneva
halit.ongen@unige.ch

Mr. Omead Ostadan
Illumina, Inc.
oostadan@illumina.com

Dr. Elaine Ostrander
National Institutes of Health
eostrand@mail.nih.gov

Dr. Toshio Ota
Kyowa Hakko Kirin
toshio.ota@kyowa-kirin.co.jp

Ms. Ningxin Ouyang
University of Michigan
nouyang@umich.edu

Dr. Athma Pai
UMass Medical School
athma.pai@umassmed.edu

Dr. Wenying Pan
Grail.Inc
wpan@grailbio.com

Dr. Qingfei Pan
St. Jude Children's Research Hospital
QINGFEI.PAN@HOTMAIL.COM

Dr. Kevin Pang
Genome Biology
kevin.pang@genomebiology.com

Dr. Nikolaos Panousis
Wellcome Sanger Institute
np12@sanger.ac.uk

Dr. Sung-Joon Park
The University of Tokyo
sjpark@ims.u-tokyo.ac.jp

Dr. Bogdan Pasaniuc
UCLA
pasaniuc@ucla.edu

Dr. Ayoti Patra
Johns Hopkins University
apatra2@jhu.edu

Dr. Dana Pe'er
Sloan Kettering Institute, MSKCC
peerster@gmail.com

Dr. Guangsheng Pei
The University of Texas Health Science Center at Houston
Guangsheng.Pei@uth.tmc.edu

Mr. Bernard Pereira
Novartis Institutes for Biomedical Research
bernard.pereira@novartis.com

Dr. Giulio Pergola
Lieber Institute for Brain Development, JHU
giulio.pergola@libd.org

Ms. Hannah Perrin
University of North Carolina at Chapel hill
hjperrin@email.unc.edu

Mr. Trenton Peters-Clarke
University of Wisconsin-Madison
petersclarke@wisc.edu

Dr. Lon Phan
NIH/NCBI
lonphan@ncbi.nlm.nih.gov

Dr. Olli Pietilainen
Harvard University
ollip@broadinstitute.org

Mr. Stefan Pietrzak
University of Wisconsin-Madison
spietrzak@wisc.edu

Dr. Luca Pinello
MGH/ Harvard
lpinello@mgh.harvard.edu

Dr. Lenore Pipes
University of California-Berkeley
lpipes@berkeley.edu

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Dr. Rahul Pisupati
Gregor Mendel Institute
rahul.pisupati@gmi.oeaw.ac.at

Mr. Anders Pitman
University of Utah
tapitman11@gmail.com

Dr. David Porubsky
University of Washingon
porubsky@uw.edu

Dr. Jeremie Poschmann
Centre de Recherche en Transplantation et Immunolo
jeremie.poschmann@univ-nantes.fr

Ms. Laurie Prelot
ETH Zurich
laurie.prelot@inf.ethz.ch

Dr. Amanda Price
Lieber Institute for Brain Development
amanda.joy.price@gmail.com

Dr. Molly Przeworski
Columbia University
mp3284@columbia.edu

Dr. Yi Qiao
University of Utah
yi.qiao@genetics.utah.edu

Dr. Letu Qingge
University of North Carolina at Charlotte
letu.qingge@uncc.edu

Prof. Francis Quetier
GENOPOLE-Evry & University of of Evry
francis.quetier@gmail.com

Dr. Aaron Quinlan
University of Utah
aaronquinlan@gmail.com

Mr. Nima Rafati
Uppsala University
nima.rafati@igp.uu.se

Dr. Arjun Raj
University of Pennsylvania
arjunraj@seas.upenn.edu

Dr. Gokul Ramaswami
UCLA
GRamaswami@mednet.ucla.edu

Dr. Gunnar Ratsch
ETH Zurich
raetsch@ethz.ch

Dr. Sarah Ratzel
American Journal of Human Genetics
sratzel@ajhg.net

Dr. Jüri Reimand
Ontario Institute for Cancer Research
juri.reimand@utoronto.ca

Dr. Jie Ren
Cold Spring Harbor Laboratory
jren@cshl.edu

Ms. Justyna Resztak
Wayne State University
ga6092@wayne.edu

Dr. Arang Rhie
National Institutes of Health
rhiea@nih.gov

Ms. Teomie Rivera-Miranda
University of Puerto Rico Rio Piedras
teomie.rivera@upr.edu

Dr. Patrizia Rizzu
DZNE
patrizia.rizzu@dzne.de

Dr. Roland Roberts
Public Library of Science (PLOS)
rroberts@plos.org

Ms. Catherine Robertson
University of Virginia
ccr5ju@virginia.edu

Dr. Christopher Robles
UCLA
chrisrobles@ucla.edu

Ms. Nicole Rockweiler
Washington University School of Medicine
nrockweiler@wustl.edu

Dr. Juan Rodriguez-Flores
Weill Cornell Medical College
jur2014@med.cornell.edu

Dr. Jose Rodriguez-Martinez
University of Puerto Rico - Rio Piedras
jose.rodriguez233@upr.edu

Dr. Masa Roller
EMBL-EBI
roller@ebi.ac.uk

Dr. Mostafa Ronaghi
Illumina, Inc.
mronaghi@illumina.com

Dr. Ian Rose
Cornell University
imr923@gmail.com

Dr. Jeffrey Rosenfeld
Rutger Cancer Institute of NJ
jeffrey.rosenfeld@rutgers.edu

Dr. Mahboubeh Rostami
Weill Cornell
mrr2006@med.cornell.edu

Dr. Troy Rowan
University of Missouri
troyrowan@mail.missouri.edu

Dr. Tanmoy Roychowdhury
University of Michigan
tanmoy63@gmail.com

Dr. Joel Rozowsky
Yale University
joel.rozowsky@yale.edu

Dr. Marina Ruiz-Romero
Center for Genomic Regulation (CRG)
marina.ruiz@crg.eu

Dr. Henry Sadowski
Bionano Genomics
hsadowski@bionanogenomics.com

Mr. Yoshitaka Sakamoto
The University of Tokyo
sakamoto_yoshitaka_18@stu-cbms.k.u-tokyo.ac.jp

Ms. Sharadha Sakthikumar
Uppsala University
sakthi@broadinstitute.org

Dr. Leonidas Salichos
Yale University
leonidas.salichos@yale.edu

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Dr. Ashley Sanders
EMBL
ashley.sanders@embl.de

Mr. Ioannis Sarropoulos
University of Heidelberg
i.sarropoulos@zmbh.uni-heidelberg.de

Ms. Ambra Sartori
University of Geneva
ambra.sartori@unige.ch

Mr. Thomas Sasani
University of Utah
tom.sasani@utah.edu

Ms. Natalia Savytska
DZNE(German Center for Neurodegenerative Diseases)
Natalia.Savytska@dzne.de

Dr. Michael Schatz
CSHL + JHU
mschatz@cshl.edu

Dr. Mikkel Schierup
Aarhus University
mheide@birc.au.dk

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Dr. David Schwartz
University of Wisconsin-Madison
dcschwartz@wisc.edu

Dr. Daniel Seaton
European Bioinformatics Institute
dseaton@ebi.ac.uk

Dr. Karl Sebby
xD Bio, Inc.
karl.sebby@xdbio.com

Dr. Shobana Sekar
Mayo Clinic
sekar.shobana@mayo.edu

Dr. Guy Sella
Columbia University
gs2747@columbia.edu

Mr. Arko Sen
Salk Institute for Biological Studies
asen@salk.edu

Mr. Sarun Sereewattanawoot
The University of Tokyo
sereewattanawoot_sarun_15@stu-cbms.k.u-tokyo.ac.jp

Dr. Brittany Sexton
New England Biolabs
bsexton@neb.com

Ms. Minita Shah
New York Genome Center
mshah@nygenome.org

Mr. Maxwell Shapiro
Stony Brook University
maxwell.shapiro@stonybrook.edu

Dr. Jason Sheltzer
Cold Spring Harbor Laboratory
sheltzer@cshl.edu

Dr. Ning Shen
Fulcrum Therapeutics
nshen@fulcrumtx.com

Dr. Jay Shendure
University of Washington
shendure@uw.edu

Dr. Xinghua Shi
UNC Charlotte
x.shi@uncc.edu

Dr. Dustin Shigaki
Johns Hopkins University
dshigak1@jhu.edu

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Ms. Olga Sigalova
EMBL
olga.sigalova@embl.de

Dr. Yuval Simons
Stanford University
Yuval.simons@stanford.edu

Dr. Alexei Slesarev
Bioreliance
alexei.slesarev@gmail.com

Dr. Jeramiah Smith
University of Kentucky
jjsmit3@uky.edu

Dr. Valentina Snetkova
Lawrence Berkeley National Lab
vsnetkova@lbl.gov

Dr. Michael Snyder
Stanford University
mpsnyder@stanford.edu

Dr. Caetano Souto-Maior
National Heart Lung and Blood Institute - NIH
caetano.soutomaior@nih.gov

Mr. Max Stammnitz
University of Cambridge
mrs72@cam.ac.uk

Dr. Heather Stark
Pacific Biosciences
hstark@pacb.com

Dr. Cody Steely
University of Utah
csteely4@gmail.com

Dr. Arnold Stein
Purdue University
steina@purdue.edu

Dr. Joshua Stein
Cold Spring Harbor Laboratory
steinj@cshl.edu

Dr. Kathleen Steinmann
Broad Institute
kathleen@broadinstitute.org

Dr. J Seth Strattan
Stanford University
strattan@gmail.com

Dr. Richard Straub
Lieber Institute for Brain Development
richard.straub@libd.org

Mr. Benjamin Strober
Johns Hopkins University
bstrober3@gmail.com

Dr. Hillary Sussman
Genome Research, Executive Editor
hsussman@cshl.edu

Dr. Milovan Suvakov
Mayo Clinic
suvakov.milovan@mayo.edu

Dr. Yutaka Suzuki
University of Tokyo
ysuzuki@hgc.jp

Dr. Dmitry Svetlichnyy
National University Higher School of Economics
dvsvetlichny@gmail.com

Dr. David Symer
Univ. Texas MD Anderson Cancer Center
desymer@mdanderson.org

Dr. Michael Talkowski
Massachusetts General Hospital
talkowski@chgr.mgh.harvard.edu

Mr. Yosuke Tanigawa
Stanford University
ytanigaw@stanford.edu

Ms. Georgette Tanner
Leeds Institute of Medical Research at St James's
medgnt@leeds.ac.uk

Dr. Manuel Tardaguila
Wellcome Sanger Institute
mt19@sanger.ac.uk

Dr. James Taylor
Johns Hopkins University
james@jamestaylor.org

Dr. Martin Taylor
University of Edinburgh
martin.taylor@igmm.ed.ac.uk

Dr. Natalie Telis
AncestryDNA, LLC
natalie.telis@gmail.com

Dr. Marcela Tello-Ruiz
Cold Spring Harbor Laboratory
telloruiz@cshl.edu

Mr. Serhat Tetikol
Seven Bridges
serhat.tetikol@sbgenomics.com

Mr. William Thomas
Stony Brook University
william.thomas@stonybrook.edu

Dr. James Thomas
NIH/NHGRI
thomasjw4@mail.nih.gov

Dr. Ammon Thompson
University of California, Davis
ammonthompson@gmail.com

Dr. Joe Thornton
University of Chicago
joet1@uchicago.edu

Dr. David Thybert
Earlham Institute
david.thybert@earlham.ac.uk

Dr. Hagen Tilgner
Weill Cornell Medicine
hut2006@med.cornell.edu

Dr. Nataliya Timoshevskaya
University of Kentucky
timnatevg@gmail.com

Ms. Susan Tousi
Illumina, Inc.
stousi@illumina.com

Dr. Michela Traglia
University of San Francisco
michela.traglia@ucsf.edu

Dr. Holly Trochet
University of Montreal
holly.trochet@umontreal.ca

Mr. Alex Tseng
Stanford University
amtseng@stanford.edu

Mr. Issei Tsuchiya
The University of Tokyo
isseiinchicago@yahoo.co.jp

Dr. Taru Tukiainen
University of Helsinki
taru.tukiainen@helsinki.fi

Prof. Jenny Tung
Duke University
jt5@duke.edu

Dr. Katarzyna Tyc
Rutgers University
katarzyna.tyc@rutgers.edu

Dr. Jason Underwood
PacBio
junderwood@pacb.com

Ms. Lea Urpa
Institute for Molecular Medicine, Finland
lea.urpa@helsinki.fi

Dr. Vladimir Vacic
23andMe, Inc.
vvacic@23andme.com

Dr. Anton Valouev
Grail Bio
valouev@gmail.com

Dr. Cristopher Van Hout
Regeneron Genetics Center
cristopher.vanhout@regeneron.com

Dr. Krishna Veeramah
Stony Brook University
krishna.veeramah@stonybrook.edu

Dr. Oliver Venn
GRAIL, INC.
ovenn@grailbio.com

Dr. Shefali Verma
Univerity of Pennsylvania
shefali.setiaverma@pennmedicine.upenn.edu

Ms. Elizabeth Vincent
Johns Hopkins School of Medicine
evincen4@jhmi.edu

Dr. Ana Vinuela
University of Geneva
ana.vinuela@unige.ch

Dr. Kristoffer Vitting-Seerup
University of Copenhangen, Denmark
kristoffer.vittingseerup@bio.ku.dk

Ms. Nadezda Volkova
EMBL-EBI
nvolkova@ebi.ac.uk

Dr. Dragana Vuckovic
Wellcome Sanger Institute
dv3@sanger.ac.uk

Dr. Deven Vyas
Stony Brook Univeristy
deven.vyas@stonybrook.edu

Ms. Rebecca Walker
UCLA
rebwalker@g.ucla.edu

Dr. Joselynn Wallace
Brown University
joselynn_wallace@brown.edu

Dr. Liya Wang
Cold Spring Harbor Laboratory
wangli@cshl.edu

Dr. Gao Wang
University of Chicago
wangow@gmail.com

Dr. Bo Wang
Cold Spring Harbor Lab
bwang@cshl.edu

Dr. Alistair Ward
University of Utah
AlistairNWard@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory/USDA/ARS
ware@cshl.edu

Dr. Varun Warrier
University of Cambridge
vw260@medschl.cam.ac.uk

Ms. Courtney Waterbury
University of Kentucky
ckwaterbury16@gmail.com

Dr. Robert Waterston
University of Washington
watersto@uw.edu

Ms. Sarah Watkins
University of Bristol
s.h.watkins@bristol.ac.uk

Ms. Kris Wetterstrand
National Human Genome Research Institute
wettersk@mail.nih.gov

Dr. Cristen Willer
University of Michigan
cristen@umich.edu

Dr. Louise Williams
New England Biolabs
lwilliams@neb.com

Dr. Richard Wilson
Nationwide Children's Hospital
richard.wilson@nationwidechildrens.org

Dr. Bartosz Wojtas
Nencki Institute of Experimental Biology
b.wojtas@nencki.gov.pl

Dr. John Wolters
University of Wisconsin-Madison
jwolters@wisc.edu

Dr. Hyejung Won
University of North Carolina at Chapel Hill
hyejung_won@med.unc.edu

Dr. Emily Wong
University of Queensland
emily.wong@uq.edu.au

Dr. Carrie Wright
Lieber Institute for Brain Development
carrie.wright@libd.org

Mr. Travis Wrightsman
Cornell University
tw493@cornell.edu

Dr. Catherine Wu
Dana-Farber Cancer Institute
cwu@partners.org

Dr. Yannick Wurm
Queen Mary University of London
y.wurm@qmul.ac.uk

Ms. Dana Wyman
University of California, Irvine
dwyman@uci.edu

Dr. Joao Xavier
Memorial Sloan Kettering Cancer Center
xavierj@mskcc.org

Dr. Wang Xi
Johns Hopkins University
wxi1@jhu.edu

Dr. Bingqing Xie
University of Chicago
bxie@uchicago.edu

Ms. Liu Xu
The University of Tokyo
leahxuliu@gmail.com

Prof. Qi Xu
Chinese academy of medical sciences
shalz_pumc@163.com

Dr. Alexander Xue
Cold Spring Harbor Laboratory
xanderxue@gmail.com

Prof. Huanming Yang
BGI-China
yanghm@genomics.cn

Dr. Xiaofei Yang
The Jackson Laboratory for Genomic Medicine
xiaofei.Yang@jax.org

Dr. Qiuming Yao
BCH, MGH, HMS
yao.ornl@gmail.com

Ms. Bin Ye
Regeneron Pharmacutical Inc.
bin.ye@regeneron.com

Dr. Muh-ching Yee
Invitae
muhching.yee@invitae.com

Mr. Toshiyuki Yokoyama
The University of Tokyo
2762556237@edu.k.u-tokyo.ac.jp

Dr. Jun Yoshimura
The University of Tokyo
yoshimura@cb.k.u-tokyo.ac.jp

Dr. Noha Yousri
Weill Cornell Medicine-Qatar
nay2005@qatar-med.cornell.edu

Ms. Zhezhen Yu
Cold Spring Harbor Laboratory
zhezhen@cshl.edu

Dr. George Yuan
Pacific Biosciences
gyuan@pacb.com

Dr. Laura Zahn
AAAS/ Science
lzahn@aaas.org

Dr. Noah Zaitlen
UCLA
nzaitlen@ucla.edu

Dr. Daniel Zerbino
EMBL-EBI
zerbino@ebi.ac.uk

Mr. Zijun Zhang
UCLA
zj.z@ucla.edu

Dr. Yunxia Zhao
Huazhong Agriculture University
yxzhao@mail.hzau.edu.cn

Dr. Shuhong Zhao
Huazhong Agriculture university
shzhao@mail.hzau.edu.cn

Ms. Yilin Zhao
Albert Einstein College of Medicine
yzhao2@mail.einstein.yu.edu

Dr. Yixin Zhao
Cold Spring Harbor Laboratory
yizhao@cshl.edu

Dr. Zhongming Zhao
University of Texas Health Science Center Houston
zhongming.zhao@uth.tmc.edu

Mr. Nanxiang (Samuel) Zhao
University of Michigan
samzhao@umich.edu

Dr. Martine Zilversmit
American Museum of Natural History
zmartine@gmail.com

Dr. Alexey Zimin
Johns Hopkins University
aleksey.zimin@gmail.com

Dr. Michael Zody
New York Genome Center
mczody@nygenome.org

## 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 | **631-351-2000** |
| **Dentists**<br>Dr. William Berg<br>Dr. Robert Zeman | <br>**631-271-2310**<br>**631-271-8090** |
| **Doctor**<br>MediCenter<br>234 W. Jericho Tpke., Huntington Station | **631-423-5400** |
| **Drugs - 24 hours, 7 days**<br>Rite-Aid<br>391 W. Main Street, Huntington | **631-549-9400** |

## GENERAL INFORMATION

**Books, Gifts, Snacks, Clothing, Newspapers**

*BOOKSTORE* 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

**Photocopiers, Journals, Periodicals, Books, Newspapers**

*Photocopying – Main Library*
*Hours:* 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
***Helpful tips* – Use PIN# 57080** to enter Library after hours.
See Library staff for photocopier code.

**Computers, E-mail, Internet access**

Grace Auditorium
Upper level: E-mail and printing in the business center area
STMP server address: mail.optonline.net
*To access your E-mail, you must know the name of your home server.*

**Dining, Bar**

Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late (Cash Only)
***Helpful tip*** - If there is a line at the upper dining area, try the lower dining room

**Messages, Mail, Faxes, ATM**

Message Board, Grace, lower level

**Swimming, Tennis, Jogging, Hiking**

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

**Russell Fitness Center**

Dolan Hall, east wing, lower level
***PIN#:*** **Press 57080 (then enter #)**

**Meetings & Courses Front Office**

**Hours during meetings: 8am – 7pm, until 9pm on arrival day**
***After hours – From guest house phones, dial x8870 for assistance***

**Pay Phones, House Phones**

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

**CSHL's Green Campus**

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

## 1-800 Access Numbers

**AT&T** **9-1-800-321-0288**

**Local Interest**

| | |
|---|---|
| Fish Hatchery | 631-692-6758 |
| Sagamore Hill | 516-922-4788 |
| Whaling Museum | 631-367-3418 |
| Heckscher Museum | 631-351-3250 |
| CSHL DNA Learning Center | x 5170 |

**New York City**

***Helpful tip -***

Take Syosset Taxi to Syosset Train Station
($9.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue).
Train ride about one hour.

## TRANSPORTATION

**Limo, Taxi**

| | |
|---|---|
| Syosset Limousine | 516-364-9681 |
| Executive Limo Service | 631-696-8000 |
| Super Shuttle | 800-957-4533 |
| Limos Long Island | 800-545-4667,ext.3 |

To head west of CSHL - Syosset train station

| | |
|---|---|
| Syosset Taxi | 516-921-2141 |

To head east of CSHL - Huntington Village

| | |
|---|---|
| Orange & White Taxi | 631-271-3600 |

**Trains**

| | |
|---|---|
| Long Island Rail Road | 822-LIRR |

*Schedules available from the Meetings & Courses Front Office.*

| | |
|---|---|
| Amtrak | 800-872-7245 |
| MetroNorth | 877-690-5114 |
| New Jersey Transit | 973-275-5555 |

**Ferries**

| | |
|---|---|
| Bridgeport / Port Jefferson | 631-473-0286 |
| Orient Point/ New London | 631-323-2525 |

**Car Rentals**

| | |
|---|---|
| Avis | 631-271-9300 |
| Enterprise | 631-424-8300 |
| Hertz | 631-427-6106 |

**Airlines**

| | |
|---|---|
| American | 800-433-7300 |
| British Airways | 800-247-9297 |
| Delta | 800-221-1212 |
| Japan Airlines | 800-525-3663 |
| Jet Blue | 800-538-2583 |
| KLM | 800-374-7747 |
| Lufthansa | 800-645-3880 |
| Southwest Airlines | 800-435-9792 |
| United | 800-241-6522 |
| Virgin American | 877-359-9792 |

# Code of Conduct

Cold Spring Harbor Laboratory is dedicated to pursuing its twin missions of research and education in the biological sciences. The Laboratory is committed to fostering a working environment that encourages and supports unfettered scientific inquiry and the free and open exchange of ideas that are the hallmarks of academic freedom. The Laboratory aims to maintain a safe and respectful environment for all attendees of our meetings and courses and associated support staff, by providing an environment free from discrimination and harassment, in accordance with federal, state, and local laws.

Consistent with the Laboratory's missions, commitments and policies, the purpose of this Code is to set forth the Laboratory's expectations for the professional conduct of individuals participating in the Laboratory's meetings program, including organizers, session chairs, invited speakers, presenters, attendees and sponsors.

By attending a CSHL meeting, participants agree to:

1. Treat fellow meeting participants and CSHL staff with respect, civility and fairness, and without bias based on race, color, religion, sex, national origin, citizenship status, sexual orientation, gender identity or expression, age, disability, marital status, veteran status, genetic information, or any other criteria prohibited under applicable federal, state or local law.

2. Use all CSHL facilities, equipment, computers, supplies and resources responsibly and appropriately, as you would at your home institution.

Meeting participants agree to refrain from:

1. Discrimination in violation of Laboratory policy based on age, gender, race, ethnicity, national origin, religion, disability, or sexual orientation.

2. Behavior that is disrespectful of others, and unprofessional interpersonal behavior that interferes with the working and learning environment.

3. Unwanted physical contact with others, or threats of such contact.

4. Sexual harassment or harassment based on age, gender, race, ethnicity, national origin, religion, disability or sexual orientation.

5. Loss of civility that interferes with the working and learning environment (for example shouting, personal attacks or insults, throwing objects or other displays of temper).

6. Misappropriation of Laboratory property or excessive personal use of resources.

## Breaches or Violations of the Code of Conduct

Cold Spring Harbor Laboratory aims to maintain a conference environment in accordance with the principles and expectations outlined in the Code of Conduct. Meeting organizers are tasked with providing leadership during each meeting, and may be approached informally about any breach or violation. Breaches or violations should also be reported to program leadership in person or by email:

- Dr. David Stewart, Grace Auditorium Room 204, 516-367-8801 or x8801 from a campus phone, stewart@cshl.edu

- Dr. Charla Lambert, Hershey Laboratory Room 214, 516-367-5058 or x5058 from a campus phone, clambert@cshl.edu

The Laboratory will take action as needed to resolve the matter, up to and including immediate expulsion of the offending participant(s) from the meeting, dismissal from the Laboratory, and exclusion from future academic events offered by CSHL.