

Abstracts of papers presented
at the 2023 meeting on

THE BIOLOGY OF GENOMES

May 9–May 13, 2023



Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2023 meeting on

THE BIOLOGY OF GENOMES

May 9–May 13, 2023

Christina Curtis, *Stanford University*

Hopi Hoekstra, *Harvard University*

Tuuli Lappalainen, *New York Genome Center & SciLife Lab*

John Marioni, *European Bioinformatics Institute (EMBL), UK*



Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Support for this meeting was provided in part by the **National Human Genome Research Institute (NHGRI)**, a branch of the **National Institutes of Health**; **PacBio**; **Oxford Nanopore Technologies**; **Complete Genomics**; and the **James P. Taylor Foundation for Open Science Scholarship Fund**.

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

Corporate Benefactors

Estée Lauder Companies
Regeneron
ThermoFisher Scientific

Corporate Sponsors

Agilent Technologies
Bayer
Biogen
Bristol-Myers Squibb Company
Calico Labs
Genentech *A member of the Roche Group*
Merck & Co., Inc.
Novartis Institutes for Biomedical Research

Corporate Partners

Alexandria Real Estate
Enzo Life Sciences
Gilead Sciences

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.

THE BIOLOGY OF GENOMES
Tuesday, May 9– Saturday, May 13, 2023

Tuesday	7:30 pm – 10:30 pm	1 Population Genomics
Wednesday	9:00 am – 12:00 pm	2 Developmental and Single Cell Genomics
Wednesday	1:00 pm – 1:45 pm	NHGRI Discussion Panel
Wednesday	2:00 pm – 5:00 pm	3 Computational Genomics
Wednesday	5:00 pm	<i>Wine & Cheese Party</i>
Wednesday	7:30 pm – 10:30 pm	Poster Session I
Thursday	9:00 am – 12:00 pm	4 Complex Traits and Microbiome
Thursday	1:30 pm – 4:30 pm	5 Functional Genomics and Epigenetics
Thursday	5:00 pm – 6:00 pm	ELSI Panel and Discussion
Thursday	7:30 pm – 10:30 pm	Poster Session II
Friday	8:30 am – 10:00 am	6 Systems Genetics
Friday	10:15 am – 12:15 pm	7 Evolutionary and Non-human Genomics
Friday	2:00 pm – 4:30 pm	Poster Session III
Friday	4:30 pm – 6:00 pm	GUEST SPEAKERS
Friday	6:30 pm	<i>Cocktails and Banquet</i>
Saturday	9:00 am – 12:00 pm	8 Cancer and Medical Genomics

Workshops (immediately following morning sessions)

Oxford Nanopore, Wednesday, May 10

PacBio, Thursday, May 11

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

All times shown are US Eastern: [Time Zone Converter](#)

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](#).



For further details as well as [Definitions and Examples](#) and how to [Report Violations](#), please see 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 9—7:30 PM

SESSION 1 POPULATION GENOMICS

Chairpersons: **Iain Mathieson**, University of Pennsylvania, Philadelphia
Shamil Sunyaev, Brigham and Women's Hospital,
Harvard Medical School, Boston, Massachusetts

Using ancient DNA to detect and understand recent natural selection in humans

Iain Mathieson.

Presenter affiliation: University of Pennsylvania, Philadelphia,
Pennsylvania.

1

Discovering epistasis between germline mutator alleles in mice

Thomas A. Sasani, Aaron R. Quinlan, Kelley Harris.

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

2

The landscape of tolerated genetic variation in humans and primates

Hong Gao, Tobias Hamp, Jeffrey Ede, Joshua G. Schraiber, Jeremy McRae, Moriel Singer-Berk, Yanshen Yang, Anastasia Dietrich, Petko Fiziev, Lukas Kuderna, Laksshman Sundaram, Yibing Wu, Aashish Adhikari, Yair Field, Jinbo Xu, Jeffrey Rogers, Tomas Marques-Bonet, Kyle Farh.

Presenter affiliation: Illumina, Inc., Foster City, California.

3

Genetic and evolutionary basis of population differences in immune response to respiratory viruses

Maxime Rotival, Yann Aquino, Aurélie Bisiaux, Zhi Li, Mary O'Neill, Javier Mendoza Revilla, Etienne Patin, Lluis Quintana-Murci.

Presenter affiliation: Institut Pasteur, Université Paris Cité, CNRS UMR2000, Paris, France.

4

Shamil Sunyaev.

Presenter affiliation: Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

Sex differences in genetic effects on complex traits

Carrie Zhu, Matthew J. Ming, Jared M. Cole, Michael D. Edge, Mark Kirkpatrick, Arbel Harpak.

Presenter affiliation: The University of Texas at Austin, Austin, Texas. 5

A Pan-pangenome captures the full spectrum of genetic variation and ancient trans-species structural polymorphism in humans, chimpanzees and bonobos

Joana L. Rocha, Juan M. Vazquez, Alison Killilea, Runyang Nicolas Lou, Stacy Li, Matthew W. Mitchell, Kendra Hoekzema, Evan E. Eichler, Peter H. Sudmant.

Presenter affiliation: UC Berkeley, Berkeley, California. 6

Single-cell RNA-seq in cell villages enables discovery of state-dependent eQTLs

Elizabeth T. Roberts, Susie Song, Matthew Tegtmeyer, Tommy Casolaro, Maddie Murphy, Xiaolan Zhang, Ralda Nehme, Eric Lander, Vidya Subramanian, Elisa Donnard, Thouis R. Jones.

Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 7

WEDNESDAY, May 10—9:00 AM

SESSION 2 DEVELOPMENTAL AND SINGLE CELL GENOMICS

Chairpersons: **Dan Landau**, Weill Cornell Medicine, New York, New York
Marta Mele, Barcelona Supercomputing Center, Spain

Transcriptional and epigenetic impact of cigarette smoking across human tissues

Jose Miguel Ramirez, Rogério Ribéiro, Oleksandra Soldatkina, Raquel Garcia, Pedro G. Ferreira, Marta Melé.

Presenter affiliation: Barcelona Supercomputing Center, Barcelona, Spain. 8

The function and decline of the female reproductive tract at single-cell resolution

Ivana Winkler, Alexander Tolkachov, Duncan T. Odom, Angela Goncalves.

Presenter affiliation: German Cancer Center, Heidelberg, Germany. 9

Genomic regulatory structure of the ENCODE4 mouse postnatal developmental time course at single-cell resolution reveals homologous regulatory topics within and across tissues

Elisabeth Rebboah, Narges Rezaie, Brian Williams, Annika K. Weimer, Heidi Y. Liang, Diane Trout, Fairlie Reese, Jasmine Sakr, Michael Snyder, Barbara Wold, Ali Mortazavi.

Presenter affiliation: UC Irvine, Irvine, California.

10

No cell left behind—Dynamic studies of gene regulation in humans

Yoav Gilad.

Presenter affiliation: University of Chicago, Chicago, Illinois.

11

Mapping somatic evolution with single cell multi-omics

Dan Landau.

Presenter affiliation: Weill Cornell Medicine, New York, New York.

Regression modeling of multiome data identifies functional enhancers and enables chromatin potential analysis

Sneha Mitra, Yuri Pritykin, Christina Leslie.

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

12

A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells

Li Li, Sarah Bowling, Qi Yu, Sean E. McGahey, Karel Alcedo, Bianca Lemke, Mark Ferreira, Allon M. Klein, Shou-Wen Wang, Fernando D. Camargo.

Presenter affiliation: Westlake University, Hangzhou, China.

13

Single cell atlas of cortical neuron development *in vitro*

Adithi Sundaresan, Dimitri Meistermann, Riina Lampela, Rosa

Woldegebriel, Pau Puigdevall Costa, Helena Kilpinen.

Presenter affiliation: University of Helsinki, Helsinki, Finland.

14

WEDNESDAY, May 10—1:00 PM

NHGRI LUNCHTIME PANEL DISCUSSION SESSION

Trainees and early career researchers: Tips for team science

Moderator: **Carolyn Hutter**, NHGRI, National Institutes of Health

Panelists: **Tuuli Lappalainen**, New York Genome Center

Doug Fowler, University of Washington, Seattle

Adam Felsenfeld, NHGRI, NIH

Are you a trainee or early career researcher who participates in, or is interested in, team science? Do you want to learn more about what makes or breaks large, interdisciplinary research collaborations? This session will discuss features that lead to successful collaborations – both from the perspective of researchers who have been involved in team science, as well as from NIH Program Officers who have helped manage large genomics initiatives. The session will include considerations relevant to early career stages. A panel discussion will provide you with the opportunity to ask your own questions.

WEDNESDAY, May 10—2:00 PM

SESSION 3 COMPUTATIONAL GENOMICS

Chairpersons: **Xihong Lin**, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

James Zou, Stanford University, California

Discovering disease-relevant spatial cellular motifs with graph deep learning on spatial omics

James Zou.

Presenter affiliation: Stanford University, Stanford, California.

15

PerturbDecode, a probabilistic analysis framework to recover regulatory circuits and predict genetic interactions from large-scale Perturb-Seq screens

Basak Eraslan, Kathryn Geiger-Schuller, Olena Kuksenko, Pratiksha Thakore, Ozge Karayel, Andrea Yung, Anugraha Rajagopalan, Ana Meireles de Sous, Karren Dai Yang, Liat Amir-Zilberstein, Toni Delorey, Devan Phillips, Rakimma Raychowdhury, Christine Moussion, Nir Hacohen, Caroline Uhler, Orit Rozenblatt-Rosen, Aviv Regev.
Presenter affiliation: Genentech, South San Francisco, California;
Broad Institute of MIT and Harvard, Cambridge, Massachusetts.

16

Learning multimodal cell trajectories in differentiating systems using single cell multiomic data

Alireza Karbalayghareh, Christopher Chin, Darko Barisic, Martin Rivas Ari Melnick, Christina Leslie.
Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York.

17

Complete genomes of a multi-generational pedigree to expand studies of genetic and epigenetic inheritance

Monika Cechova, Sergey Koren, Julian K. Lucas, Rebecca Serra Mari, Mobin Asri, David Porubsky, Andrey Bzikadze, Christopher Markovic, Tamara Potapova, Jennifer L. Gerton, Evan E. Eichler, Benedict Paten, Adam M. Phillippy, Ting Wang, Nathan O. Stitzel, Robert S. Fulton, Tobias Marschall, Karen H. Miga.

Presenter affiliation: UC Santa Cruz Genomics Institute, Santa Cruz, California.

18

cellSTAAR—Incorporating single-cell based cell-type specific functional data in rare variant association testing of non-coding regions of whole genome sequencing studies

Eric Van Buren, NHLBI Trans-Omics for Precision Medicine (TOPMed), Xihong Lin.

Presenter affiliation: Harvard TH Chan School of Public Health, Boston, Massachusetts; Harvard University, Cambridge, Massachusetts.

19

Genomics at scale with the NHGRI AnVIL

Michael Schatz, Anthony Philippakis.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

20

A novel method to account for fine-scale population structure in large-scale genomic analysis

Ruhollah Shemirani, Sinead Cullina, Gillian M. Belbin, Christopher R. Gignoux, Noah A. Zaitlen, Eimear E. Kenny.

Presenter affiliation: Icahn School of Medicine at Mount Sinai, New York, New York.

21

Cell-type-specific co-expression inference from single cell RNA-sequencing data

Chang Su, Zichun Xu, Xinning Shan, Biao Cai, Hongyu Zhao, Jingfei Zhang.

Presenter affiliation: Yale University, New Haven, Connecticut.

22

WEDNESDAY, May 10—5:00 PM

Wine and Cheese Party

WEDNESDAY, May 10—7:30 PM

POSTER SESSION I

See *p. xviii* for List of Posters

THURSDAY, May 11—9:00 AM

SESSION 4 COMPLEX TRAITS AND MICROBIOME

Chairpersons: **Ran Blekhman**, University of Chicago, Illinois

Melina Claussnitzer, Broad Institute, MGH / Harvard Medical School, Boston, Massachusetts

Ran Blekhman.

Presenter affiliation: University of Chicago, Chicago, Illinois.

Discovering stimulatory state specific T2D GWAS mechanisms with single cell multi-omics on iPSC-derived FAP villages

Christa Ventresca, Arushi Varshney, Peter Orchard, Yao-chang Tsan, Andre Monteiro da Rocha, Markku Laakso, Jaakko Tuomilehto, Timo A. Lakka, Karen L. Mohlke, Michael Boehnke, Laura Scott, Heikki A. Koistinen, Francis S. Collins, Todd Herron, Stephanie Bielas, Stephen Parker.

Presenter affiliation: University of Michigan, Ann Arbor, Michigan.

23

Efficient and accurate detection of viral sequences in bulk and single-cell transcriptomics data

Laura Luebbert, Delaney K. Sullivan, Lior Pachter.

Presenter affiliation: California Institute of Technology, Pasadena, California.

24

Functional dissection of complex and molecular trait variants at single nucleotide resolution

Hayley Siraj, Hannah Dewey, Susan Kales, Steven Reilly, Hilary Finucane, Jacob Ulirsch, Ryan Tewhey.

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

25

PRS2F—Linking polygenic risk scores for metabolic disease to biological function

Melina Claussnitzer.

Presenter affiliation: Broad Institute, MGH / Harvard Medical School, Boston, Massachusetts.

Diversity improves all aspects of genomic research—Lessons from 700,000 human exomes, genomes, and genotypes

Julia Sealock, Rahul Gupta, Katherine Chao, Masahiro Kanai, Siwei Chen, Kristin Tsuo, Wenhan Lu, gnomAD Consortium, Pan-UKB Team, Benjamin Neale, Heidi Rehm, Kaitlin Samocha, Mark Daly, Alicia Martin, Konrad Karczewski.

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

26

Milo2.0 unlocks population genetic analyses of cell state abundance using a counts-based mixed model

Alice Kluzer, John C. Marioni, Michael D. Morgan.

Presenter affiliation: Institute of Medical Sciences, Aberdeen, United Kingdom.

27

Cell type-specific and disease-associated eQTL in the human lung

Heini M. Natri, Christina Del Azodi, Lance Peter, Chase Taylor, Sagrika Chugh, Robert Kendle, Jonathan Kropski, Davis McCarthy, Nicholas E. Banovich.

Presenter affiliation: Translational Genomics Research Institute (TGen), Phoenix, Arizona.

28

SESSION 5 FUNCTIONAL GENOMICS AND EPIGENETICS

Chairpersons: **Douglas Fowler**, University of Washington, Seattle
Lars Velten, Centre for Genomic Regulation (CRG),
Barcelona, Spain

Targeted single cell methylome profiling reveals epigenetic encoding of hematopoietic stem cell fate

Michael Scherer, Indranil Singh, Agostina Bianchi, Chelsea Szu-Tu,
Roser Zaurin, Renée Beekman, Alejo Rodriguez-Fraticelli, Lars Velten.
Presenter affiliation: Centre for Genomic Regulation (CRG),
Barcelona, Spain.

29

Single cell sequencing as a universal variant interpretation assay

Dan Cao, Ken Jean-Baptiste, Mohan Sun, Xingyan Zhou, Sandy
Wong, Ling Chen, Hongxia Xu, Francois Aguet, Kyle Farh.
Presenter affiliation: Illumina, Inc., Foster City, California.

30

A chromosome-scale CRISPR screen to identify essential elements in the human noncoding genome

Xinyi Guo, Chirag Lakhani, Noa Liscovitch-Brauer, Qiyao Lin, Christina
Caragine, Giuseppe Narzisi, Edward Paik, Caroline Yu, David
Knowles, Neville E. Sanjana.

Presenter affiliation: New York Genome Center, New York, New York;
New York University, New York, New York.

31

Machine-guided design of *cis*-regulatory elements with cell-type specificity

Rodrigo I. Castro, Sager J. Gosai, Natalia Fuentes, Kousuke Mouri,
Pardis C. Sabeti, Steven K. Reilly, Ryan Tewhey.

Presenter affiliation: The Jackson Laboratory, Bar Harbor, Maine.

32

Understanding genetic variants at scale—From technology development to the clinic

Douglas Fowler.

Presenter affiliation: University of Washington, Seattle.

Extensive natural variation and silencing of lncRNAs in *Arabidopsis thaliana*

Aleksandra E. Kornienko, Viktoria Nizhynska, Magnus Nordborg.
Presenter affiliation: Gregor Mendel Institute, Vienna, Austria.

33

DragoNNFruit—Learning cis- and trans-regulation of chromatin accessibility at single base and single cell resolution

Jacob Schreiber, Surag Nair, Anshul Kundaje.

Presenter affiliation: Stanford University, Stanford, California.

34

Mapping the convergence of genes for coronary artery disease onto endothelial cell programs

Gavin R. Schnitzler, Helen Kang, Vivian S. Lee-Kim, Rosa X. Ma, Tony Zeng, Ramcharan S. Angom, Shi Fang, Shamsudheen K. Vellarikkal, Ronghao Zhou, Katherine Guo, Oscar Sias-Garcia, Alex Bloemendal, Glen Munson, Debabrata Mukhopadhyay, Eric S. Lander, Hilary K. Finucane, Rajat M. Gupta, Jesse M. Engreitz.

Presenter affiliation: Broad Institute, Cambridge, Massachusetts; Stanford University, Stanford, California; Betty Irene Moore Children's Heart Center, Stanford, California.

35

THURSDAY, May 11—5:00 PM

ELSI PANEL and DISCUSSION

Scientists' roles and responsibilities combatting the misuse of genomic research

Moderator: **Dave Kaufman**, NHGRI, National Institutes of Health

Panelists: **Kushan Dasgupta**, UCLA Institute for Society and Genetics
Daphne Martschenko, Stanford Center for Biomedical Ethics
Ashley Smart, Massachusetts Institute of Technology
Robbee Wedow, Purdue University

Intentional and unintentional misinterpretations and misuses of genomic research findings that disparage groups of people, often thought of as part of the field's troubled and distant past, have resurfaced. The most extreme examples are typified by people reimagining current genomic findings as justification for beliefs and actions of white supremacists. Subtler but perhaps more pernicious are suggestions that leaders in areas like education, immigration and judicial sentencing factor genomic backgrounds into policies and decision-making.

The American Society of Human Genetics has recently noted the harm that may accrue if the field remains silent on “modern attempts to use human genetics to advance racism, assert other forms of ‘superiority,’ or perpetuate discrimination”. This panel seeks to raise awareness of the issue and present attendees with ideas of how to communicate with colleagues, the media and others when discussing these misuses of genomics. Researchers across the field can bring their understanding of the nuances, strengths and limitations of genomic research findings to these discussions.

Panelists include a genomic scientist versed in these conversations, two ELSI scholars who work in concert with genomic researchers, and an expert in scientific communication. Together they will explain some of the conceptual roots of these misuses, present empirical data on the existence of the problem today, discuss how issues may persist even after findings have been published and publicized, and provide researchers with tools to talk about discriminatory uses of genomics, regardless of whether your research is the subject of misinterpretation.

THURSDAY, May 11—7:30 PM

POSTER SESSION II

See [p. xxxi for List of Posters](#)

FRIDAY, May 12—8:30 AM

SESSION 6 SYSTEMS GENETICS

Chairperson: **Ben Lehner**, Wellcome Sanger Institute, Cambridge, United Kingdom

Mutate everything—Globally mapping allosteric communication in proteins

Ben Lehner.

Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom; CRG, Barcelona, Spain.

36

Understanding complex genotype-phenotype maps

Carlos Martí-Gómez, David McCandlish.

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

37

Profiling the neurobiology underlying brain structure in living human subjects

Anina N. Lund, Noam D. Beckmann, Alexander W. Charney.

Presenter affiliation: Icahn School of Medicine at Mount Sinai, New York, New York.

38

Precise modulation of transcription factor levels reveals non-linear dosage responses within transcriptional networks

Julia Domingo, Mariia Minaeva, Marcello Ziosi, John A. Morris, Tuuli Lappalainen.

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

39

FRIDAY, May 12—10:15 AM

SESSION 7 EVOLUTIONARY AND NON-HUMAN GENETICS

Chairperson: **Doris Bachtrog**, University of California, Berkeley

Dynamic gene content evolution on *Drosophila* Y chromosomes

Matthew J. Nalley, Doris Bachtrog.

Presenter affiliation: University of California-Berkeley, Berkeley, California.

40

Genetic causes and phenotypic consequences of newly evolved adrenal cell type

Natalie Niepoth, Jennifer Merritt, Michelle Uminski, Sarah Wacker, Stefano Lutzu, Stephanie Rudolph, Andres Bendesky.

Presenter affiliation: Columbia University, New York, New York.

41

Forging new regulatory elements during the 500 million years of evolution leading to humans

Riley J. Mangan, Christiana Fauci, Yanting Luo, Craig B. Lowe.

Presenter affiliation: Duke University School of Medicine, Durham, North Carolina.

42

Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements

Ardian Ferraj, Peter A. Audano, Parithi Balachandran, Anne Czechanski, Jacob I. Flores, Alexander A. Radecki, Varun Mosur, David S. Gordon, Isha A. Walawalkar, Evan E. Eichler, Christine R. Beck.

Presenter affiliation: University of Connecticut Health Center, Farmington, Connecticut; The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut.

43

Investigating the role of insertions in the human genome

Yanting Lu, Craig Lowe.

Presenter affiliation: Duke University Medical Center, Durham, North Carolina.

44

Telomere-to-telomere ape sex chromosome assemblies unravel rapid evolution on the Y and conservative evolution on the X

Kateryna D. Makova, Brandon D. Pickett, Sergey Nurk, DongAhn Yoo, Hyeonsoo Jeong, Barbara McGrath, Robert S. Harris, Monika Cechova, Gabrielle A. Hartley, Jessica M. Storer, Patrick Grady, Tamara Potapova, Matthew Borchers, Sergey Koren, Jennifer L. Gerton, Rachel O'Neill, Evan E. Eichler, Adam M. Phillippy.

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

45

FRIDAY, May 12—2:00 PM

POSTER SESSION III

See p. xliv for List of Posters

FRIDAY, May 12—4:30 PM

GUEST SPEAKERS

Evan Eichler

University of Washington / HHMI

Erich Jarvis

Rockefeller University / HHMI

FRIDAY, May 12—6:30 PM

COCKTAILS and BANQUET

SATURDAY, May 13—9:00 AM

SESSION 8 CANCER AND MEDICAL GENOMICS

Chairpersons: **Ekta Khurana**, Weill Cornell Medicine, New York, New York
Peter van Loo, The University of Texas MD Anderson Cancer Center, Houston

Identifying oncogenic cis-regulatory elements by integration of patient epigenetic data and high-throughput CRISPR screens
Ekta Khurana.

Presenter affiliation: Weill Cornell Medicine, New York, New York.

The mutational landscape of normal gastric epithelium

Tim Coorens, Grace Collard, Hyunchul Jung, Yichen Wang, Suet Yi Leung, Michael Stratton.

Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom; Broad Institute of MIT and Harvard, Cambridge, Massachusetts.

46

Germline-mediated immunoediting sculpts breast cancer subtypes

Kathleen E. Houlahan, Aziz Khan, Noah F. Greenwald, Michael Angelo, Christina Curtis.

Presenter affiliation: Stanford University School of Medicine, Stanford, California.

47

Oncogenes outside chromosomes

King L. Hung, Howard Y. Chang.

Presenter affiliation: Stanford University School of Medicine, Stanford, California.

48

Molecular archeology of cancer

Peter Van Loo.

Presenter affiliation: The University of Texas MD Anderson Cancer Center, Houston, Texas; The Francis Crick Institute, London, United Kingdom.

49

Human genetics of endocrine-related brain anatomy using phenotypes from large-scale biomedical imaging	
Hannah Currant, Christoph Arthofer, Stephen Smith, Teresa Ferreira, Christoffer Nellaker, Gwenaëlle Douaud, Andreas Bartsch, Jesper Andersson, Margaret F. Lippincott, Yee-Ming Chan, Stephanie B. Seminara, Thomas E. Nichols, Søren Brunak, Frederik J. Lange, Cecilia M. Lindgren.	
Presenter affiliation: University of Copenhagen, Copenhagen, Denmark.	50
Evolutionary trajectories of complex genome rearrangements in cancer	
Jose Espejo Valle-Inclan, Solange de Noon, Katherine Trevers, Hillary Elrick, Adrienne M. Flanagan, Isidro Cortes-Ciriano.	
Presenter affiliation: European Molecular Biology Laboratory, Hinxton, Cambridge, United Kingdom.	51
Intercellular extrachromosomal DNA copy number heterogeneity drives cancer cell state diversity	
Maja C. Stöber, Rocío Chamorro González, Lotte Brückner, Thomas Conrad, Nadine Wittstruck, Annabell Szymansky, Angelika Eggert, Johannes H. Schulte, Richard P. Koche, Anton G. Henssen, Roland F. Schwarz, <u>Kerstin Haase</u> .	
Presenter affiliation: Charité-Universitätsmedizin Berlin, Germany German Cancer Research Center (DKFZ), Heidelberg, Germany.	52
 <u>POSTER SESSION I</u>	
Personalized transcription factor binding from deep learning algorithms offers a new framework to identify disease associations	
<u>Temidayo Adeluwa</u> , Saideep Gona, Boxiang Liu, Ravi Madduri, Tiffany Amariuta, Hae Kyung Im.	
Presenter affiliation: The University of Chicago, Chicago, Illinois.	53
How to estimate cloud costs for genomics analyses	
Enis Afgan, Keith Suderman, Nuwan Goonasekera, Michele Savage, Michael Schatz.	
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.	54

Role of chromatin architecture of human cells in the response to infection by Sars-CoV-2

Saumya Agrawal, Kosuke Miyauchi, Kokoro Ozaki, Saera Fujiki, Prashanti Jeyamohan, Hidehiro Fukuyama, Fumihiro Ishikawa, Masato Kubo, Michiel de Hoon.

Presenter affiliation: RIKEN, Yokohama, Japan.

55

Enrichment of Native American and African haplotypes following the Columbian interchange

Richard Ågren, Hugo Zeberg.

Presenter affiliation: Karolinska Institutet, Stockholm, Sweden.

56

Tigerfish and FISHtank—An open-source toolkit for oligonucleotide probe design against repetitive DNA in emerging genomes

Robin Aguilar, Conor K. Campisso, Chelsey Lin, Karen H. Miga, William S. Noble, Brian J. Beliveau.

Presenter affiliation: University of Washington, Seattle, Washington.

57

Modeling the structure and function of gene regulatory networks—From graph properties to expression variation

Matthew Aguirre, Guy Sella, Jonathan K. Pritchard.

Presenter affiliation: Stanford University, Stanford, California.

58

Role of alternative polyadenylation in driving noradrenergic-to-mesenchymal transition in neuroblastoma

Rhea Ahluwalia, Quang Trinh, Fupan Yao, Gabrielle Persad, Brent Derry, Lincoln Stein.

Presenter affiliation: University of Toronto, Toronto, Canada; The Hospital for Sick Children, Toronto, Canada; Ontario Institute for Cancer Research, Toronto, Canada.

59

Compressed linear pangenome indexes for taxonomic classification and genotyping

Omar Ahmed, Naga Sai Kavya Vaddadi, Taher Mun, Ben Langmead.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

60

Intratumoral heterogeneity and clonal evolution induced by HPV integration

Keiko Akagi, David E. Symer, Medhat Mahmoud, Bo Jiang, Sara Goodwin, Darawalee Wangsa, Zhengke Li, Weihong Xiao, Joe D. Dunn, Thomas Ried, Kevin R. Coombes, Fritz J. Sedlazeck, Maura L. Gillison.

Presenter affiliation: MD Anderson Cancer Center, Houston, Texas.

61

The genetic regulation of proteins and post-translational modifications across tissues and cancer

Yo Akiyama, Yifat Geffen, Shankara Anand, Özgün Babur, Meric Kinali, Kisan Thapa, Clinical Proteomic Tumor Analysis Consortium (CP), François Aguet, Gad Getz.

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

62

Allele specific expression in brain of genes involved in psychiatric disorder heritability and cortical thickness

Nirmala Akula, Siyuan Liu, Shrey Shah, Teja N. Peddada, Heejong Sung, Armin Raznahan, Francis J. McMahon.

Presenter affiliation: Human Genetics Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland.

63

Enabling variant annotation and displays across multiple gene annotated human assemblies in Ensembl

Jamie Allen, Olanrewaju Austine-Orimoloye, Gurpreet Ghattaoraya, S. Nakib Hossain, Diana Lemos, Diego Marques-Coelho, Anne Parker, Nuno Saraiva-Agostinho, Likhitha Surapaneni, Thomas Walsh, Leanne Haggerty, Stephen J. Trevanion, David Thybert, Sarah E. Hunt, Andrew Yates, Fiona Cunningham, Fergal J. Martin.

Presenter affiliation: European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom.

64

Signals of epistatic interactions in time series genomic data

Nathan W. Anderson, Carol E. Lee, Aaron P. Ragsdale.

Presenter affiliation: UW - Madison, Madison, Wisconsin.

65

Positive selection in the genomes of two Papua New Guinean populations at distinct altitude levels

Mathilde André, Nicolas Brucato, Georgi Hudjasov, Vasili Pankratov, Danat Yermakovich, Rita Kreevan, Jason Kariwiga, John Muke, Anne Boland, Jean-François Deleuze, Nicholas Evans, Murray P. Cox, Matthew Leavesley, Michael Dannemann, Tõnis Org, Mait Metspalu, Mayukh Mondal, François-Xavier Ricaut.

Presenter affiliation: University of Tartu, Tartu, Estonia.

66

Mapping variants from multiplex assays of variant effect (MAVEs) to human reference sequences

Jeremy A. Arbesfeld, Kori Kuzma, Kevin Riehle, Julia Foreman, Sumaiya Iqbal, Melissa Cline, Alan F. Rubin, Alex H. Wagner.

Presenter affiliation: The Ohio State University, Columbus, Ohio; Nationwide Children's Hospital, Columbus, Ohio.

67

The distribution of distances between heterozygous sites in diploid species allows to efficiently infer demographic history	
<u>Peter F. Arndt, Florian Massip, Michael Sheinman.</u>	
Presenter affiliation: Max Planck Institute for Molecular Genetics, Berlin, Germany.	68
Long-read structural variant breakpoints are altered by small polymorphisms	
<u>Peter A. Audano, Christine R. Beck.</u>	
Presenter affiliation: The Jackson Laboratory, Farmington, Connecticut.	69
Copy-number variants as modulators of common disease susceptibility	
<u>Chiara Auwerx, Maarja Jõeloo, Nicolò Tesio, Alexandre Reymond, Zoltán Kutalik.</u>	
Presenter affiliation: University of Lausanne, Lausanne, Switzerland; University of Lausanne, Lausanne, Switzerland.	70
Quantitative trait gene discovery by genome-wide reciprocal hemizygote scanning	
<u>Randi R. Avery, Sheila Lutz, Frank W. Albert.</u>	
Presenter affiliation: University of Minnesota, Minneapolis, Minnesota.	71
Automated cancer cell line identification from RNA-seq data	
<u>Milad Alasady, Elizabeth T. Bartom.</u>	
Presenter affiliation: Northwestern University, Chicago, Illinois.	72
Genome organization and noncoding RNAs synergistically control the timing of <i>Hox</i> gene transcription during development	
<u>Philippe J. Batut, Michael S. Levine.</u>	
Presenter affiliation: Princeton University, Princeton, New Jersey.	73
Genetic and environmental contributions to ancestry differences in gene expression in the human brain	
<u>Kynon J M. Benjamin, Qiang Chen, Nicholas J. Eagles, Louise A. Huuki-Myers, Leonardo Collado-Torres, Joshua M. Stoltz, Joo Heon Shin, Apuā C M. Paquola, Thomas M. Hyde, Joel E. Kleinman, Andrew E. Jaffe, Shizhong Han, Daniel R. Weinberger.</u>	
Presenter affiliation: Lieber Institute for Brain Development, Baltimore, Maryland; Johns Hopkins University School of Medicine, Baltimore, Maryland.	74

Utilizing ultra-deep WGS to investigate somatic mosaicism in healthy tissue of a BRCA variant carrier

Gage Black, Andrew Farrell, Xiaomeng Huang, Gabor Marth.

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

75

m6A patterns are consistent across different *Drosophila* datasets

George Boateng-Sarfo, Sarah Signor, Lijuan Kan, Eric Lai.

Presenter affiliation: North Dakota State University, Fargo, North Dakota.

76

Targeted deep coverage epigenetic profiles with single-molecule and single-nucleotide precision

Stephanie C. Bohaczuk, Morgan O. Hamm, Chang Li, Mitchell R.

Vollger, Anupama Jha, Benjamin J. Mallory, Jane Ranchalis, Katherine M. Munson, Andre Lieber, Andrew B. Stergachis.

Presenter affiliation: University of Washington School of Medicine, Seattle, Washington.

77

Beyond star alleles—Genetic variant burden scoring framework for pharmacogenomics

Małgorzata Borczyk, Jacek Hajto, Marcin Piechota, Michał Korostynski.

Presenter affiliation: Maj Institute of Pharmacology Polish Academy of Sciences, Krakow, Poland.

78

Isoform Inspector—A JBrowse 2 plugin for visualization and analysis of RNA-splicing patterns

Caroline Bridge, Scott Cain, Colin Diesh, Robert Buels, Garrett Stevens, Lincoln Stein, Ian Holmes.

Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada.

79

eQTL analysis of canine testes suggests novel gene associations with morphological trait loci

Reuben M. Buckley, Alex C. Harris, Susan E. Lana, Elaine A. Ostrander.

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

80

Chromosome substitution for characterizing epistasis in the budding yeast *Saccharomyces cerevisiae*

Cassandra Buzby, Federica Sartori, Mark Siegal.

Presenter affiliation: New York University, New York, New York.

81

An intron motif-aware pipeline for the assembly of spliced transcripts in species of the non-model organism <i>Trichomonas</i>	
<u>Francisco Callejas-Hernández</u> , Mari Shiratori, Krithika Shankar, Frances Blow, Jane M. Carlton.	
Presenter affiliation: Center for Genomics and Systems Biology, New York University, New York, New York.	82
Towards a complete characterization of human polymorphic inversions and their functional effects	
<u>Elena Campoy</u> , Jon Lerga-Jaso, Marta Puig, Ruth Gómez Graciani, Illya Yakymenko, Teresa Soos, Alba Vilella-Figuerola, Ricardo Moreira, Alejandra Delprat, Marina Laplana, Mario Cáceres.	
Presenter affiliation: Institut de Biotecnologia i de Biomedicina, Bellaterra, Spain.	83
Transcriptome-wide co-expression of small non-coding RNAs and genes in cancer	
<u>Taylor B. Cavazos</u> , Aiden M. Sababi, Jeffrey Wang, Alexander J. Lazar, Patrick A. Goodarzi, Hani Goodarzi, Fereydoun Hormozdiari, Babak Alipanahi.	
Presenter affiliation: Exai Bio, Palo Alto, California.	84
The landscape of regional missense intolerance quantified from 125,748 exomes	
<u>Katherine Chao</u> , Lily Wang, Konrad Karczewski, Mark Daly, Kaitlin Samocha.	
Presenter affiliation: Broad Institute, Cambridge, Massachusetts.	85
Uncovering single-cell spatial relationships with highly multiplexed imaging	
<u>Erin Chung</u> , Harald Voehringer, Anastasiia Horlova.	
Presenter affiliation: European Molecular Biology Laboratory, Heidelberg, Germany.	86
Differential cell-type-specific gene expression by type 2 diabetes status in human skeletal muscle	
<u>Dan L. Ciotlos</u> , Sarah C. Hanks, Arushi Varshney, Michael R. Erdos, Nandini Manickam, Anne U. Jackson, Heather M. Stringham, Narisu Narisu, Lori Bonnycastle, Markku Laakso, Jaakko Tuomilehto, Timo A. Lakka, Karen L. Mohlke, Michael Boehnke, Heikki A. Koistinen, Francis S. Collins, Stephen C J. Parker, Laura J. Scott.	
Presenter affiliation: University of Michigan, Ann Arbor, Michigan.	87

Predicting and spatially localising bulk RNA-seq from histology across 39 healthy human tissues	88
<u>Francesco Cisternino</u> , Soumick Chatterjee, Adam P. Levine, Craig A. Glastonbury. Presenter affiliation: Human Technopole, Milan, Italy.	
ProCapNet—Dissecting the cis-regulatory syntax of transcription initiation with deep learning	89
<u>Kelly Cochran</u> , Melody Yin, Anshul Kundaje. Presenter affiliation: Stanford University, Stanford, California.	
The battle of the sexes in humans is highly polygenic	90
<u>Jared M. Cole</u> , Peter R. Golightly, Carly B. Scott, Mackenzie M. Johnson, Jediah Carlson, Matthew J. Ming, Arbel Harpak, Mark Kirkpatrick. Presenter affiliation: University of Texas at Austin, Austin, Texas.	
Decoding the intercellular signals underlying human microglia and astrocyte plasticity	91
<u>Natacha Comandante-Lou</u> , Masashi Fujita, Gilad S. Green, David A. Bennett, Naomi Habib, Vilas Menon, Philip L. De Jager. Presenter affiliation: Columbia University Medical Center, New York, New York.	
The fat-tailed dunnart genome reveals cis-regulatory drivers of developmental heterochrony between marsupials and mammals	92
<u>Laura E. Cook</u> , Charles Y. Feigin, Davide M. Vespaiani, Andrew J. Pask, Irene Gallego Romero. Presenter affiliation: Lawrence Berkeley National Laboratory, Berkeley, California; University of Melbourne, Melbourne, Australia.	
A phenotypic patient matching algorithm to improve diagnoses in rare disease cohorts	93
<u>Isabelle B. Cooperstein</u> , Alistair Ward, Gabor Marth. Presenter affiliation: University of Utah School of Medicine, Salt Lake City, Utah.	
Direct conversion of pediatric ALL to AML after CAR-T cell and blinatumomab therapy	94
<u>Tim Coorens</u> , Grace Collard, Taryn Treger, Stuart Adams, Emily Mitchell, Barbara Newman, Gad Getz, Anna Godfrey, Jack Bartram, Sam Behjati. Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts.	

Deep whole-genome sequencing of GTEx tissues reveals developmental patterns and somatic evolution

Tim Coorens, Danielle Firer, Oliver Priebe, Julian Hess, Gad Getz, Francois Aguet, Kristin Ardlie.

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

95

Genetic architecture of immune cell DNA methylation in free-ranging rhesus macaques

Christina E. Costa, Marina M. Watowich, Rachel M. Peterson, Elisabeth A. Goldman, Kirstin Sterner, Michael J. Montague, Michael Platt, Josue E. Negron-Del Valle, Daniel Phillips, Lauren J. Brent, James P. Higham, Noah Snyder-Mackler, Amanda J. Lea.

Presenter affiliation: New York University, New York, New York.

96

Detection of genetic and epigenetic alterations driven by loss of TET proteins at single base resolution

Hugo Sepulveda, Robert Crawford, Fabio Puddu, Yang Liu, Ankita Singhal, Gary Yalloway, Helen Sansom, Jens Fullgrabe, Nikolay Pchelintsev, Lidia Prieto-Lafuente, Audrey Vandomme, Philippa Burns, David M. Morley, Rosie Spencer, Páidí Creed, Joanna D. Holbrook, Anjana Rao.

Presenter affiliation: La Jolla Institute for Immunology, San Diego, California.

97

Simultaneous measurement of genetics and epigenetics enables new biological insight

Nicholas J. Harding, Páidí Creed, David Currie, Casper K. Lumby, David M. Morley, Fabio Puddu, Jean Teyssandier, Michael Wilson, Jens Fullgrabe, Audrey Vandomme, Aurel Negrea, Alexandra Palmer, Philippa Burns, Shirong Yu, Diljeet Gill, Aled Parry, Wolf Reik, Joanna D. Holbrook.

Presenter affiliation: Cambridge Epigenetix Ltd, Cambridge, United Kingdom.

98

Detecting somatic LINE-1 retrotransposon insertions in single neurons

Michael S. Cuoco, Meiyang Wang, Rohini Gadde, Iryna Gallina, Reicardo Jacomini, Daniel R. Weinberger, Jennifer A. Erwin, Eran A. Mukamel, Apua Paquola, Fred H. Gage.

Presenter affiliation: Salk Institute for Biological Studies, La Jolla, California; University of California, San Diego, La Jolla, California.

99

Retrocopies—GENE copies identified in vertebrates and invertebrates' genomes and their orthology

Helena B. da Conceição, Rafael L. Mercuri, Matheus P. Castro, Daniel T. Ohara, Gabriela Guardia, Pedro F. Galante.

Presenter affiliation: Hospital Sírio Libanês, São Paulo, Brazil;
University of São Paulo, São Paulo, Brazil.

100

Exploring the regulatory landscape of cancer adaption in acidic extracellular matrix

Yifan Dai, Arnaud Stigliani, Jiayi Yao, Renata Lalchina, Dominika Czaplinska, Stine F. Pedersen, Albin G. Sandelin.

Presenter affiliation: University of Copenhagen, Copenhagen,
Denmark.

101

Diversity and representation of South Asian genomes

Arun Das, Michael C. Schatz.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

102

Multi-modal assessment of functional impact of mutations on the genome

Maitreya Das, Deepro Banerjee, Jiawan Sun, Saie Mogre, Ayaan Hossain, Adam Glick, Santhosh Girirajan.

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

103

Discovery of type 2 diabetes genes using an accessible tissue

David Davtian, Theo Dupuis, Dina Mansour-Aly, Naeimeh Atabaki-Pasdar, Mark Walker, Paul W. Franks, Femke Rutters, Hae Kyung Im, Ewan R. Pearson, Martijn Van de Bunt, Ana Viñuela, Andrew A. Brown.

Presenter affiliation: University of Dundee, Dundee, United Kingdom.

104

Biochemical activity is the default DNA state in eukaryotes

Ishika Luthra, Xinyi E. Chen, Cassandra Jensen, Abdul Muntakim Rafi, Asfar Lathif Salaudeen, Carl G. de Boer.

Presenter affiliation: University of British Columbia, Vancouver,
Canada.

105

Unidirectional expression of enhancers with cell type-dependent direction of transcription

Emi Kanamaru, Yoriko Saito, Fumihiko Ishikawa, Michiel J. de Hoon.

Presenter affiliation: RIKEN, Yokohama, Japan.

106

Understanding the context-dependent role of PBAF-specific subunit PBRM1 in chromatin regulation and cancer

Alisha Dhiman, Emily Dykhuizen.

Presenter affiliation: Purdue University, West Lafayette, Indiana.

107

Rhinovirus infected epithelial cells drive genetic susceptibility to childhood-onset asthma

Sarah Djeddi, Daniela Fernandez-Salinas, George Huang, Chitrasen Mohanty, Christina Kendzierski, Joshua Boyce, James Gern, Nora Barrett, Maria Gutierrez-Arcelus.

Presenter affiliation: Boston Children's Hospital, Boston, Massachusetts; Broad Institute of MIT and Harvard, Boston, Massachusetts.

108

The genetic architecture of adaptive pigmentation traits in swordtail (*Xiphophorus*) fishes

Tristram O. Dodge, Daniel L. Powell, John J. Baczenas, Theresa R. Gunn, Shreya M. Banerjee, Manfred Schartl, Molly Schumer.

Presenter affiliation: Stanford University, Stanford, California.

109

Leveraging polygenic enrichments for risk gene prioritisation from GWAS summary statistics

Theo Dupuis, Will Macnair, Andrew A. Brown, Martin Ebeling, Julien Bryois.

Presenter affiliation: Roche Pharma Research and Early Development, Basel, Switzerland; University of Dundee, School of Medicine, Dundee, United Kingdom.

110

StratoMod—Predicting sequencing and variant calling errors with interpretable machine learning

Nathan Dwarshuis, Peter Tonner, Nathan Olson, Fritz Sedlazeck, Justin Wagner, Justin Zook.

Presenter affiliation: National Institute of Standards and Technology, Gaithersburg, Maryland.

111

Promoter sequence and architecture determine expression variability and confer robustness to genetic variants

Hjörleifur Einarsson, Marco Salvatore, Christian Vaagensen, Nicolas Alcaraz, Sarah Rennie, Jette Bornholdt, Robin Andersson.

Presenter affiliation: University of Copenhagen, Copenhagen, Denmark.

112

Discovering macroevolutionary trends for human cell types

Christiana Fauci, Craig B. Lowe.

Presenter affiliation: Duke University Medical Center, Durham, North Carolina.

113

Chromosome-scale and haplotype-resolved sequence assembly of *Hordeum bulbosum* genomes

Jia-Wu Feng, Maria Cuacos, Hélène Pidon, Thomas Lux, Heidrun Gundlach, Yi-Tzu Kuo, Jörg Fuchs, Axel Himmelbach, Manuel Spannagl, Jochen Kumlehn, Stefan Heckmann, Andreas Houben, Frank Blattner, Nils Stein, Martin Mascher.

Presenter affiliation: Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany.

114

Widespread transposable element dysregulation in human aging brains

Yayan Feng, Feixiong Cheng.

Presenter affiliation: Cleveland Clinic, Cleveland, Ohio.

115

Segmental duplication-mediated variation across diverse mouse genomes

Eden Francoeur, Ardian Ferraj, Peter A. Audano, Parithi Balachandran, Christine R. Beck.

Presenter affiliation: The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut; University of Connecticut Health Center, Farmington, Connecticut.

116

Proteome-scale probabilistic modeling of human genetic variation

Jonathan Frazer.

Presenter affiliation: The Centre for Genomic Regulation, Barcelona, Spain.

117

High-throughput multi-omic profiling of oncofusion proteins with single-cell genomics

Max Frenkel, Zachary Morris, Vatsan Raman.

Presenter affiliation: University of Wisconsin, Madison, Wisconsin.

118

The sex-specific genetic architecture of childhood asthma

Amelie Fritz, Anders Ulrik Eliasen, Kasper Rasmussen, Casper Emil Tingskov Pedersen, Klaus Bønnelykke, Anders Gorm Pedersen.

Presenter affiliation: Technical University of Denmark, Kgs. Lyngby, Denmark; Copenhagen University Hospital, Herlev-Gentofte, Gentofte, Denmark.

119

Species-aware DNA language modelDennis Gankin, Alexander Karollus, Julien Gagneur.Presenter affiliation: Technical University of Munich, Munich,
Germany.

120

**Investigating RNA-based gene duplication, a major force
generating genetic novelties in human and other genomes**Pedro A. Galante, Helena B. Conceição, Gabriela D. Guardia, Rafael
L. Mercuri.

Presenter affiliation: Hospital Sírio-Libanês, São Paulo, Brazil.

121

**Machine learning based patient stratification to enhance PheWAS
analyses using UK Biobank data**Manik Garg, Marcin Karpinski, Ryan S. Dhindsa, Dorota Matelska,
Amanda O'Neill, Quanli Wang, Andrew Harper, Slavé Petrovski,
Dimitrios Vitsios.Presenter affiliation: Centre for Genomics Research, Cambridge,
United Kingdom.

122

**Recombination between heterologous human acrocentric
chromosomes**Andrea Guerracino, Silvia Buonaiuto, Leonardo Gomes de Lima,
Tamara Potapova, Arang Rhee, Sergey Koren, Boris Rubinstein,
Christian Fischer, Jennifer L. Gerton, Adam M. Phillippy, Vincenza
Colonna, Erik Garrison.Presenter affiliation: University of Tennessee Health Science Center,
Genetics, Genomics and Informatics, Tennessee.

123

**Detecting decoherent gene co-expression patterns associated
with a rural-to-urban lifestyle transition in Turkana**Kristina M. Garske, Diogo Melo, Marina M. Watowich, Varada
Abhyankar, Echwa John, Michael Gurven, John Kahumbu, Joseph
Kamau, Patricia Kinyua, Dino J. Martins, Charles Miano, Benjamin
Muhoya, Julie Peng, Amanda J. Lea, Julien F. Ayroles.Presenter affiliation: Princeton University, Princeton, New Jersey;
Mpala Research Centre, Nanyuki, Kenya.

124

**Using long-read sequencing to identify methylation differences
related to Alzheimer's disease**Rylee M. Genner, Melissa M. Meredith, Kimberley J. Billingsley, Pilar
Alvarez Jerez, Laksh Malik, Winston Timp, Miten Jain, Cornelis
Blauwendaart.Presenter affiliation: Center for Alzheimer's and Related Dementias,
Bethesda, Maryland; Johns Hopkins University, Baltimore, Maryland.

125

The topography of nullomer-resurfacing mutations and their relevance to human disease

Candace Chan, Ioannis Mouratidis, Georgios G. Tsatsianis, Sarah Fong, Martin Hemberg, Nadav Ahituv, Ilias Georgakopoulos-Soares.
Presenter affiliation: The Pennsylvania State University, Hershey, Pennsylvania.

126

Effects of parental age and polymer composition on short tandem repeat *de novo* mutation rates

Michael E. Goldberg, Michelle D. Noyes, Evan E. Eichler, Aaron R. Quinlan, Kelley Harris.

Presenter affiliation: University of Utah, Salt Lake City, Utah; University of Washington, Seattle, Washington.

127

The role of recombination in the origin and evolution of human inversions.

Ruth Gómez-Graciani, Antonio Barbadilla, Mario Cáceres.

Presenter affiliation: Institut de Biotecnologia i de Biomedicina, Bellaterra (Barcelona), Spain.

128

Identification of haplotype epistasis in gene expression regulation using deep learning models

Saideep Gona, Temidayo Adeluwa, Andy Dahl, Boxiang Liu, Ravi Madduri, Hae Kyung Im.

Presenter affiliation: University of Chicago, Chicago, Illinois.

129

Terminator: Rise of the Plant Machine—How large-scale characterization of plant terminators reveals species-specific expression levels

Sayeh Gorjifard, Tobias Jores, Jackson Tonnes, Nicholas A. Meuth, Kerry Bubb, Travis Wrightsman, Joshua T. Cuperus, Edward S. Buckler, Stanley Fields, Christine Queitsch.

Presenter affiliation: University of Washington, Seattle, Washington.

130

Integration of 170,000 samples reveals global patterns of gut microbiome diversity

Samantha P. Graham, Richard J. Abdill, Frank W. Albert, Ran Blekhman.

Presenter affiliation: University of Minnesota, Minneapolis, Minnesota.

131

Regulatory enhancer-gene interactions in the human genome

Andreas R. Gschwind, Kristy Mualim, Alireza Karbalayghareh, Maya U. Sheth, Kushal K. Dey, Ramil N. Nurtdinov, Evelyn Jagoda, Wang Xi, Alkes L. Price, Michael Beer, Roderic Guigo, Lars M. Steinmetz, Christina Leslie, John Stamatoyannopoulos, Erez Aiden, William J. Greenleaf, Anshul Kundaje, Jesse M. Engreitz.

Presenter affiliation: Stanford University, Stanford, California.

132

The rapid evolution of TBC1D3-a gene implicated in human cortical expansion.

Xavi Guitart, Philip Dishuck, David Porubsky, PingHsun Hsieh, Evan Eichler.

Presenter affiliation: University of Washington, Seattle, Washington.

133

POSTER SESSION II**single cell multiome analysis reveals cancer cell plasticity**

Yasuhiko Haga, Masahide Seki, Ayako Suzuki, Yutaka Suzuki.

Presenter affiliation: The University of Tokyo, Kashiwa, Japan.

134

Transcriptome analysis of familial dysautonomia reveals tissue-specific gene expression disruption in the peripheral nervous system

Ricardo S. Harrapaul, Elisabetta Morini, Monica Salani, Emily Logan, Emily G. Kirchner, Jessica Bolduc, Anil Chekuri, Benjamin Currall, Rachita Yadav, Serkan Erdin, Michael E. Talkowski, Dadi Gao, Susan Slaugenhaupt.

Presenter affiliation: Massachusetts General Hospital, Cambridge, Massachusetts.

135

Centuries of genome instability and evolution in soft-shell clam transmissible cancer

Samuel F. Hart, Marisa A. Yonemitsu, Rachael M. Giersch, Brian F. Beal, Gloria Arriagada, Brian W. Davis, Elaine A. Ostrander, Stephen P. Goff, Michael J. Metzger.

Presenter affiliation: Pacific Northwest Research Institute, Seattle, Washington; University of Washington, Seattle, Washington.

136

Whole genome burden testing in 333,100 individuals of rare non-coding genetic variation on complex phenotypes

Gareth Hawkes, Robin N. Beaumont, Zilin Li, Ravi Mandla, Xihao Li, Alisa K. Manning, Xihong Lin, Caroline F. Wright, Andrew R. Wood, Timothy M. Frayling, Michael N. Weedon.

Presenter affiliation: University of Exeter, Exeter, United Kingdom.

137

How does somatic mosaicism vary along the length of the human colon?

Laurel Hiatt, Jason Kunisaki, Suchita Lulla, Xichen Nie, James Hotaling, Aaron Quinlan.

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

138

The impact of genomic variation on function catalog

Benjamin C. Hitz, Pedro R. de Assis, Idan Gabdank, Shencheng Dong, Meenakshi S. Kagda, Mingjie Li, Otto Jolanki, Jennifer Jou, Kalina Andreeva, Khine Lin, Ian Whaling, Wenjin Zhang, Xiaowen Ma, Daofeng Li, Heather Lawson, Feng Yue, Ting Wang, J Michael Cherry.

Presenter affiliation: Stanford School of Medicine, Stanford, California.

139

Transcriptional profiling of 3D lower airway tissue culture model exposed to phylogenetically diverse microbes

Mian Horvath, Elizabeth Fleming, Ruoyu Yang, Diana Cadena Castaneda, Megan Callendar, Jose Fachi, Marco Colonna, Karolina Palucka, Julia Oh.

Presenter affiliation: The Jackson Laboratory, Farmington, Connecticut; UConn Health, Farmington, Connecticut.

140

Human disease genetics and real-world patient data fuel target and drug discovery in European and African American ancestries

Yuan Hou, Pengyue Zhang, Jeffrey Cummings, Andrew A. Pieper, James B. Leverenz, Feixiong Cheng.

Presenter affiliation: Cleveland Clinic, Cleveland, Ohio.

141

Comprehensive map of introgressed structural variation in the human genome

PingHsun Hsieh, William T. Harvey, Katherine M. Munson, Kendra Hoekzema, Francois-Xavier Ricaut, Nicolas Brucato, Irene G. Romero, Murray Cox, Evan E. Eichler.

Presenter affiliation: University of Minnesota, Twin Cities, Minnesota.

142

Cell division history encodes directional information of fate transitions

Kun Wang, Liangzhen Hou, Xionglei He, Christina Curtis, Da Zhou, Zheng Hu.

Presenter affiliation: Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

143

Linking process-specific polygenic risk scores of lipodystrophy to differentiation-dependent adipocyte subpopulations

Yi Huang, Joaquín Pérez-Schindler, Thiago M. Batista, Hesam Dashti, Miriam S. Udler, Melina Claussnitzer.

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

144

Investigating the effects of *Indy* reduction on a fly model of Alzheimer's disease

Billy J. Huggins, Jacob Macro, Kali Meadows, Blanka Rogina.

Presenter affiliation: UConn Health, Farmington, Connecticut.

145

Single-nucleoid architecture reveals heterogeneous packaging of human mitochondrial DNA

R. Stefan Isaac, Thomas W. Tullius, Katja G. Hansen, Danilo Dubocanin, Mary Couvillion, Andrew B. Stergachis, L. Stirling Churchman.

Presenter affiliation: Blavatnik Institute, Harvard Medical School, Boston, Massachusetts.

146

Producing artificial antibodies from FFPE lung tissue

Sadahiro Iwabuchi, Shinichi Hashimoto.

Presenter affiliation: Wakayama Medical University, Wakayama, Japan.

147

Assessing tissue-specific effects of rare and structural variants towards gene regulation with the EN-TEx personal genome resource

Matthew Jensen, Tai Michaels, Anna Su, Timur Galeev, Sushant Kumar, Kun Xiong, Beatrice Borsari, Joel Rozowsky, Mark Gerstein. Presenter affiliation: Yale University, New Haven, Connecticut.

148

Novel epistatic interactions between *KIT* and *MITF* cause breakthrough pigmentation in regions of white on the coat in cattle

Swati Jivanji, Anna Yeates, Chad Harland, Charlotte Gray, Christine Couldrey, Gemma Worth, Isabelle Gamache, John A A. Tabares, Lorna McNaughton, Marie-Pier Cloutier, Jade Desjardins, Mitra Crowan, Tony Fransen, Tracey Monehan, Richard Spelman, Richard Mort, Yojiro Yamanaka, Mathew D. Littlejohn.

Presenter affiliation: Livestock Improvement Corporation, Hamilton, New Zealand.

149

Application of network-based heterogeneity clustering for investigation of genotype-phenotype correlations in BioMe BioBank

Meltem Ece Kars, Yiming Wu, Cigdem Sevim Bayrak, Bruce D. Gelb, Yuval Itan.

Presenter affiliation: Icahn School of Medicine at Mount Sinai, New York, New York.

150

Single cell analysis of immune cell landscape of healthy individuals

Yukie Kashima, Yutaka Suzuki.

Presenter affiliation: The University of Tokyo, Kashiwa, Japan.

151

Robust differential expression testing for single-cell CRISPR screens

Timothy Barry, Kathryn Roeder, Eugene Katsevich.

Presenter affiliation: University of Pennsylvania, Philadelphia, Pennsylvania.

152

Chromatin modifiers as drivers in endometrial cancer

Katarzyna Z. Kedzierska, Yannick Comoglio, Matthew W. Brown, Endometrial Cancer GeCIP Domain, Dan J. Woodcock, David N. Church.

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

153

Common genetic variation related to Schizophrenia is associated with cognitive performance in children and adolescents

Gianluca C. Kikidis, Alessandra Raio, Nora Penzel, Leonardo Sportelli, Linda A. Antonucci, Alessandro Bertolino, Qiang Chen, Pierluigi Selvaggi, Antonio Rampino, Giulio Pergola.

Presenter affiliation: University of Bari Aldo Moro, Bari, Italy; Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, Maryland.

154

Inference of admixture origins in indigenous African cattle

Kwondo Kim, Donghee Kim, Olivier Hanotte, Charles Lee, Heebal Kim, Choongwon Jeong.

Presenter affiliation: The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut; Seoul National University, Seoul, South Korea.

155

Randomizing the human genome by engineering recombination between repeat elements

Jonas Koeppel, Raphael Ferreira, Fabio G. Liberante, Thomas Vanderstichele, Gareth Girling, George Church, Tom Ellis, Leopold Parts.

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

156

Systematic mutagenesis reveals functional complexity of human enhancers *in vivo*

Michael Kosicki, Stella Tran, Jennifer A. Akiyama, Ingrid Plajzer-Frick, Catherine S. Novak, Yiwen Zhu, Momoe Kato, Anne Harrington, Riana D. Hunter, Kianna von Maydell, Janeth Godoy, Eman M. Meky, Sarah Bartoni, Erik Beckman, Diane E. Dickel, Axel Visel, Len A. Pennacchio.

Presenter affiliation: Lawrence Berkeley Lab, Berkeley, California.

157

Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses

Kyungyong Seong, Ksenia V. Krasileva.

Presenter affiliation: University of California, Berkeley, California.

158

(Context-) transcription factors create cooperative regulatory environments and mediate enhancer communication

Judith F. Kribelbauer, Olga Pushkarev, Julie Russeil, Vincent Gardeux, Guido van Mierlo, Bart Deplancke.

Presenter affiliation: EPFL, Lausanne, Switzerland; Swiss Institute of Bioinformatics, Lausanne, Switzerland.

159

Cloud-scale training and education in the NHGRI Analysis, Visualization, and Informatics Lab-space (AnVIL)

Natalie Kucher, Michael C. Schatz, Anthony Philippakis, AnVIL Team.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

160

Identification of conserved sequence elements across 242 primate genomes with deep learning

Sabrina Rashid, Lukas F.K. Kuderna, Jacob Ullirsch, Mo Ameen, Laksshman Sundaram, Glenn Hickey, Anthony J. Cox, Hong Gao, Arvind Kumar, Francois Aguet, Primate Conservation Sequencing Initiative, Benedict Paten, Kerstin Lindblad-Toh, Jeffrey Rogers, Tomas Marques Bonet, Kyle Kai-How Farh.

Presenter affiliation: Illumina, Inc., Foster City, California.

161

Prioritizing coronary artery disease risk variants in atherosclerosis using deep learning models of chromatin accessibility in mouse

Soumya Kundu, Robert Wirka, Daniel Li, Joao Monteiro, Disha Sharma, Laksshman Sundaram, Thomas Quertermous, Anshul Kundaje.

Presenter affiliation: Stanford, Stanford, California.

162

Leveraging duplex DNA sequencing to comprehensively investigate germline mutations in longitudinally sampled bulk sperm

Jason Kunisaki, Suchita Lulla, Michael Goldberg, Kenneth Aston, Jim Hotaling, Aaron Quinlan.

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

163

A programmed Mendelian violation maintains heterozygosity in a parthenogenetic ant

Kip D. Lacy, Daniel J. Kronauer.

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

164

Automated Reference Genome Assembly by Galaxy and the Vertebrate Genome Project

Delphine Lariviere, Giulio Formenti, Alex Ostrovsky, Cristobal Gallardo, Linelle Abueg, Nadolina Brajuka, Marc Palmada-Flores, Anton Nekrutenko, Bjorn Grüning, Michael Schatz.

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

165

Massively parallel profiling of androgen receptor protein-coding variants with SCAnnEd

Ceejay Lee, Tristan Tay, Hui Si Kwok, Simon P. Shen, Calvin Hu, Jason D. Buenrostro, Brian B. Liau.

Presenter affiliation: Harvard University, Cambridge, Massachusetts.

166

Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants

Jessica C. McAfee, Sool Lee, Jiseok Lee Lee, Jessica L. Bell, Oleh Krupa, Jessica Davis, Kimberly Insigne, Marielle L. Bond, Nanxiang Zhao, Hyejung Won.

Presenter affiliation: University of North Carolina, Chapel Hill, North Carolina.

167

The genomic landscape across 624 surgically accessible epileptogenic human brain lesions

Costin Leu, Christian M. Boßelmann, Jean Khoury, Lucas Hoffmann, Sara Baldassari, Robyn M. Busch, Stéphanie Baulac, Peter Nürnberg, Imad Najm, Ingmar Blümcke, Dennis Lal.

Presenter affiliation: Cleveland Clinic, Cleveland, Ohio.

168

Identification of cell types and cellular dynamics genetically associated with brain disorders and cognitive traits

Ang Li, Irina Voineagu, Ryan Lister, Naomi R. Wray, Jian Zeng.

Presenter affiliation: University of Queensland, Brisbane, Australia.

169

Changes in astrocytes transcriptome and proteome with aging in normal and Alzheimer's Disease mice brain

Jiangtao Li, Michelle Olsen.

Presenter affiliation: Virginia Tech, Blacksburg, Virginia.

170

Associating cancer and stromal genomes with transcriptomes by high-throughput single-cell sequencing

Siran Li, Joan Alexander, Jude Kendall, Gary Goldberg, Dan Levy, Michael Wigler.

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

171

Employing linked read sequencing (haplotagging) to profile de novo structural variation in sperm through aging

Stacy Li, Joana Rocha, Peter H. Sudmant.

Presenter affiliation: UC Berkeley, Berkeley, California.

172

Multi-omic Bayesian hierarchical modeling reveals trait-relevant rare genetic variation

Taibo Li, Rebecca Keener, Rachel Ungar, Nicole Ferraro, Matilde Cirnigliaro, Stephanie Arteaga, Bohan Ni, Jerome Rotter, Stephen Rich, Dan Arking, Daniel Geschwind, Stephen Montgomery, Alexis Battle.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

173

A novel pathway analysis method for scRNA-seq and spatial transcriptomics data

Qingnan Liang, Ken Chen.

Presenter affiliation: The University of Texas, MD Anderson Cancer Center, Houston, Texas.

174

Phenome-wide PGS portability in the Colorado Center for Personalized Medicine biobank suggests overlooked challenges in diverse populations

Meng Lin, Christopher H. Arehart, Nicholas Rafaels, Kristy R. Crooks, Nikita Pozdeyev, Audrey Hendricks, Sridharan Raghavan, Christopher R. Gignoux.

Presenter affiliation: University of Colorado Anschutz Medical Campus, Aurora, Colorado.

175

scGraph2Vec—A new method for gene embedding augmented by Graph Neural Network and single-cell omics data

Shiqi Lin, Peilin Jia.

Presenter affiliation: Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China; University of Chinese Academy of Sciences, Beijing, China.

176

Dynamic complexity of genetic regulatory effects in response to a high cholesterol, high fat diet in baboons

Wenhe Lin, Ge Li, John VandeBerg, Deborah Newman, Michael Olivier, Mark Abney, Jeff Wall, Laura A. Cox, Yoav Gilad.

Presenter affiliation: The University of Chicago, Chicago, Illinois.

177

An integrative study to identify the link between dysregulated intercellular signalings and genetic variants in Alzheimer's disease

Andi Liu, Xiaoyang Li, Yulin Dai, Zhongming Zhao.

Presenter affiliation: The University of Texas Health Science Center at Houston, Houston, Texas.

178

Single-cell RNA-seq links cell type-specific regulation of splicing to autoimmune diseases

Chi Tian, Yuntian Zhang, Yihan Tong, Boxiang Liu.

Presenter affiliation: National University of Singapore, Singapore.

179

Mapping and functional characterization of structural variation in 1,060 pig genomes

Liu Yang, Lijing Bai, Hongwei Yin, Kui Li, George E. Liu, Lingzhao Fang.

Presenter affiliation: USDA ARS, Beltsville, Maryland.

180

DNA-sequence and epigenomic determinants of local rates of transcription elongation

Lingjie Liu, Yixin Zhao, Adam Siepel.

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

181

Correcting and classifying SARS-CoV-2 RNA expression in single cells

Wendao Liu, Zhongming Zhao.

Presenter affiliation: The University of Texas MD Anderson Cancer Center UTHHealth Houston, Houston, Texas; The University of Texas Health Science Center at Houston, Houston, Texas.

182

The chromatin regulatory landscape of mouse liver regeneration

Palmira Llorens-Giralt, Marina Ruiz-Romero, Macarena Herranz-Itúrbide, Ramil Nurtdinov, A Silvina Nacht, Guillermo P. Vicent, Florenci Serras, Isabel Fabregat, Montserrat Corominas.

Presenter affiliation: Universitat de Barcelona, Barcelona, Spain.

183

Unraveling the genetic basis of rapid diversification in rockfish

Runyang Nicolas Lou, Laura Timm, Stacy Li, Katie D'Amelio, Kirby Karpan, Nathan Sykes, Gregory Owens, Wesley Larson, Peter Sudmant.

Presenter affiliation: UC Berkeley, Berkeley, California.

184

Characterization of housekeeping regulatory elements in the human genome

Martin Loza, Alexis Vandenbon, Kenta Nakai.

Presenter affiliation: The University of Tokyo, Tokyo, Japan.

185

Characterization of SVs in the human pangenome reference

Shuangjia Lu, Wen-wei Liao, Haley J. Abel, Heng Li, Ira Hall.

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

186

Strategies for identifying high-confidence de novo mutations in somatic and germline cells through duplex sequencing of diverse tissue types

Suchita Lulla, Jason Kunisaki, Laurel Hiatt, Michael Goldberg, Kenneth Aston, Aaron Quinlan.

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

187

Nucleotide determinants of derived enhancer function in hominin evolution

Riley J. Mangan, Craig B. Lowe.

Presenter affiliation: Duke University Medical Center, Durham, North Carolina.

188

Single-cell multi-omics in fetal Down's syndrome reveals the impact of aneuploidy in cellular differentiation and gene regulation

Andrew R. Marderstein, Marco De Zuani, Haoliang Xue, Jon Bezney, Shuo Wong, Stephen B. Montgomery, Ana Cvejic.

Presenter affiliation: Stanford, Stanford, California.

189

Cross-ancestry, cell-type-informed atlas of gene, isoform, and splicing regulation in the developing human brain

Cindy Wen, Michael Margolis, Rujia Dai, Pan Zhang, Pawel Przytycki, Daniel Vo, Bogdan Pasaniuc, Jason Stein, Michael Love, Katherine Pollard, Chunyu Liu, Michael Gandal.

Presenter affiliation: David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California.

190

The high-coverage genome of a male Neandertal

Diyendo Massilani, Stéphane Peyrégne, Cesare De Filippo, Leonardo N. Iasi, Alba Bossoms Mesa, Divyaratna Popli, Arev Pelin Sümer, Christian Heide, Maxim B. Kozlikin, Michael V. Shunkov, Anatoly P. Derevianko, Samantha Brown, Thomas Higham, Katerina Douka, Matthias Meyer, Hugo Zeberg, Janet Kelso, Svante Pääbo.

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

191

LT-Free—A novel method for leveraging family history in genetic association studies of arbitrarily complex diseases.

Jamie Matthews, Mike Thompson, Noah Zaitlen.

Presenter affiliation: UCLA, Los Angeles, California.

192

Analysis of ancient bone DNA samples from excavations at St Peter's burial ground, Blackburn

Shakhawan Mawlood, Catriona Pickard, Benjamin Pickard.

Presenter affiliation: Shakhawan Mawlood, Glasgow, United Kingdom.

193

Social environmental effects on gene regulation and aging in a large cohort of companion dogs

Brianah M. McCoy, Layla Brassington, Beth Slickas, The Dog Aging Project Consortium, Noah Snyder-Mackler.

Presenter affiliation: Arizona State University, Tempe, Arizona.

194

Beyond the exome—A genomics-based undiagnosed genetic disease research program

Stephen Meyn, Bryn D. Webb, Derek Pavelec, Heather Motiff, Jadin Heilmann, Xiang Qiang Shao, Vanessa Horner, April Hall.

Presenter affiliation: University of Wisconsin - Madison, Madison, Wisconsin.

195

Hypoxia leads to unique mtDNA transcriptional patterns and affects mito-nuclear regulatory coordination

Noam Shtolz, Sara Dadon, Dan Mishmar.

Presenter affiliation: Ben-Gurion University of the Negev, Beer-Sheva, Israel.

196

Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data

Ziyi Mo, Adam Siepel.

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

197

A unified computing environment for genomics data storage, management, and analysis—NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL)

Stephen L. Mosher, Michael C. Schatz, Anthony Philippakis, AnVIL Team.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

198

Accurate *de novo* detection of somatic mutations in single-cell genomics and transcriptomics data

Francesc Muyas, Ruoyan Li, Thomas J. Mitchell, Sahand Hormoz, Isidro Cortés-Ciriano.

Presenter affiliation: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom.

199

Intraspecific variation of transposable element dynamics and protein families in a fungal phytopathogen reveal differences in the evolutionary history of its various pathotypes

Anne A. Nakamoto, Pierre M. Joubert, Daniil M. Prigozhin, Ksenia V. Krasileva.

Presenter affiliation: University of California, Berkeley, Berkeley, California.

200

Using snakes rearrangement display to visualize pairwise alignments on the UCSC Genome Browser

Luis R. Nassar, Brian J. Raney, Mark Diekhans, Maximilian Haeussler, William J. Kent, Galt P. Barber, Jonathan Casper, Hiram Clawson, Clay Fischer, Jairo Navarro Gonzalez, Angie S. Hinrichs, Christopher M. Lee, Gerardo Perez.

Presenter affiliation: University of California Santa Cruz, Santa Cruz, California.

201

Modeling the impact of rare structural variants on gene expression in rare disease cases

Bohan Ni, Tanner Jensen, Pagé Goddard, Rachel Ungar, Benjamin Strober, Nicole Ersaro, Taibo Li, Euan A. Ashley, Matthew Wheeler, Stephen B. Montgomery, Michael C. Schatz, Alexis Battle.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 202

BaboonXcan, a framework to identify genes associated with phenotypes in Baboons

Festus Nyasimi, Wenhe Lin, Ellen Quillen, John VandeBerg, Deborah Newman, Michael Olivier, Jeff Wall, Laura A. Cox, Yoav Gilad, Hae Kyung Im.

Presenter affiliation: The University of Chicago, Chicago, Illinois. 203

Identification of host genes associated with COVID-19 risk and severity by ancestry-aware trans-layer multi-omic analysis

Meritxell Oliva, Justyna A. Resztak, Sabah Kadri, Jacob Degner.

Presenter affiliation: AbbVie Inc., Chicago, Illinois. 204

Systematic characterization of regulatory variants of blood pressure genes

Winona Oliveros, Kate Delfosse, Daniella F. Lato, Katerina Kiriaikopoulos, Milad Mokhtaridoost, Abdelrahman Said, Brandon J. McMurray, Jared W. Browning, Kaia Mattioli, Guoliang Meng, James Ellis, Seema Mital, Marta Melé, Philipp G. Maass.

Presenter affiliation: Barcelona Supercomputing Center, Barcelona, Spain. 205

The alpha-2A adrenergic receptor (ADRA2A) modulates susceptibility to dysautonomia and Raynaud's disease

Anniina Tervi, Markus Räsänen, Samuel E. Jones, Caroline Heckman, Erik Abner, Tonu Esko, Jacqueline M. Lane, Matthew Maher, FinnGen FinnGen, Estonian Biobank Research Team, Richa Saxena, Thomas Quertermous, Hanna M. Olliila.

Presenter affiliation: University of Helsinki, Helsinki, Finland; Massachusetts General Hospital, Center for Genomic Medicine, Boston, Massachusetts. 206

Retrospective lineage tracing and phenotypic profiling in human tissues by droplet single cell microsatellite sequencing

Nathaniel D. Omans, Tamara Prieto, John Zinno, Jake Qiu, Shu Wang, Lucy A. Godley, Dan A. Landau.

Presenter affiliation: Weill Cornell Medicine, New York, New York; The New York Genome Center, New York, New York. 207

From "Alaskan Thunderfck" to "Maui Wowie"—The genetic architecture of cannabinoid concentration in 500 strains of pot	208
<u>Sara J. Oppenheim</u> , Armin Scheben, Dean M. Bobo, Robert DeSalle.	
Presenter affiliation: American Museum of Natural History, NYC, New York.	
Metabolic and behavioral effects of a modern human-specific amino acid substitution in adenylosuccinate lyase	
Xiangchun Ju, ShinYu Lee, Chika Azama, Tomomi Miyamoto, Agnieszka Kubik-Zahorodna, Ronald Naumann, Victor Wiebe, Jeanette Frommolt, Rowina Voigtlaender, Michael C. Roy, Wulf Hevers, Izumi Fukunaga, <u>Svante Pääbo</u> .	
Presenter affiliation: Okinawa Institute of Science and Technology, Onna-son, Japan; Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.	209
Utilizing temporal proteomics of iPSC-derived neuronal cell states for study of disease-specific pathways in mental disorders	
<u>Petra Páleníková</u> , Greta Pintacuda, Yu-Han Hsu, Julia Biagini, Daya Mena, Joshua Ching, Travis Botts, Nadine Fornelos, Kasper Lage.	
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts.	210
The impact of repair context on mutations generated by Cas9	
<u>Ananth Pallaseni</u> , Elin Madli Peets, Özdemirhan Serçin, Balca Mardin, Michael Kosicki, Leopold Parts.	
Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom.	211
Clonally selected lines after CRISPR/Cas editing are not isogenic	
<u>Arijit Panda</u> , Milovan Suvakov, Jessica Mariani, Kristen L. Drucker, Yohan Park, Yeongjun Jang, Thomas M. Kollmeyer, Gobinda Sarkar, Taejeong Bae, Jean J. Kim, Wan Hee Yoon, Robert B. Jenkins, Flora M. Vaccarino, Alexej Abyzov.	
Presenter affiliation: Mayo Clinic, Rochester, Minnesota.	212
Systematic identification of silencers in the human genome	
<u>Baoxu Pang</u> , Michael Snyder.	
Presenter affiliation: Leiden University Medical Center, Leiden, Netherlands.	213
Alternative splicing of RPS24 gene is a prognostic biomarker in kidney renal clear cell carcinoma	
<u>Jiyeon Park</u> , Yeun-Jun Chung.	
Presenter affiliation: The Catholic University of Korea, Seoul, South Korea.	214

Expanding cancer therapy options through genome-scale identification of synthetic lethal paralog pairs

Phoebe C. Parrish, Austin M. Gabel, Daniel Grosø, Shriya Kamlapurkar, James D. Thomas, Robert K. Bradley, Alice H. Berger.
Presenter affiliation: Fred Hutchinson Cancer Center, Seattle, Washington; University of Washington, Seattle, Washington.

215

Reconstructing the *cis*-regulatory landscape of archaic hominids using deep learning

Aman Patel, Georgi Marinov, Anshul Kundaje.
Presenter affiliation: Stanford University, Stanford, California.

216

CATE—An accelerated and scalable solution for large-scale genomic data processing through GPU and CPU-based parallelization

Deshan Perera, Elsa Reisenhofer, Said Hussain, Eve Higgins, Christian D. Huber, Quan Long.
Presenter affiliation: University of Calgary, Calgary, Canada.

217

The epigenetic logic of gene activation

Beatrice Borsari, Amaya Abad, Cecilia C. Klein, Ramil Nurtdinov, Vasilis F. Ntasis, Alexandre Esteban, Emilio Palumbo, Marina Ruiz-Romero, Raúl G. Veiga, Maria Sanz, Bruna R. Correa, Rory Johnson, Sílvia Pérez-Lluch, Roderic Guigó.

Presenter affiliation: Centre de Regulació Genòmica, Barcelona, Spain.

218

Slower B cell transdifferentiation in human than in mouse is the by-product of the human specific Alu-repeat expansion

Ramil Nurtdinov, Maria Sanz, Amaya Abad, Carme Arnan, Alexandre Esteban, Sebastian Ullrich, Rory Johnson, Sílvia Pérez-Lluch, Roderic Guigó.

Presenter affiliation: Centre de Regulació Genòmica, Barcelona, Spain.

219

CRISPR-CLEAR—In-situ investigation of genotype-to-phenotype relationship with nucleotide level resolution CRISPR saturation mutagenesis screens

Becerra Basheer, Martin Jankowiak, Sandra Wittibschlager, Anzhelika Karjalainen, Ana Patricia Kutschat, Ting Wu, Marlena Starrs, Zain Patel, Daniel Bauer, Davide Seruggia, Luca Pinello.

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

220

An accessible chromatin atlas of the human intestine identifies disease specific regulatory features and delineates the impact of non-coding variants in Crohn's disease

Yu Zhao, Ran Zhou, Bingqing Xie, Candace M. Cham, Jason Koval, Xin He, Eugene B. Chang, Anindita Basu, Sebastian Pott.

Presenter affiliation: University of Chicago, Chicago, Illinois.

221

It takes two (strands) to make a thing go right. Right?

Aaron Quinlan, Michael Goldberg, Brent Pedersen, Jason Kunisaki, Suchita Lulla, Laurel Hiatt.

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

222

POSTER SESSION III

Plasma proteomic determinants of common causes of mortality

Anurag Sethi, Anil Raj, Kevin Wright, Eugene Melamud.

Presenter affiliation: Calico Life Sciences LLC, South San Francisco, California.

223

Uncovering higher-order motif interactions with biologically interpretable neural networks

Chandana Rajesh, Rohan Ghotra, Steven Yu, Peter Koo.

Presenter affiliation: Simons Center for Quantitative Biology, Cold Spring Harbor, New York.

224

The landscape of transcriptomic and epigenetic variation across human traits

Raquel García-Pérez, Jose Miguel Ramirez, Aida Ripoll-Cladellas, Ruben Chazarra-Gil, Winona Oliveros, Oleksandra Soldatkina, Mattia Bosio, Paul Joris Rognon, Salvador Capella, Miquel Calvo, Ferran Reverter, Roderic Guigó, François Aguet, Pedro G. Ferreira, Kristin G. Ardlie, Marta Melé.

Presenter affiliation: Barcelona Supercomputing Center, Barcelona, Spain.

225

Identification of function variants associated with adolescent idiopathic scoliosis

Darius Ramhalawan, Gloria Montoya-Vazquez, Kayla Ernst, Nadja Makki.

Presenter affiliation: University of Florida, Gainesville, Florida.

226

A test of Mother's Curse with deep mtDNA divergence and outbred nuclear backgrounds in Drosophila

David M. Rand, Faye A. Lemieux, Kenneth Bradley, Lindsay Marmor.

Presenter affiliation: Brown University, Providence, Rhode Island.

227

Single-cell profiling of transcriptional changes associated with neighborhood stress in immune cells of adolescents with asthma

A. Ranjbaran, J. Wei, J. Resztak, S. Nirmalan, A. Alazizi, H. Mair-Meijers, J. Bruinsma, X. Wen, R. Slatcher, S. Zilioli, R. Pique-Regi, F. Luca.

Presenter affiliation: Wayne State University, Detroit, Michigan.

228

Direct long-read RNA sequencing uncovers functional genetic variation affecting transcripts expression.

Aline Réal, Andrew Brown, Gisella Puga Yung, Christelle Borel, Nikolaos Lykoskoufis, Jörg D. Seebach, Emmanouil T. Dermitzakis, Anna Ramisch, Ana Viñuela.

Presenter affiliation: University of Geneva Medical School, Geneva, Switzerland; University Hospitals and Medical Faculty, Geneva, Switzerland; New York Genome Center, New York, New York.

229

The IGVF “8-cubed” single nucleus RNA-seq dataset of transcriptional variation across mouse genotypes

Elisabeth Rebboah, Sina Booreshaghi, Heidi Y. Liang, Delaney Sullivan, Diane Trout, Maria Carilli, Ghassan Filimban, Parvin Mahdipoor, Jasmine Sakr, Fairlie Reese, Brian Williams, Ingileif Hallgrimsdottir, Shimako Kawauchi, Grant McGregor, Kim Green, Lior Patcher, Barbara J. Wold, Ali Mortazavi.

Presenter affiliation: University of California, Irvine, Irvine, California.

230

The ENCODE4 long-read RNA-seq dataset reveals distinct classes of transcript structure diversity

Fairlie Reese, Brian Williams, Elisabeth Rebboah, Narges Rezaie, Diane Trout, Heidi Liang, The ENCODE RNA Working Group, Barbara J. Wold, Ali Mortazavi.

Presenter affiliation: University of California, Irvine, Irvine, California.

231

A robust statistical framework for gene-wise single-cell differential expression meta-analysis in the context of population-based single-cell studies

Aida Ripoll-Cladellas, Monique G.P. van der Wijst, Marc Jan Bonder, Lude Franke, Marta Melé.

Presenter affiliation: Barcelona Supercomputing Center, Barcelona, Spain.

232

Systematic computational search for ORF gene function

Ellen Tsai, Bin Ye, Suganthi Balasubramanian, Alan R. Shuldiner, Juan L. Rodriguez-Flores.

Presenter affiliation: Regeneron Pharmaceuticals Inc, Tarrytown, New York.

233

Bacterial pangenomes shape essentiality and gene-phenotype associations and thereby drive genome-evolution

Federico Rosconi, Tim van Opijken.

Presenter affiliation: Boston College, Chestnut Hill, Massachusetts.

234

A human atlas of imprinting and allele-specific methylation

Jonathan Rosenski, Ayelet Peretz, Judith Magenheimer, Ruth Shemer, Yuval Dor, Benjamin Glaser, Tommy Kaplan.

Presenter affiliation: The Hebrew University of Jerusalem, Jerusalem, Israel.

235

Spatial transcriptomics in the AstroPath platform

Jeffrey S. Roskes, Elizabeth L. Engle, Long Yuan, Atul Deshpande, Joel C. Sunshine, Kellie Smith, Drew M. Pardoll, Janis M. Taube, Alexander S. Szalay.

Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

236

Enhancing gene expression prediction in major psychiatric disorders via co-expression models

Fabiana Rossi, Madhur Parikh, Leonardo Sportelli, Loredana Bellantuono, Nora Penzel, Joel E. Kleinman, Joo Heon Shin, Thomas M. Hyde, Daniel R. Weinberger, Giulio Pergola.

Presenter affiliation: Lieber Institute for Brain Development, Johns Hopkins University, Baltimore, Maryland; University of Bari Aldo Moro, Bari, Italy.

237

A comprehensive rRNA variation atlas in health and disease

Daphna Rothschild, Teodorus T. Susanto, Jeffrey P. Spence, Naomi R. Genuth, Nasa Sinnott-Armstrong, Jonathan K. Pritchard§, Maria Barna§.

Presenter affiliation: Stanford University, Stanford, California.

238

Genetic variants associated with immune cell population abundances in single-cell data

Laurie Rumker, Saori Sakaue, Joyce B. Kang, Yakir Reshef, Cristian Valencia, Seyhan Yazar, José Alquicira-Hernandez, Joseph Powell, Soumya Raychaudhuri.

Presenter affiliation: Brigham and Women's Hospital, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts.

239

The CZ CELLxGENE Discover suite is an analytical platform and the largest repository of standardized single-cell data

Erica M. Rutherford, Jim Chaffer, Jenny Chien, Lian Morales, Jennifer L. Zamanian, Jason Hilton, Michael Cherry, CZI Single Cell Biology Team.

Presenter affiliation: Stanford University, Palo Alto, California.

240

Age and social status are associated with Th1 and Th2 immune gene regulatory responses in rhesus macaques

Mitchell Sanchez Rosado, Marina M. Watowich, Laura Newman, Melissa A. Pavez-Fox, Erin R. Siracusa, Macaela Skelton, Josue E. Negron-Del Valle, Daniel Phillips, James P. Higham, Lauren Brent, Amanda J. Lea, Carlos Sariol, Noah Snyder-Mackler.

Presenter affiliation: University of Puerto Rico-Medical Sciences, San Juan, Puerto Rico.

241

Application of Nanopore sequencing for liquid biopsy analysis in children with cancer

Carolin M. Sauer, Nicholas Tovey, Debbie Hughes, Marwane Bourdim, Reda Stankunaite, Joanne Stockton, Claire Lynn, Harvey Che, Michael Hubank, John Anderson, Andrew D. Beggs, Louis Chesler, Isidro Cortés-Ciriano.

Presenter affiliation: EMBL, Cambridge, United Kingdom.

242

Isoform-specific functions of RNA-binding proteins

Megan D. Schertzer, Stella H. Park, Erin D. Jeffery, Gloria Sheynkman, David A. Knowles.

Presenter affiliation: New York Genome Center, New York, New York; Columbia University, New York, New York.

243

Defining ancestry, heritability and plasticity of cellular phenotypes in somatic evolution

Joshua S. Schiffman, Andrew R. D'Avino, Tamara Prieto, Yakun Pang, Yilin Fan, Srinivas Rajagopalan, Catherine Potenski, Toshiro Hara, Mario L. Suva, Charles Gawad, Dan A. Landau.

Presenter affiliation: New York Genome Center, New York, New York; Weill Cornell Medicine, New York, New York.

244

CGC1, a new gap-free and telomere-to-telomere reference genome for *Caenorhabditis elegans*

Kazuki Ichikawa, Massa J. Shoura, Karen L. Artiles, Chie Owa, Haruka Kobayashi, Manami Kanamori, Yu Toyoshima, Yuichi Iino, Ann E. Rougvie, Andrew Z. Fire, Erich M. Schwarz, Shinichi Morishita.

Presenter affiliation: Cornell University, Ithaca, New York.

245

Improving precision cancer treatment selection by integrating deep omic tumor characterization and patient-specific drug screening

Casey Sederman, Tony Di Sera, Gabor T. Marth.

Presenter affiliation: University of Utah School of Medicine, Salt Lake City, Utah.

246

A surrogate modeling framework for interpreting deep neural networks in functional genomics

Evan Seitz, Peter Koo, Justin Kinney.

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

247

Dissecting the functional drivers of complex phenotypes by *in vivo* protein-interaction QTL mapping (pi-QTL)

Adrian Serohijos.

Presenter affiliation: University of Montreal, Montreal, Canada.

248

Mitochondrial haplotype and mito-nuclear matching drive somatic mutation and selection through aging

Isabel M. Serrano, Misa Hirose, Clint Valentine, Sharie Austin, Jesse Salk, Saleh Ibrahim, Peter H. Sudmant.

Presenter affiliation: University of California, Berkeley, Berkeley, California.

249

Meta-analysis of epigenetic age in dogs modulated by breed lifespan

Aitor Serres-Armero, Matteo Pellegrini, Steve Horvath, Elaine A. Ostrander.

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

250

DUX4 double whammy—The transcription factor that causes a rare muscular dystrophy also kills the precursors of the human nose

Kaoru Inoue, Hamed Bostan, MaKenna R. Browne, Owen F. Bevis, Carl D. Bortner, Steven A. Moore, Aaron A. Stence, Negin P. Martin, Shih-Heng Chen, Adam B. Burkholder, Jian-Liang Li, Natalie D. Shaw.

Presenter affiliation: NIEHS/NIH, Durham, North Carolina.

251

Transcriptome-wide and proteome-wide association study of Tourette's syndrome

Sudhanshu Shekhar, Peristera Paschou.

Presenter affiliation: Purdue University, West Lafayette, Indiana.

252

EASTR—Improving RNA-seq alignment for accurate transcriptome assembly and reference annotation curation
Ida Shinder, Richard Hu, Hyun Joo Ji, Kuan-Hao Chao, Mihaela Pertea.

Presenter affiliation: Johns Hopkins School of Medicine, Baltimore, Maryland. 253

Changes in circulating cell-free DNA as a biomarker of immune response to short-duration spaceflight

Karolina Sienkiewicz, Kirill Grigorev, Namita Damle, Deena Najjar, Sebastian Garcia Medina, JangKeun Kim, Jonathan Foox, Eliah G. Overbey, Kelly Blease, Juan Moreno, Junhua Zhao, Bryan Lajoie, Andrew Altomare, Semyon Kruglyak, Ari M. Melnick, Jaime Mateus, Christopher E. Mason.

Presenter affiliation: Weill Cornell Medicine (WCM), New York, New York. 254

Gene deserts harbor noncoding functions critically required for normal development

Neil Slaven, Yiwen Zhou, Fabrice Darbellay, Jennifer Akiyama, Ingrid Plazjer-Frick, Catherine Novak, Momoe Kato, Axel Visel, Len Pennacchio.

Presenter affiliation: Lawrence Berkeley National Lab, Berkeley, California. 255

Surveying specific and shared responses of human islets to T2D-associated stressors

Eishani K. Sokolowski, Redwan M. Bhuiyan, Romy Kursawe, Michael L. Stitzel, Duygu Ucar.

Presenter affiliation: University of Connecticut Health Center, Farmington, Connecticut; The Jackson Laboratory For Genomic Medicine, Farmington, Connecticut. 256

A unified framework of realistic in silico data generation and statistical model inference for single-cell and spatial omics

Dongyuan Song, Qingyang Wang, Jingyi Li.

Presenter affiliation: UCLA, Los Angeles, California. 257

T1K—Efficient and accurate KIR and HLA genotyping with next-generation sequencing data

Li Song, Gali Bai, X. Shirley Liu, Bo Li, Heng Li.

Presenter affiliation: Dartmouth College, Lebanon, New Hampshire. 258

Enhancer repression in gene expression fine-tuning

Wei Song, Ivan Ovcharenko.

Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 259

Transcriptional attenuation of CNV amplification and consequences for gene regulatory networks

Pieter Speelman, Grace Avecilla, Julia Matthews, David Gresham.
Presenter affiliation: New York University, New York, New York.

260

The evolutionary history of 17q21.31 structural haplotypes in ancient and modern humans

Samvardhini Sridharan, Peter H. Sudmant.
Presenter affiliation: University of California, Berkeley, Berkeley, California.

261

A machine learning approach to identify functionally relevant endogenous mRNA targets of piRNAs in *C. elegans*

Margaret R. Starostik, Charlotte P. Choi, Rebecca J. Tay, Lars K. Benner, Brooke E. Montgomery, Taiowa A. Montgomery, Michael C. Schatz, John K. Kim.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

262

A method to sequence true full-length capped RNA

Jamie Auxillois, Arnaud Stigliani, Albin Sandelin.
Presenter affiliation: University of Copenhagen, Copenhagen, Denmark.

263

The small RNA transcriptome and its genetic regulation across human tissues

Tim Coorens, Petar Stojanov, Juan Carlos Fernandez del Castillo, Scott Steelman, Sarah Young, Chad Nussbaum, Gad Getz, Kristin Ardlie, François Aguet.
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts.

264

ABM—Benchmarking bioinformatics tools

Keith Suderman, Enis Afgan, Nuwan Goonasekera, Michael Schatz.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland.

265

Single-cell eQTLs mapping in brain cell types reveal context specific genetic regulation and implications in AD genetics

Na Sun, Yongjin Park, Carles Boix, Lei Hou, Xushen Xiong, Yosuke Tanigawa, Xikun Han, Manolis Kellis.
Presenter affiliation: Massachusetts Institute of Technology, Cambridge, Massachusetts; The Broad Institute of Harvard and MIT, Cambridge, Massachusetts.

266

High intraspecies allelic diversity in plant immune receptors is associated with distinct genomic and epigenomic features

Chandler A. Sutherland, Daniil M. Prigozhin, J. G. Monroe, Ksenia V. Krasileva.

Presenter affiliation: University of California, Berkeley, Berkeley, California.

267

Transcriptome-wide meta-analysis of codon usage in *Escherichia coli*

Anima Sutradhar, Jonathan Pointon, Christopher Lennon, Giovanni Stracquadanio.

Presenter affiliation: University of Edinburgh, Edinburgh, United Kingdom.

268

Neotelomeres and telomere-spanning chromosomal arm fusions in cancer genomes revealed by long-read sequencing

Kar-Tong Tan, Michael K. Slevin, Mitchell L. Leibowitz, Max Garrity-Janger, Heng Li, Matthew Meyerson.

Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Harvard Medical School, Boston, Massachusetts.

269

Modeling gene expression through proximal and distal sequence elements using deep learning

Shushan Toneyan, Peter Koo.

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

270

Age-related accumulation of *de novo* indels in mitochondrial DNA of mice and macaques

Edmundo Torres-Gonzalez, Barbara Arbeithuber, Kateryna D. Makova.

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

271

Systematic interpretation of genetic variants that disrupt CTCF binding sites

Colby Tubbs, Mary Lauren Benton, Evonne McArthur, John Capra, Douglas Ruderfer.

Presenter affiliation: Vanderbilt University, Nashville, Tennessee.

272

Strong conservation of promoter-like elements, but not active enhancers, between circulating porcine and human immune cells

Ryan J. Corbett, Juber H. Uribe, Jinyan Teng, Kristen Byrne, Haibo Liu, Houcheng Li, Zhe Zhang, James E. Koltes, Catherine W. Ernst, Crystal L. Loving, Lingzhao Fang, Christopher K. Tuggle.

Presenter affiliation: Iowa State University, Ames, Iowa.

273

Modeling single-cell activation states enhances power to identify ex vivo stimulation response eQTLs

Cristian Valencia, Aparna Nathan, Joyce Kang, Laurie Rumker, Soumya Raychaudhuri.

Presenter affiliation: Brigham & Women's Hospital, Boston, Massachusetts; Harvard University, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts.

274

Spatial transcriptomics analysis reveals pathology-specific cell and molecular changes in pulmonary fibrosis

Annika Vannan, Ruqian Liu, Arianna L. Williams, Evan D. Mee, Mei-i Chung, Saahithi Mallapragada, Jonathan A. Kropski, Davis J. McCarthy, Nicholas E. Banovich.

Presenter affiliation: Translational Genomics Research Institute, Phoenix, Arizona.

275

The distinct genetic determinants of reproductive hormones and infertility

Samvida S. Venkatesh, Laura B. L Wittemans, Benjamin M. Jacobs, Jessica F. Campos de Jesus, Minna Karjalainen, Anu Pasanen, Ahmed Elhakeem, Deborah A. Lawlor, Nicholas J. Timpson, Triin Laisk, Hannele Laivuori, David van Heel, Cecilia M. Lindgren.

Presenter affiliation: Big Data Institute, Oxford, United Kingdom.

276

300 billion associations—Genetic architecture of 2,071 phenotypes in 658,582 individuals of diverse ancestry in the VA Million Veteran Program

Anurag Verma, Jennifer E. Huffman, Alex Rodriguez, Yuk-Lam Ho, Mitchell Conery, Molei Liu, Benjamin Voight, Tianxi Cai, Ravi K. Madduri, Scott M. Damrauer, Katherine P. Liao.

Presenter affiliation: CMC VA Medical Center, Philadelphia, Pennsylvania.

277

Genetic variation shaping the transcriptomic immune response to *Yersinia pestis*

Tauras P. Vilgalys, Mari Shiratori, Anne Dumaine, Luis B. Barreiro.

Presenter affiliation: University of Chicago, Chicago, Illinois.

278

Genetic impacts on DNA methylation help dissect the interplay between genetics, epigenetics and disease

Sergio Villicaña, Juan Castillo-Fernandez, Eilis Hannon, Colette Christiansen, Pei-Chien Tsai, Jane Maddock, Diana Kuh, Matthew Suderman, Christine Power, Caroline Relton, George Ploubidis, Andrew Wong, Rebecca Hardy, Alissa Goodman, Ken K. Ong, Jordana T. Bell.

Presenter affiliation: King's College London, London, United Kingdom.

279

Dissecting the interplay between ageing, sex and body mass index on a molecular level

T D. Michaletou, MG Hong, J Fernandez-Tajes, S Sharma, C A. Brorsson, R W. Koivula, J Adamski, S Brunak, P W. Franks, E T. Dermitzakis, E R. Person, J M. Schwenk, M Walker, A A. Brown, A Viñuela.

Presenter affiliation: Newcastle University, Newcastle, United Kingdom.

280

scMetaBrain—Federated single-cell consortium for cell-type specific eQTL analysis of neurological disease variants

Martijn Vochteloo, Roy Oelen, Drew R. Neavin, Robert Warmerdam, Urmo Võsa, Maryna Korshevniuk, Dan Kaptijn, Monique van der Wijst, Marc Jan Bonder, scEQTLGen Consortium, Tõnu Esko, Julien Bryois, Ellen A. Tsai, Heiko Runz, Lude Franke, Harm-Jan Westra.

Presenter affiliation: University Medical Center Groningen, University of Groningen, Groningen, Netherlands; Oncode Investigator, Utrecht, Netherlands.

281

The poor portability of polygenic scores is only partially attributable to genetic ancestry

Joyce Y. Wang, Michael Zietz, Jason Mares, Paul J. Rathouz, Vagheesh M. Narasimhan, Molly F. Przeworski, Arbel Harpak.

Presenter affiliation: The University of Texas at Austin, Austin, Texas.

282

CoSpar identifies early cell fate biases from single cell transcriptomic and lineage information

Shou-Wen Wang, Michael J. Herriges, Kilian Hurley, Darrell N. Kotton, Allon M. Klein.

Presenter affiliation: Harvard Medical School, Blavatnik Institute, Boston, Massachusetts; Westlake University, Hangzhou, China.

283

The role of race and genetic ancestry in KRAS mutation and subtypes in non-small cell lung cancer

Xinan Wang, Kangcheng Hou, Rounak Dey, Biagio Ricciuti, Xihong Lin, Bruce E. Johnson, David C. Christiani.

Presenter affiliation: Harvard University, Boston, Massachusetts.

284

Impact of individual level uncertainty of lung cancer polygenic risk score on risk stratification and prediction

Xinan Wang, Ziwei Zhang, Tony Chen, Yi Ding, Xihong Lin, David C. Christiani.

Presenter affiliation: Harvard University, Boston, Massachusetts.

285

Calypso—Longitudinal genomic diagnostic care with innovative web tools	
Alistair Ward, Isabelle Cooperstein, Tony Di Sera, Stephanie Georges, Anders Pitman, Marti Tristani-Firouzi, Gabor Marth.	
Presenter affiliation: University of Utah, Salt Lake City, Utah; Frameshift Labs, Cambridge, Massachusetts.	286
Top2B-inhibiting breast cancer drugs including anthracyclines affect cardiomyocyte health through a shared gene expression response signature	
Elizabeth R. Matthews, Omar D. Johnson, Kandace J. Horn, Jose A. Gutierrez, Simon Powell, <u>Michelle C. Ward</u> .	
Presenter affiliation: University of Texas Medical Branch, Galveston, Texas.	287
Transforming genomics research through community engagement and return of results—A case study from French Polynesia	
<u>Kaja A. Wasik</u> , Sarah LeBaron von Baeyer, Keolu Fox, Tristan Pascart, Tehani Mairai, Vehia Wheeler.	
Presenter affiliation: Variant Bio, Seattle, Washington.	288
Rural and urban lifestyles are associated with differential immune gene regulation in Turkana	
<u>Marina M. Watowich</u> , Kristina M. Garske, Varada Abhyankar, Echwa John, Michael Gurven, John Kahumbu, Joseph Kamau, Dino J. Martins, Charles Miano, Benjamin Muhoya, Julie Peng, Jenny Tung, Julien F. Ayroles, Amanda J. Lea.	
Presenter affiliation: Vanderbilt University, Nashville, Tennessee.	289
A computational toolkit to integrate multi-omics time-series data across species in brain development	
Beatrice Borsari, <u>Eve S. Wattenberg</u> , Ke Xu, Xuezhu Yu, Mor Frank, Susanna Liu, Mark Gerstein.	
Presenter affiliation: Yale University, New Haven, Connecticut.	290
Methylome, transcriptome and alternative splicing profiling of neurons, astrocytes, and microglia	
Xiaoran Wei, Michelle L. Olsen.	
Presenter affiliation: Virginia Tech, Blacksburg, Virginia.	291

Bayesian causal inference of gene regulatory networks from CRISPR perturbations in CD4+ T cells

Joshua S. Weinstock, Maya Arce, Jacob W. Freimer, Mineto Ota, Alexander Marson, Alexis Battle, Jonathan K. Pritchard.

Presenter affiliation: Stanford University, Stanford, California; Johns Hopkins University, Baltimore, Maryland.

292

Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants

Jonas Koeppel, Juliane Weller, Elin Madli Peets, Ananth Pallasani, Ivan Kuzmin, Uku Raudvere, Hedi Peterson, Fabio Liberante, Leopold Parts.

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

293

Uncovering gene fusions with 3D genomics—From clinical validation to actionable insights for undiagnosable solid tumors

Allyson Whittaker, Kristin Sikkink, Anthony Schmitt, Kristyn Galbraith, Michelle Perez-Arreola, Misha Movahed-Ezazi, George Jour, Matija Snuderl.

Presenter affiliation: Arima Genomics, Carlsbad, California.

294

Spatial regression models for the analysis of chromatin conformation data

Wilfred Wong, Julian Pulecio, Renhe Luo, Nan Zhang, Jielin Yan, Effie Apostolou, Danwei Huangfu, Christina Leslie.

Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York; Tri-Institutional Training Program, New York, New York.

295

Structure-aware annotation of solenoid protein domains

Boyan Xu, Daven Lim, Christopher J. Tralie, Alois Cerbu, Daniil Prigozhin, Ksenia Krasileva.

Presenter affiliation: UC Berkeley, Berkeley, California.

296

Structural variants drive context dependent oncogene activation in cancer

Zhichao Xu, Dong-Sung Lee, Sahaana Chandran, Victoria T. Le, Rosalind Bump, Jean Yasis, Sofia Dallarda, Samantha Marcotte, Benjamin Clock, Nicholas Haghani, Chae Yun Cho, Kadir Akdemir, Selene Tyndale, P. Andrew Futreal, Graham McVicker, Geoffrey M. Wahl, Jesse R. Dixon.

Presenter affiliation: Salk Institute for Biological Studies, La Jolla, California.

297

Mapping the <i>cis</i>- and <i>trans</i>-regulatory landscape of the immune checkpoint PD-L1 with paired genetic screens	
<u>Xinhe Xue</u> , Zoran Gajic, Christina Caragine, Mateusz Legut, Conor Walker, James Kim, Hans-Hermann Wessels, Congyi Lu, Gamze Gursoy, Neville E. Sanjana.	
Presenter affiliation: New York Genome Center, New York, New York; New York University, New York, New York.	298
Predicting the structure and function of alternative proteins	
<u>Feriel Yala</u> , Sebastien Leblanc, Xavier Roucou.	
Presenter affiliation: University of Sherbrooke, Sherbrooke, Canada.	299
Epiphany—Predicting the Hi-C contact map from 1D epigenomic data	
<u>Rui Yang</u> , Arnav Das, Vianne R. Gao, Alireza Karbalayghareh, William S. Noble, Jeffrey A. Bilmes, Christina S. Leslie.	
Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York City, New York.	300
High level of structural complexity of canine olfactory receptor gene families revealed by genome assemblies of six dog breeds	
<u>Feyza Yilmaz</u> , Kwondo Kim, Pille Hallast, Wonyeong Kang, Qihui Zhu, Charles Lee.	
Presenter affiliation: The Jackson Laboratory, Farmington, Connecticut.	301
Proteomic analyses reveal mechanistic links between clonal hematopoiesis of indeterminate potential and coronary artery disease	
<u>Zhi Yu</u> , Bing Yu, Amélie Vromman, Ngoc Quynh H. Nguyen, Alexander G. Bick, Benjamin L. Ebert, Rajat M. Gupta, Peter Libby, Robert E. Gerszten, Pradeep Natarajan.	
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts.	302
Accurate quantification of multi-mapping reads in single-cell RNA-seq with STARsolo	
<u>Dinar Yunusov</u> , Nathan Castro-Pacheco, Alexander Dobin.	
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.	303

Understanding caudal developmental abnormalities using single-nucleus multi-omics data from wild type and Danforth's short tail mouse E9.5 tailbuds

Cynthia K. Zajac, Ricardo D. Albanus, Nandini Manickam, Erika Curka, Catherine Keegan, Stephen C. Parker.

Presenter affiliation: University of Michigan, Ann Arbor, Michigan.

304

Denisovan and Neandertal gene variants influence brain morphology in present-day people

Hugo Zeberg.

Presenter affiliation: Karolinska Institutet, Stockholm, Sweden; Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

305

Genetic perturbation of PU.1 binding in microglia and disease risk association

Falak Sher, Lu Zeng, Hans-Ulrich Klein, Julie J. McInvale, Philip L. De Jager.

Presenter affiliation: Columbia University Irving Medical Center, New York, New York.

306

Model-based characterization of the equilibrium dynamics of transcription initiation and promoter-proximal pausing in human cells

Yixin Zhao, Lingjie Liu, Adam Siepel.

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

307

Multiomic analysis reveals cellular and epigenetic plasticity in intestinal pouches of ulcerative colitis patients

Yu Zhao, Ran Zhou, Bingqing Xie, Cambrian Y. Liu, Martin Kalski, Candace M. Cham, Jason Koval, Christopher R. Weber, Jingwen Xu, David T. Rubin, Mitch Sogin, Sean Crosson, Jun Huang, Aretha Fiebig, Sushila Dalal, Eugene B. Chang, Anindita Basu, Sebastian Pott.

Presenter affiliation: University of Chicago, Chicago, Illinois.

308

Genome in a Bottle benchmarks in the era of complete human genomes

Nathan D. Olson, Justin Wagner, Nathan Dwarshuis, Adam English, Fritz J. Sedlazeck, Justin M. Zook, Genome in a Bottle Consortium. Presenter affiliation: National Institute of Standards and Technology, Gaithersburg, Maryland.

309

AUTHOR INDEX

- Abad, Amaya, 218, 219
Abdill, Richard J., 131
Abel, Haley J., 186
Abhyankar, Varada, 124, 289
Abner, Erik, 206
Abney, Mark, 177
Abueg, Linelle, 165
Abyzov, Alexej, 212
Adams, Stuart, 94
Adamski, J, 280
Adeluwa, Temidayo, 53, 129
Adhikari, Aashish, 3
Afgan, Enis, 54, 265
Agrawal, Saumya, 55
Ågren, Richard, 56
Aguet, Francois, 30, 62, 95, 161,
 225, 264
Aguilar, Robin, 57
Aguirre, Matthew, 58
Ahituv, Nadav, 126
Ahluwalia, Rhea, 59
Ahmed, Omar, 60
Aiden, Erez, 132
Akagi, Keiko, 61
Akdemir, Kadir, 297
Akiyama, Jennifer, 157, 255
Akiyama, Yo, 62
Akula, Nirmala, 63
Alasadys, Milad, 72
Alazizi, A., 228
Albanus, Ricardo D., 304
Albert, Frank W., 71, 131
Alcaraz, Nicolas, 112
Alcedo, Karel, 13
Alexander, Joan, 171
Alipanahi, Babak, 84
Allen, Jamie, 64
Alquicira-Hernandez, José, 239
Altomare, Andrew, 254
Alvarez Jerez, Pilar, 125
Amariuta, Tiffany, 53
Ameen, Mo, 161
Amir-Zilberstein, Liat, 16
Anand, Shankara, 62
Anderson, John, 242
Anderson, Nathan W., 65
Andersson, Jesper, 50
Andersson, Robin, 112
André, Mathilde, 66
Andreeva, Kalina, 139
Angelo, Michael, 47
Angom, Ramcharan S., 35
Antonucci, Linda A., 154
Apostolou, Effie, 295
Aquino, Yann, 4
Arbeithuber, Barbara, 271
Arbesfeld, Jeremy A., 67
Arce, Maya, 292
Ardlie, Kristin, 95, 225, 264
Arehart, Christopher H., 175
Arking, Dan, 173
Arnan, Carme, 219
Arndt, Peter F., 68
Arriagada, Gloria, 136
Arteaga, Stephanie, 173
Arthofer, Christoph, 50
Artiles, Karen L., 245
Ashley, Euan A., 202
Asri, Mobin, 18
Aston, Kenneth, 163, 187
Atabaki-Pasdar, Naeimeh, 104
Audano, Peter A., 43, 69, 116
Austin, Sharie, 249
Austine-Orimoloye, Olanrewaju,
 64
Auwerx, Chiara, 70
Auxillos, Jamie, 263
Avecilla, Grace, 260
Avery, Randi R., 71
Ayroles, Julien F., 124, 289
Azama, Chika, 209
Babur, Özgün, 62
Bachtrog, Doris, 40
Baczenas, John J., 109
Bae, Taejeong, 212
Bai, Gali, 258
Bai, Lijing, 180
Balachandran, Parithi, 43, 116
Balasubramanian, Suganthi, 233

- Baldassari, Sara, 168
Banerjee, Deepro, 103
Banerjee, Shreya M., 109
Banovich, Nicholas E., 28, 275
Barbadilla, Antonio, 128
Barber, Galt P., 201
Barna, Maria, 238
Barreiro, Luis B., 278
Barrett, Nora, 108
Barry, Timothy, 152
Bartom, Elizabeth T., 72
Bartonn, Sarah, 157
Bartram, Jack, 94
Bartsch, Andreas, 50
Basheer, Becerra, 220
Basu, Anindita, 221, 308
Batista, Thiago M., 144
Battle, Alexis, 173, 202, 292
Batut, Philippe J., 73
Bauer, Daniel, 220
Baulac, Stéphanie, 168
Beal, Brian F., 136
Beaumont, Robin N., 137
Beck, Christine R., 43, 69, 116
Beckman, Erik, 157
Beckmann, Noam D., 38
Beekman, Renée, 29
Beer, Michael, 132
Beggs, Andrew D., 242
Behjati, Sam, 94
Belbin, Gillian M., 21
Beliveau, Brian J., 57
Bell, Jessica L., 167
Bell, Jordana T., 279
Bellantuono, Loredana, 237
Bendesky, Andres, 41
Benjamin, Kynon J M., 74
Benner, Lars K., 262
Bennett, David A., 91
Benton, Mary Lauren, 272
Berger, Alice H., 215
Bertolino, Alessandro, 154
Bevis, Owen F., 251
Bezney, Jon, 189
Bhuiyan, Redwan M., 256
Biagini, Julia, 210
Bianchi, Agostina, 29
Bick, Alexander G., 302
Bielas, Stephanie, 23
Billingsley, Kimberley J., 125
Bilmes, Jeffrey A., 300
Bisiaux, Aurélie, 4
Black, Gage, 75
Blattner, Frank, 114
Blauwendraat, Cornelis, 125
Blease, Kelly, 254
Blekhman, Ran, 131
Bloemendaal, Alex, 35
Blow, Frances, 82
Blümcke, Ingmar, 168
Boateng-Sarfo, George, 76
Bobo, Dean M., 208
Boehnke, Michael, 23, 87
Bohaczuk, Stephanie C., 77
Boix, Carles, 266
Boland, Anne, 66
Bolduc, Jessica, 135
Bond, Marielle L., 167
Bonder, Marc Jan, 232, 281
Bønnelykke, Klaus, 119
Bonnycastle, Lori, 87
Booeshaghi, Sina, 230
Borchers, Matthew, 45
Borczyk, Małgorzata, 78
Borel, Christelle, 229
Bornholdt, Jette, 112
Borsari, Beatrice, 148, 218, 290
Bortner, Carl D., 251
Bosio, Mattia, 225
Boßelmann, Christian M., 168
Bossoms Mesa, Alba, 191
Bostan, Hamed, 251
Botts, Travis, 210
Bourdim, Marwane, 242
Bowling, Sarah, 13
Boyce, Joshua, 108
Bradley, Kenneth, 227
Bradley, Robert K., 215
Brajuka, Nadolina, 165
Brassington, Layla, 194
Brent, Lauren, 96, 241
Bridge, Caroline, 79
Brorsson, C A., 280
Brown, Andrew A., 104, 110,
 229, 280
Brown, Matthew W., 153

- Brown, Samantha, 191
Browne, MaKenna R., 251
Browning, Jared W., 205
Brucato, Nicolas, 66, 142
Brückner, Lotte, 52
Bruinsma, J., 228
Brunak, Søren, 50, 280
Bryois, Julien, 110, 281
Bubb, Kerry, 130
Buckler, Edward S., 130
Buckley, Reuben M., 80
Buels, Robert, 79
Buenrostro, Jason D., 166
Bump, Rosalind, 297
Buonaiuto, Silvia, 123
Burkholder, Adam B., 251
Burns, Philippa, 97, 98
Busch, Robyn M., 168
Buzby, Cassandra, 81
Byrne, Kristen, 273
Bzikadze, Andrey, 18
- Cáceres, Mario, 83, 128
Cadena Castaneda, Diana, 140
Cai, Biao, 22
Cai, Tianxi, 277
Cain, Scott, 79
Callejas-Hernández, Francisco, 82
Callendar, Megan, 140
Calvo, Miquel, 225
Camargo, Fernando D., 13
Campilsson, Conor K., 57
Campos de Jesus, Jessica F., 276
Campoy, Elena, 83
Cao, Dan, 30
Capella, Salvador, 225
Capra, John, 272
Caragine, Christina, 31, 298
Carilli, Maria, 230
Carlson, Jedidiah, 90
Carlton, Jane M., 82
Casolaro, Tommy, 7
Casper, Jonathan, 201
Castillo-Fernandez, Juan, 279
Castro, Matheus P., 100
Castro, Rodrigo I., 32
- Castro-Pacheco, Nathan, 303
Cavazos, Taylor B., 84
Cechova, Monika, 18, 45
Cerbu, Alois, 296
Chaffer, Jim, 240
Cham, Candace M., 221, 308
Chamorro González, Rocío, 52
Chan, Candace, 126
Chan, Yee-Ming, 50
Chandran, Sahaana, 297
Chang, Eugene B., 221, 308
Chang, Howard Y., 48
Chao, Katherine, 26, 85
Chao, Kuan-Hao, 253
Charney, Alexander W., 38
Chatterjee, Soumick, 88
Chazarra-Gil, Ruben, 225
Che, Harvey, 242
Chekuri, Anil, 135
Chen, Ken, 174
Chen, Ling, 30
Chen, Qiang, 74, 154
Chen, Shih-Heng, 251
Chen, Siwei, 26
Chen, Tony, 285
Chen, Xinyi E., 105
Cheng, Feixiong, 115, 141
Cherry, J Michael, 139, 240
Chesler, Louis, 242
Chien, Jenny, 240
Chin, Christopher, 17
Ching, Joshua, 210
Cho, Chae Yun, 297
Choi, Charlotte P., 262
Christiani, David C., 284, 285
Christiansen, Colette, 279
Chugh, Sagrika, 28
Chung, Erin, 86
Chung, Mei-i, 275
Chung, Yeun-Jun, 214
Church, David N., 153
Church, George, 156
Churchman, L. Stirling, 146
Ciotlos, Dan L., 87
Cirnigliaro, Matilde, 173
Cisternino, Francesco, 88
Claussnitzer, Melina, 144
Clawson, Hiram, 201

- Cline, Melissa, 67
Clock, Benjamin, 297
Cloutier, Marie-Pier, 149
Cochran, Kelly, 89
Cole, Jared M., 59
Collado-Torres, Leonardo, 74
Collins, Francis S., 23, 87
Collord, Grace, 46, 94
Colonna, Marco, 140
Colonna, Vincenza, 123
Comandante-Lou, Natacha, 91
Comoglio, Yannick, 153
Conceição, Helena B., 100, 121
Conery, Mitchell, 277
Conrad, Thomas, 52
Cook, Laura E., 92
Coombes, Kevin R., 61
Cooperstein, Isabelle, 93, 286
Coorens, Tim, 46, 94, 95, 264
Corbett, Ryan J., 273
Corominas, Montserrat, 183
Correa, Bruna R., 218
Cortes-Ciriano, Isidro, 51, 199, 242
Costa, Christina E., 96
Couldrey, Christine, 149
Couvillion, Mary, 146
Cox, Anthony J., 161
Cox, Laura A., 177, 203
Cox, Murray, 66, 142
Crawford, Robert, 97
Creed, Páidí, 97, 98
Crooks, Kristy R., 175
Crosson, Sean, 308
Crowan, Mitra, 149
Cuacos, Maria, 114
Cullina, Sinead, 21
Cummings, Jeffrey, 141
Cunningham, Fiona, 64
Cuoco, Michael S., 99
Cuperus, Joshua T., 130
Curka, Erika, 304
Currall, Benjamin, 135
Currant, Hannah, 50
Currie, David, 98
Curtis, Christina, 47, 143
Cvejic, Ana, 189
Czaplinska, Dominika, 101
Czechanski, Anne, 43
Dadon, Sara, 196
Dahl, Andy, 129
Dai Yang, Karren, 16
Dai, Ruija, 190
Dai, Yifan, 101
Dai, Yulin, 178
Dalal, Sushila, 308
Dallarda, Sofia, 297
Daly, Mark, 26, 85
D'Amelio, Katie, 184
Damle, Namita, 254
Damrauer, Scott M., 277
Dannemann, Michael, 66
Darbellay, Fabrice, 255
Das, Arnav, 300
Das, Arun, 102
Das, Maitreya, 103
Dashti, Hesam, 144
D'Avino, Andrew R., 244
Davis, Brian W., 136
Davis, Jessica, 167
Davtian, David, 104
de Assis, Pedro R., 139
de Boer, Carl G., 105
De Filippo, Cesare, 191
de Hoon, Michiel, 55, 106
De Jager, Philip L., 91, 306
de Noon, Solange, 51
De Zuani, Marco, 189
Degner, Jacob, 204
Del Azodi, Christina, 28
Deleuze, Jean-François, 66
Delfosse, Kate, 205
Delorey, Toni, 16
Delprat, Alejandra, 83
Deplancke, Bart, 159
Derevianko, Anatoly P., 191
Dermitzakis, Emmanouil T., 229, 280
Derry, Brent, 59
DeSalle, Robert, 208
Deshpande, Atul, 236
Desjardins, Jade, 149
Dewey, Hannah, 25
Dey, Kushal K., 132
Dey, Rounak, 284

- Dhiman, Alisha, 107
Dhindsa, Ryan S., 122
Di Sera, Tony, 246, 286
Dickel, Diane E., 157
Diekhans, Mark, 201
Diesh, Colin, 79
Dietrich, Anastasia, 3
Ding, Yi, 285
Dishuck, Philip, 133
Dixon, Jesse R., 297
Djeddi, Sarah, 108
Dobin, Alexander, 303
Dodge, Tristram O., 109
Domingo, Julia, 39
Dong, Shencheng, 139
Donnard, Elisa, 7
Dor, Yuval, 235
Douaud, Gwenaëlle, 50
Douka, Katerina, 191
Drucker, Kristen L., 212
Dubocanin, Danilo, 146
Dumaine, Anne, 278
Dunn, Joe D., 61
Dupuis, Theo, 104, 110
Dwarshuis, Nathan, 111, 309
Dykhuijen, Emily, 107
- Eagles, Nicholas J., 74
Ebeling, Martin, 110
Ebert, Benjamin L., 302
Ede, Jeffrey, 3
Edge, Michael D., 5
Eggert, Angelika, 52
Eichler, Evan E., 6, 18, 43, 45, 127, 133, 142
Einarsson, Hjörleifur, 112
Elhakeem, Ahmed, 276
Eliasen, Anders Ulrik, 119
Ellis, James, 205
Ellis, Tom, 156
Elrick, Hillary, 51
Engle, Elizabeth L., 236
English, Adam, 309
Engreitz, Jesse M., 35, 132
Eraslan, Basak, 16
Erdin, Serkan, 135
Erdos, Michael R., 87
Ernst, Catherine W., 273
- Ernst, Kayla, 226
Ersaro, Nicole, 202
Erwin, Jennifer A., 99
Esko, Tonu, 206, 281
Espejo Valle-Inclan, Jose, 51
Esteban, Alexandre, 218, 219
Evans, Nicholas, 66
- Fabregat, Isabel, 183
Fachi, Jose, 140
Fan, Yilin, 244
Fang, Lingzhao, 180, 273
Fang, Shi, 35
Farh, Kyle, 3, 30, 161
Farrell, Andrew, 75
Fauci, Christiana, 42, 113
Feigin, Charles Y., 92
Feng, Jia-Wu, 114
Feng, Yayan, 115
Fernandez del Castillo, Juan Carlos, 264
Fernandez-Salinas, Daniela, 108
Fernandez-Tajes, J., 280
Ferraj, Ardian, 43, 116
Ferraro, Nicole, 173
Ferreira, Mark, 13
Ferreira, Pedro G., 8, 225
Ferreira, Raphael, 156
Ferreira, Teresa, 50
Fiebig, Aretha, 308
Field, Yair, 3
Fields, Stanley, 130
Filimban, Ghassan, 230
Finucane, Hilary, 25, 35
Fire, Andrew Z., 245
Firer, Danielle, 95
Fischer, Christian, 123
Fischer, Clay, 201
Fiziev, Petko, 3
Flanagan, Adrienne M., 51
Fleming, Elizabeth, 140
Flores, Jacob I., 43
Fong, Sarah, 126
Footh, Jonathan, 254
Foreman, Julia, 67
Formenti, Giulio, 165
Fornelos, Nadine, 210
Fox, Keolu, 288

- Francoeur, Eden, 116
Frank, Mor, 290
Franke, Lude, 232, 281
Franks, Paul W., 104, 280
Fransen, Tony, 149
Frayling, Timothy M., 137
Frazer, Jonathan, 117
Freimer, Jacob W., 292
Frenkel, Max, 118
Fritz, Amelie, 119
Frommolt, Jeanette, 209
Fuchs, Jörg, 114
Fuentes, Natalia, 32
Fujiki, Saera, 55
Fujita, Masashi, 91
Fukunaga, Izumi, 209
Fukuyama, Hidehiro, 55
Fullgrabe, Jens, 97, 98
Fulton, Robert S., 18
Futreal, P. Andrew, 297
- Gabdank, Idan, 139
Gabel, Austin M., 215
Gadde, Rohini, 99
Gage, Fred H., 99
Gagneur, Julien, 120
Gajic, Zoran, 298
Galante, Pedro, 100, 121
Galbraith, Kristyn, 294
Galeev, Timur, 148
Gallardo, Cristobal, 165
Gallego Romero, Irene, 92
Gallina, Iryna, 99
Gamache, Isabelle, 149
Gandal, Michael, 190
Gankin, Dennis, 120
Gao, Dadi, 135
Gao, Hong, 3, 161
Gao, Vianne R., 300
Garcia Medina, Sebastian, 254
Garcia, Raquel, 8
García-Pérez, Raquel, 225
Gardeux, Vincent, 159
Garg, Manik, 122
Garrison, Erik, 123
Garrity-Janger, Max, 269
Garske, Kristina M., 124, 289
Gawad, Charles, 244
- Geffen, Yifat, 62
Geiger-Schuller, Kathryn, 16
Gelb, Bruce D., 150
Genner, Rylee M., 125
Genuth, Naomi R., 238
Georgakopoulos-Soares, Ilias, 126
Georges, Stephanie, 286
Gern, James, 108
Gerstein, Mark, 148, 290
Gerszten, Robert E., 302
Gerton, Jennifer L., 18, 45, 123
Geschwind, Daniel, 173
Getz, Gad, 62, 94, 95, 264
Ghattaoraya, Gurpreet, 64
Ghotra, Rohan, 224
Giersch, Rachael M., 136
Gignoux, Christopher R., 21, 175
Gilad, Yoav, 11, 177, 203
Gill, Diljeet, 98
Gillison, Maura L., 61
Girirajan, Santhosh, 103
Girling, Gareth, 156
Glaser, Benjamin, 235
Glastonbury, Craig A., 88
Glick, Adam, 103
Goddard, Pagé, 202
Godfrey, Anna, 94
Godley, Lucy A., 207
Godoy, Janeth, 157
Goff, Stephen P., 136
Goldberg, Gary, 171
Goldberg, Michael, 127, 163, 187, 222
Goldman, Elisabeth A., 96
Golightly, Peter R., 90
Gomes de Lima, Leonardo, 123
Gómez Graciani, Ruth, 83, 128
Gona, Saideep, 53, 129
Goncalves, Angela, 9
Goodarzi, Hani, 84
Goodarzi, Patrick A., 84
Goodman, Alissa, 279
Goodwin, Sara, 61
Goonasekera, Nuwan, 54, 265
Gordon, David S., 43
Gorjifard, Sayeh, 130
Gosai, Sager J., 32

- Grady, Patrick, 45
Graham, Samantha P., 131
Gray, Charlotte, 149
Green, Gilad S., 91
Green, Kim, 230
Greenleaf, William J., 132
Greenwald, Noah F., 47
Gresham, David, 260
Grigorev, Kirill, 254
Groso, Daniel, 215
Grüning, Bjorn, 165
Gschwind, Andreas R., 132
Guardia, Gabriela, 100, 121
Guarracino, Andrea, 123
Guigó, Roderic, 132, 218, 219, 225
Guitart, Xavi, 133
Gundlach, Heidrun, 114
Gunn, Theresa R., 109
Guo, Katherine, 35
Guo, Xinyi, 31
Gupta, Rahul, 26
Gupta, Rajat M., 35, 302
Gursoy, Gamze, 298
Gurven, Michael, 124, 289
Gutierrez, Jose A., 287
Gutierrez-Arcelus, Maria, 108
- Haase, Kerstin, 52
Habib, Naomi, 91
Hacohen, Nir, 16
Haeussler, Maximilian, 201
Haga, Yasuhiko, 134
Haggerty, Leanne, 64
Haghani, Nicholas, 297
Hajto, Jacek, 78
Hall, April, 195
Hall, Ira, 186
Hallast, Pille, 301
Hallgrimsdottir, Ingileif, 230
Hamm, Morgan O., 77
Hamp, Tobias, 3
Han, Shizhong, 74
Han, Xikun, 266
Hanks, Sarah C., 87
Hannon, Ellis, 279
Hanotte, Olivier, 155
Hansen, Katja G., 146
- Hara, Toshiro, 244
Harding, Nicholas J., 98
Hardy, Rebecca, 279
Harland, Chad, 149
Harpak, Arbel, 5, 90, 282
Harper, Andrew, 122
Harrington, Anne, 157
Harripaul, Ricardo S., 135
Harris, Alex C., 80
Harris, Kelley, 2, 127
Harris, Robert S., 45
Hart, Samuel F., 136
Hartley, Gabrielle A., 45
Harvey, William T., 142
Hashimoto, Shinichi, 147
Hawkes, Gareth, 137
He, Xin, 221
He, Xionglei, 143
Heckman, Caroline, 206
Heckmann, Stefan, 114
Heide, Christian, 191
Heilmann, Jadin, 195
Hemberg, Martin, 126
Hendricks, Audrey, 175
Henssen, Anton G., 52
Herranz-Itúrbide, Macarena, 183
Herriges, Michael J., 284
Herron, Todd, 23
Hess, Julian, 95
Hevers, Wulf, 209
Hiatt, Laurel, 138, 187, 222
Hickey, Glenn, 161
Higgins, Eve, 217
Higham, James P., 96, 241
Higham, Thomas, 191
Hilton, Jason, 240
Himmelbach, Axel, 114
Hinrichs, Angie S., 201
Hirose, Misa, 249
Hitz, Benjamin C., 139
Ho, Yuk-Lam, 277
Hoekzema, Kendra, 6, 142
Hoffmann, Lucas, 168
Holbrook, Joanna D., 97, 98
Holmes, Ian, 79
Hong, MG, 280
Horlova, Anastasiia, 86
Hormoz, Sahand, 199

- Hormozdiari, Fereydoun, 84
Horn, Kandace J., 287
Horner, Vanessa, 195
Horvath, Mian, 140
Horvath, Steve, 250
Hossain, Ayaan, 103
Hossain, S. Nakib, 64
Hotaling, James, 138, 163
Hou, Kangcheng, 284
Hou, Lei, 266
Hou, Liangzhen, 143
Hou, Yuan, 141
Houben, Andreas, 114
Houlahan, Kathleen E., 47
Hsieh, PingHsun, 133, 142
Hsu, Yu-Han, 210
Hu, Calvin, 166
Hu, Richard, 253
Hu, Zheng, 143
Huang, George, 108
Huang, Jun, 308
Huang, Xiaomeng, 75
Huang, Yi, 144
Huangfu, Danwei, 295
Hubank, Michael, 242
Huber, Christian D., 217
Hudjasov, Georgi, 66
Huffman, Jennifer E., 277
Huggins, Billy J., 145
Hughes, Debbie, 242
Hung, King L., 48
Hunt, Sarah E., 64
Hunter, Riana D., 157
Hurley, Kilian, 284
Hussain, Said, 217
Huuki-Myers, Louise A., 74
Hyde, Thomas M., 74, 237
- Ialchina, Renata, 101
Iasi, Leonardo N., 191
Ibrahim, Saleh, 249
Ichikawa, Kazuki, 245
Iino, Yuichi, 245
Im, Hae Kyung, 53, 104, 129, 203
Inoue, Kaoru, 251
Insigne, Kimberly, 167
Iqbal, Sumaiya, 67
- Isaac, R. Stefan, 146
Ishikawa, Fumihiro, 55, 106
Itan, Yuval, 150
Iwabuchi, Sadahiro, 147
- Jackson, Anne U., 87
Jacobs, Benjamin M., 276
Jacomini, Reicardo, 99
Jaffe, Andrew E., 74
Jagoda, Evelyn, 132
Jain, Miten, 125
Jang, Yeongjun, 212
Jankowiak, Martin, 220
Jean-Baptiste, Ken, 30
Jeffery, Erin D., 243
Jenkins, Robert B., 212
Jensen, Cassandra, 105
Jensen, Matthew, 148
Jensen, Tanner, 202
Jeong, Choongwon, 155
Jeong, Hyeonsoo, 45
Jeyamohan, Prashanti, 55
Jha, Anupama, 77
Jia, Peilin, 176
Jiang, Bo, 61
Jivanji, Swati, 149
Jõeloo, Maarja, 70
John, Echwa, 124, 289
Johnson, Bruce E., 284
Johnson, Mackenzie M., 90
Johnson, Omar D., 287
Johnson, Rory, 218, 219
Jolanki, Otto, 139
Jones, Samuel E., 206
Jones, Thouis R., 7
Joo Ji, Hyun, 253
Jores, Tobias, 130
Jou, Jennifer, 139
Joubert, Pierre M., 200
Jour, George, 294
Ju, Xiangchun, 209
Jung, Hyunchul, 46
- Kadri, Sabah, 204
Kagda, Meenakshi S., 139
Kahumbu, John, 124, 289
Kales, Susan, 25
Kalski, Martin, 308

- Kamau, Joseph, 124, 289
Kamlapurkar, Shriya, 215
Kan, Lijuan, 76
Kanai, Masahiro, 26
Kanamaru, Emi, 106
Kanamori, Manami, 245
Kang, Helen, 35
Kang, Joyce, 239, 274
Kang, Wonyeong, 301
Kaplan, Tommy, 235
Kaptijn, Dan, 281
Karayel, Ozge, 16
Karbalayghareh, Alireza, 17, 132, 300
Karczewski, Konrad, 26, 85
Kariwiga, Jason, 66
Karjalainen, Anzhelika, 220
Karjalainen, Minna, 276
Karollus, Alexander, 120
Karpan, Kirby, 184
Karpinski, Marcin, 122
Kars, Meltem Ece, 150
Kashima, Yukie, 151
Kato, Momoe, 157, 255
Katsevich, Eugene, 152
Kawauchi, Shimako, 230
Kedzierska, Katarzyna Z., 153
Keegan, Catherine, 304
Keener, Rebecca, 173
Kellis, Manolis, 266
Kelso, Janet, 191
Kendall, Jude, 171
Kendle, Robert, 28
Kendziorski, Christina, 108
Kenny, Eimear E., 21
Kent, William J., 201
Khan, Aziz, 47
Khoury, Jean, 168
Kikidis, Gianluca C., 154
Killilea, Alison, 6
Kilpinen, Helena, 14
Kim, Donghee, 155
Kim, Heebal, 155
Kim, James, 298
Kim, JangKeun, 254
Kim, Jean J., 212
Kim, John K., 262
Kim, Kwondo, 155, 301
Kinali, Meric, 62
Kinney, Justin, 247
Kinyua, Patriciah, 124
Kirchner, Emily G., 135
Kiriakopoulos, Katerina, 205
Kirkpatrick, Mark, 5, 90
Klein, Allon M., 13, 284
Klein, Cecilia C., 218
Klein, Hans-Ulrich, 306
Kleinman, Joel E., 74, 237
Kluzer, Alice, 27
Knowles, David, 31, 243
Kobayashi, Haruka, 245
Koche, Richard P., 52
Koeppel, Jonas, 156, 293
Koistinen, Heikki A., 23, 87
Koivula, R W., 280
Kollmeyer, Thomas M., 212
Koltes, James E., 273
Koo, Peter, 224, 247, 270
Koren, Sergey, 18, 45, 123
Kornienko, Aleksandra E., 33
Korostynski, Michal, 78
Korsheviuk, Maryna, 281
Kosicki, Michael, 157, 211
Kotton, Darrell N., 284
Koval, Jason, 221, 308
Kozlikin, Maxim B., 191
Krasileva, Ksenia V., 158, 200, 267, 296
Kreevan, Rita, 66
Kribelbauer, Judith F., 159
Kronauer, Daniel J., 164
Kropski, Jonathan, 28, 275
Kruglyak, Semyon, 254
Krupa, Oleh, 167
Kubik-Zahorodna, Agnieszka, 209
Kubo, Masato, 55
Kucher, Natalie, 160
Kuderna, Lukas, 3, 161
Kuh, Diana, 279
Kuksenko, Olena, 16
Kumar, Arvind, 161
Kumar, Sushant, 148
Kumlehn, Jochen, 114
Kundaje, Anshul, 34, 89, 132, 162, 216

- Kundu, Soumya, 162
Kunisaki, Jason, 138, 163, 187,
 222
Kuo, Yi-Tzu, 114
Kursawe, Romy, 256
Kutalik, Zoltán, 70
Kutschat, Ana Patricia, 220
Kuzma, Kori, 67
Kuzmin, Ivan, 293
Kwok, Hui Si, 166
- Laakso, Markku, 23, 87
Lacy, Kip D., 164
Lage, Kasper, 210
Lai, Eric, 76
Laisk, Triin, 276
Laivuori, Hannele, 276
Lajoie, Bryan, 254
Lakhani, Chirag, 31
Lakka, Timo A., 23, 87
Lal, Dennis, 168
Lampela, Riina, 14
Lana, Susan E., 80
Landau, Dan A., 207, 244
Lander, Eric, 7, 35
Lane, Jacqueline M., 206
Lange, Frederik J., 50
Langmead, Ben, 60
Laplana, Marina, 83
Lappalainen, Tuuli, 39
Lariviere, Delphine, 165
Larson, Wesley, 184
Lato, Daniella F., 205
Lawlor, Deborah A., 276
Lawson, Heather, 139
Lazar, Alexander J., 84
Le, Victoria T., 297
Lea, Amanda J., 96, 124, 241,
 289
Leavesley, Matthew, 66
LeBaron von Baeyer, Sarah, 288
Leblanc, Sébastien, 299
Lee, Carol E., 65
Lee, Ceejay, 166
Lee, Charles, 155, 301
Lee, Christopher M., 201
Lee, Dong-Sung, 297
Lee, Jiseok Lee, 167
- Lee, ShinYu, 209
Lee, Sool, 167
Lee-Kim, Vivian S., 35
Legut, Mateusz, 298
Lehner, Ben, 36
Leibowitz, Mitchell L., 269
Lemieux, Faye A., 227
Lemke, Bianca, 13
Lemos, Diana, 64
Lennon, Christopher, 268
Lerga-Jaso, Jon, 83
Leslie, Christina, 12, 17, 132,
 295, 300
Leu, Costin, 168
Leverenz, James B., 141
Levine, Adam P., 88
Levine, Michael S., 73
Levy, Dan, 171
Li, Ang, 169
Li, Bo, 258
Li, Chang, 77
Li, Daniel, 162
Li, Daofeng, 139
Li, Ge, 177
Li, Heng, 186, 258, 269
Li, Houcheng, 273
Li, Jiangtao, 170
Li, Jian-Liang, 251
Li, Jingyi, 257
Li, Kui, 180
Li, Li, 13
Li, Mingjie, 139
Li, Ruoyan, 199
Li, Siran, 171
Li, Stacy, 6, 172, 184
Li, Taibo, 173, 202
Li, Xiaoyang, 178
Li, Xihao, 137
Li, Zhengke, 61
Li, Zhi, 4
Li, Zilin, 137
Liang, Heidi Y., 10, 230, 231
Liang, Qingnan, 174
Liao, Katherine P., 277
Liao, Wen-wei, 186
Liau, Brian B., 166
Libby, Peter, 302
Liberante, Fabio G., 156, 293

- Lieber, Andre, 77
Lim, Daven, 296
Lin, Chelsey, 57
Lin, Khine, 139
Lin, Meng, 175
Lin, Qiyao, 31
Lin, Shiqi, 176
Lin, Wenhe, 177, 203
Lin, Xihong, 19, 137, 284, 285
Lindblad-Toh, Kerstin, 161
Lindgren, Cecilia M., 50, 276
Lippincott, Margaret F., 50
Liscovitch-Brauer, Noa, 31
Lister, Ryan, 169
Littlejohn, Mathew D., 149
Liu, Andi, 178
Liu, Boxiang, 53, 129, 179
Liu, Cambrian Y., 308
Liu, Chunyu, 190
Liu, George E., 180
Liu, Haibo, 273
Liu, Lingjie, 181, 307
Liu, Molei, 277
Liu, Ruqian, 275
Liu, Siyuan, 63
Liu, Susanna, 290
Liu, Wenda, 182
Liu, X. Shirley, 258
Liu, Yang, 97
Llorens-Giralt, Palmira, 183
Logan, Emily, 135
Long, Quan, 217
Lou, Runyang Nicolas, 6, 184
Love, Michael, 190
Loving, Crystal L., 273
Lowe, Craig B., 42, 44, 113, 188
Loza, Martin, 185
Lu, Congyi, 298
Lu, Shuangjia, 186
Lu, Wenhan, 26
Luca, F., 228
Lucas, Julian K., 18
Luebbert, Laura, 24
Lulla, Suchita, 138, 163, 187,
 222
Lumby, Casper K., 98
Lund, Anina N., 38
Luo, Renhe, 295
Luo, Yanting, 42, 44
Luthra, Ishika, 105
Lutz, Sheila, 71
Lutzu, Stefano, 41
Lux, Thomas, 114
Lykoskoufis, Nikolaos, 229
Lynn, Claire, 242
Ma, Rosa X., 35
Ma, Xiaowen, 139
Maass, Philipp G., 205
Macnair, Will, 110
Macro, Jacob, 145
Maddock, Jane, 279
Madduri, Ravi, 53, 129, 277
Magenheim, Judith, 235
Mahdipoor, Parvin, 230
Maher, Matthew, 206
Mahmoud, Medhat, 61
Mairai, Tehani, 288
Mair-Meijers, H., 228
Makki, Nadja, 226
Makova, Kateryna D., 45, 271
Malik, Laksh, 125
Mallapragada, Saahithi, 275
Mallory, Benjamin J., 77
Mandla, Ravi, 137
Mangan, Riley J., 42, 188
Manickam, Nandini, 87, 304
Manning, Alisa K., 137
Mansour-Aly, Dina, 104
Marcotte, Samantha, 297
Marderstein, Andrew R., 189
Mardin, Balca, 211
Mares, Jason, 282
Margolis, Michael, 190
Mariani, Jessica, 212
Marinov, Georgi, 216
Marioni, John C., 27
Markovic, Christopher, 18
Marmor, Lindsay, 227
Marques-Bonet, Tomas, 3, 161
Marques-Coelho, Diego, 64
Marschall, Tobias, 18
Marson, Alexander, 292
Marth, Gabor, 75, 93, 246, 286
Martí-Gómez, Carlos, 37
Martin, Alicia, 26

- Martin, Fergal J., 64
Martin, Negin P., 251
Martins, Dino J., 124, 289
Mascher, Martin, 114
Mason, Christopher E., 254
Massilani, Diyendo, 191
Massip, Florian, 68
Matelska, Dorota, 122
Mateus, Jaime, 254
Mathieson, Iain, 1
Matthews, Elizabeth R., 287
Matthews, Jamie, 192
Matthews, Julia, 260
Mattioli, Kaia, 205
Mawlood, Shakhawan, 193
McAfee, Jessica C., 167
McArthur, Evonne, 272
McCandlish, David, 37
McCarthy, Davis, 28, 275
McCoy, Brianah M., 194
McGeary, Sean E., 13
McGrath, Barbara, 45
McGregor, Grant, 230
McInvale, Julie J., 306
McMahon, Francis J., 63
McMurray, Brandon J., 205
McNaughton, Lorna, 149
McRae, Jeremy, 3
McVicker, Graham, 297
Meadows, Kali, 145
Mee, Evan D., 275
Meireles de Sous, Ana, 16
Meistermann, Dimitri, 14
Meky, Eman M., 157
Melamud, Eugene, 223
Melé, Marta, 8, 205, 225, 232
Melnick, Ari, 17, 254
Melo, Diogo, 124
Mena, Daya, 210
Mendoza Revilla, Javier, 4
Meng, Guoliang, 205
Menon, Vilas, 91
Mercuri, Rafael L., 100, 121
Meredith, Melissa M., 125
Merritt, Jennifer, 41
Metspalu, Mait, 66
Metzger, Michael J., 136
Meuth, Nicholas A., 130
Meyer, Matthias, 191
Meyerson, Matthew, 269
Meyn, Stephen, 195
Miano, Charles, 124, 289
Michaels, Tai, 148
Michalettou, T D., 280
Miga, Karen H., 18, 57
Minaeva, Mariia, 39
Ming, Matthew J., 5, 90
Mishmar, Dan, 196
Mital, Seema, 205
Mitchell, Emily, 94
Mitchell, Matthew W., 6
Mitchell, Thomas J., 199
Mitra, Sneha, 12
Miyamoto, Tomomi, 209
Miyauchi, Kosuke, 55
Mo, Ziyi, 197
Mogre, Saie, 103
Mohanty, Chitrasen, 108
Mohlke, Karen L., 23, 87
Mokhtaridoost, Milad, 205
Mondal, Mayukh, 66
Monehan, Tracey, 149
Monroe, J. G., 267
Montague, Michael J., 96
Monteiro da Rocha, Andre, 23
Monteiro, Joao, 162
Montgomery, Brooke E., 262
Montgomery, Stephen B., 173, 189, 202
Montgomery, Taiowa A., 262
Montoya-Vazquez, Gloria, 226
Moore, Steven A., 251
Morales, Lian, 240
Moreira, Ricardo, 83
Moreno, Juan, 254
Morgan, Michael D., 27
Morini, Elisabetta, 135
Morishita, Shinichi, 245
Morley, David M., 97, 98
Morris, John A., 39
Morris, Zachary, 118
Mort, Richard, 149
Mortazavi, Ali, 10, 230, 231
Mosher, Stephen L., 198
Mosur, Varun, 43
Motiff, Heather, 195

- Mouratidis, Ioannis, 126
Mouri, Kousuke, 32
Moussion, Christine, 16
Movahed-Ezazi, Misha, 294
Mualim, Kristy, 132
Muhyoa, Benjamin, 124, 289
Mukamel, Eran A., 99
Muke, John, 66
Mukhopadhyay, Debabrata, 35
Mun, Taher, 60
Munson, Glen, 35
Munson, Katherine M., 77, 142
Murphy, Maddie, 7
Muyas, Francesc, 199
- Nacht, A Silvina, 183
Nair, Surag, 34
Najjar, Deena, 254
Najm, Imad, 168
Nakai, Kenta, 185
Nakamoto, Anne A., 200
Nalley, Matthew J., 40
Narasimhan, Vagheesh M., 282
Narisu, Narisu, 87
Narzisi, Giuseppe, 31
Nassar, Luis R., 201
Natarajan, Pradeep, 302
Nathan, Aparna, 274
Natri, Heini M., 28
Naumann, Ronald, 209
Navarro Gonzalez, Jairo, 201
Neale, Benjamin, 26
Neavin, Drew R., 281
Negrea, Aurel, 98
Negron-Del Valle, Josue E., 96, 241
Nehme, Ralda, 7
Nekrutenko, Anton, 165
Nellaker, Christoffer, 50
Newman, Barbara, 94
Newman, Deborah, 177, 203
Newman, Laura, 241
Nguyen, Ngoc Quynh H., 302
Ni, Bohan, 173, 202
Nichols, Thomas E., 50
Nie, Xichen, 138
Niepoth, Natalie, 41
Nirmalan, S., 228
- Nizhynska, Viktoria, 33
Noble, William S., 57, 300
Nordborg, Magnus, 33
Novak, Catherine, 157, 255
Noyes, Michelle D., 127
Ntasis, Vasilis F., 218
Nurk, Sergey, 45
Nürnberg, Peter, 168
Nurtdinov, Ramil, 132, 183, 218, 219
Nussbaum, Chad, 264
Nyasimi, Festus, 203
- Odom, Duncan T., 9
Oelen, Roy, 281
Oh, Julia, 140
Ohara, Daniel T., 100
Oliva, Meritxell, 204
Oliveros, Winona, 205, 225
Olivier, Michael, 177, 203
Ollila, Hanna M., 206
Olsen, Michelle, 170, 291
Olson, Nathan, 111, 309
Omans, Nathaniel D., 207
O'Neill, Amanda, 122
O'Neill, Mary, 4
O'Neill, Rachel, 45
Ong, Ken K., 279
Oppenheim, Sara J., 208
Orchard, Peter, 23
Org, Tõnis, 66
Ostrander, Elaine A., 80, 136, 250
Ostrovsky, Alex, 165
Ota, Mineto, 292
Ovcharenko, Ivan, 259
Overbey, Eliah G., 254
Owa, Chie, 245
Owens, Gregory, 184
Ozaki, Kokoro, 55
- Pääbo, Svante, 191, 209
Pachter, Lior, 24
Paik, Edward, 31
Páleníková, Petra, 210
Pallaseni, Ananth, 211, 293
Palmada-Flores, Marc, 165
Palmer, Alexandra, 98

- Palucka, Karolina, 140
Palumbo, Emilio, 218
Panda, Arjjit, 212
Pang, Baoxu, 213
Pang, Yakun, 244
Pankratov, Vasili, 66
Paquola, Apua, 74, 99
Pardoll, Drew M., 236
Parihar, Madhur, 237
Park, Jiyeon, 214
Park, Stella H., 243
Park, Yohan, 212
Park, Yongjin, 266
Parker, Anne, 64
Parker, Stephen, 23, 87, 304
Parrish, Phoebe C., 215
Parry, Aled, 98
Parts, Leopold, 156, 211, 293
Pasanen, Anu, 276
Pasaniuc, Bogdan, 190
Pascart, Tristan, 288
Paschou, Peristera, 252
Pask, Andrew J., 92
Patcher, Lior, 230
Patel, Aman, 216
Patel, Zain, 220
Paten, Benedict, 18, 161
Patin, Etienne, 4
Pavelec, Derek, 195
Pavez-Fox, Melissa A., 241
Pchelintsev, Nikolay, 97
Pearson, Ewan R., 104
Peddada, Teja N., 63
Pedersen, Anders Gorm, 119
Pedersen, Brent, 222
Pedersen, Stine F., 101
Peets, Elin Madli, 211, 293
Pelin Sümer, Arev, 191
Pellegrini, Matteo, 250
Peng, Julie, 124, 289
Pennacchio, Len, 157, 255
Penzel, Nora, 154, 237
Perera, Deshan, 217
Peretz, Ayelet, 235
Perez, Gerardo, 201
Perez-Arreola, Michelle, 294
Pérez-Lluch, Sílvia, 218, 219
Pérez-Schindler, Joaquín, 144
Pergola, Giulio, 154, 237
Persad, Gabrielle, 59
Person, E R., 280
Pertea, Mihaela, 253
Peter, Lance, 28
Peterson, Hedi, 293
Peterson, Rachel M., 96
Petrovski, Slavé, 122
Peyrégne, Stéphane, 191
Philippakis, Anthony, 20, 160,
 198
Phillippy, Adam M., 18, 45, 123
Phillips, Devan, 16, 96, 241
Pickard, Benjamin, 193
Pickard, Catriona, 193
Pickett, Brandon D., 45
Pidon, Hélène, 114
Piechota, Marcin, 78
Pieper, Andrew A., 141
Pinello, Luca, 220
Pintacuda, Greta, 210
Pique-Regi, R., 228
Pitman, Anders, 286
Plajzer-Frick, Ingrid, 157
Platt, Michael, 96
Plazier-Frick, Ingrid, 255
Ploubidis, George, 279
Pointon, Jonathan, 268
Pollard, Katherine, 190
Popli, Divyaratana, 191
Porubsky, David, 18, 133
Potapova, Tamara, 18, 45, 123
Potenski, Catherine, 244
Pott, Sebastian, 221, 308
Powell, Daniel L., 109
Powell, Joseph, 239
Powell, Simon, 287
Power, Christine, 279
Pozdeyev, Nikita, 175
Price, Alkes L., 132
Priebe, Oliver, 95
Prieto, Tamara, 207, 244
Prieto-Lafuente, Lidia, 97
Prigozhin, Daniil M., 200, 267,
 296
Pritchard, Jonathan K., 58, 238,
 292
Pritykin, Yuri, 12

- Przeworski, Molly F., 282
Przytycki, Paweł, 190
Puddu, Fabio, 97, 98
Puga Yung, Gisella, 229
Puig, Marta, 83
Puigdevall Costa, Pau, 14
Pulecio, Julian, 295
Pushkarev, Olga, 159
- Qiu, Jake, 207
Queitsch, Christine, 130
Quertermous, Thomas, 162, 206
Quillen, Ellen, 203
Quinlan, Aaron, 2, 127, 138, 163, 187, 222
Quintana-Murci, Lluís, 4
- Radecki, Alexander A., 43
Rafaels, Nicholas, 175
Rafi, Abdul Muntakim, 105
Raghavan, Sridharan, 175
Ragsdale, Aaron P., 65
Raio, Alessandra, 154
Raj, Anil, 223
Rajagopalan, Anugraha, 16
Rajagopalan, Srinivas, 244
Rajesh, Chandana, 224
Raman, Vatsan, 118
Ramirez, Jose Miguel, 8, 225
Ramisch, Anna, 229
Ramkhalawan, Darius, 226
Rampino, Antonio, 154
Ranchalis, Jane, 77
Rand, David M., 227
Raney, Brian J., 201
Ranjbaran, A., 228
Rao, Anjana, 97
Räsänen, Markus, 206
Rashid, Sabrina, 161
Rasmussen, Kasper, 119
Rathouz, Paul J., 282
Raudvere, Uku, 293
Raychaudhuri, Soumya, 239, 274
Raychowdhury, Raktima, 16
Raznahan, Armin, 63
Réal, Aline, 229
- Rebboah, Elisabeth, 10, 230, 231
Reese, Fairlie, 10, 230, 231
Regev, Aviv, 16
Rehm, Heidi, 26
Reik, Wolf, 98
Reilly, Steven, 25, 32
Reisenhofer, Elsa, 217
Relton, Caroline, 279
Rennie, Sarah, 112
Reshef, Yakir, 239
Resztak, Justyna A., 204, 228
Reverter, Ferran, 225
Reymond, Alexandre, 70
Rezaie, Narges, 10, 231
Rhie, Arang, 123
Ribéiro, Rogério, 8
Ricaut, François-Xavier, 66, 142
Ricciuti, Biagio, 284
Rich, Stephen, 173
Ried, Thomas, 61
Riehle, Kevin, 67
Ripoll-Cladellas, Aida, 225, 232
Roberts, Elizabeth T., 7
Rocha, Joana L., 6, 172
Rodriguez, Alex, 277
Rodriguez-Flores, Juan L., 233
Rodriguez-Fraticelli, Alejo, 29
Roeder, Kathryn, 152
Rogers, Jeffrey, 3, 161
Rogina, Blanka, 145
Rognon, Paul Joris, 225
Romero, Irene G., 142
Rosconi, Federico, 234
Rosenski, Jonathan, 235
Roskes, Jeffrey S., 236
Rossi, Fabiana, 237
Rothschild, Daphna, 238
Rotival, Maxime, 4
Rotter, Jerome, 173
Roucou, Xavier, 299
Rougvie, Ann E., 245
Roy, Michael C., 209
Rozenblatt-Rosen, Orit, 16
Rozowsky, Joel, 148
Rubin, Alan F., 67
Rubin, David T., 308

- Rubinstein, Boris, 123
Ruderfer, Douglas, 272
Rudolph, Stephanie, 41
Ruiz-Romero, Marina, 183, 218
Rumker, Laurie, 239, 274
Runz, Heiko, 281
Russeil, Julie, 159
Rutherford, Erica M., 240
Rutters, Femke, 104
- Sababi, Aiden M., 84
Sabeti, Pardis C., 32
Said, Abdelrahman, 205
Saito, Yoriko, 106
Sakaue, Saori, 239
Sakr, Jasmine, 10, 230
Salani, Monica, 135
Salaudeen, Asfar Lathif, 105
Salk, Jesse, 249
Salvatore, Marco, 112
Samocha, Kaitlin, 26, 85
Sanchez Rosado, Mitchell, 241
Sandelin, Albin, 101, 263
Sanjana, Neville E., 31, 298
Sansom, Helen, 97
Sanz, Maria, 218, 219
Saraiva-Agostinho, Nuno, 64
Sariol, Carlos, 241
Sarkar, Gobinda, 212
Sartori, Federica, 81
Sasani, Thomas A., 2
Sauer, Carolin M., 242
Savage, Michele, 54
Saxena, Richa, 206
Schartl, Manfred, 109
Schatz, Michael C., 20, 54, 102,
 160, 165, 198, 202, 262, 265
Scheben, Armin, 208
Scherer, Michael, 29
Schertzer, Megan D., 243
Schiffman, Joshua S., 244
Schmitt, Anthony, 294
Schnitzler, Gavin R., 35
Schraiber, Joshua G., 3
Schreiber, Jacob, 34
Schulte, Johannes H., 52
Schumer, Molly, 109
Schwarz, Erich M., 245
- Schwarz, Roland F., 52
Schwenk, J M., 280
Scott, Carly B., 90
Scott, Laura, 23, 87
Sealock, Julia, 26
Sederman, Casey, 246
Sedlazeck, Fritz J., 61, 111, 309
Seebach, Jörg D., 229
Seitz, Evan, 247
Seki, Masahide, 134
Sella, Guy, 58
Selvaggi, Pierluigi, 154
Seminara, Stephanie B., 50
Seong, Kyungyong, 158
Sepulveda, Hugo, 97
Serçin, Özdemirhan, 211
Serohijos, Adrian, 248
Serra Mari, Rebecca, 18
Serrano, Isabel M., 249
Serras, Florenci, 183
Serres-Armero, Aitor, 250
Seruggia, Davide, 220
Sethi, Anurag, 223
Sevim Bayrak, Cigdem, 150
Shah, Shrey, 63
Shan, Xinning, 22
Shankar, Krithika, 82
Shao, Xiang Qiang, 195
Sharma, Disha, 162
Sharma, S., 280
Shaw, Natalie D., 251
Sheinman, Michael, 68
Shekhar, Sudhanshu, 252
Shemer, Ruth, 235
Shemirani, Ruhollah, 21
Shen, Simon P., 166
Sher, Falak, 306
Sheth, Maya U., 132
Sheynkman, Gloria, 243
Shin, Joo Heon, 74, 237
Shinder, Ida, 253
Shiratori, Mari, 82, 278
Shoura, Massa J., 245
Shtolz, Noam, 196
Shuldiner, Alan R., 233
Shunkov, Michael V., 191
Sias-Garcia, Oscar, 35
Siegal, Mark, 81

- Sienkiewicz, Karolina, 254
Siepel, Adam, 181, 197, 307
Signor, Sarah, 76
Sikkink, Kristin, 294
Singer-Berk, Moriel, 3
Singh, Indranil, 29
Singhal, Ankita, 97
Sinnott-Armstrong, Nasa, 238
Siracusa, Erin R., 241
Siraj, Layla, 25
Skelton, Macaela, 241
Slatcher, R., 228
Slaugenhouette, Susan, 135
Slaven, Neil, 255
Slevin, Michael K., 269
Slickas, Beth, 194
Smith, Kellie, 236
Smith, Stephen, 50
Snuderl, Matija, 294
Snyder, Michael, 10, 213
Snyder-Mackler, Noah, 96, 194, 241
Sogin, Mitch, 308
Sokolowski, Eishani K., 256
Soldatkina, Oleksandra, 8, 225
Song, Dongyuan, 257
Song, Li, 258
Song, Susie, 7
Song, Wei, 259
Soos, Teresa, 83
Spannagl, Manuel, 114
Speelman, Pieter, 260
Spelman, Richard, 149
Spence, Jeffrey P., 238
Spencer, Rosie, 97
Sportelli, Leonardo, 154, 237
Sridharan, Samvardhini, 261
Stamatoyannopoulos, John, 132
Stankunaite, Reda, 242
Starostik, Margaret R., 262
Starrings, Marlena, 220
Steelman, Scott, 264
Stein, Jason, 190
Stein, Lincoln, 59, 79
Stein, Nils, 114
Steinmetz, Lars M., 132
Stence, Aaron A., 251
Stergachis, Andrew B., 77, 146
Sterner, Kirstin, 96
Stevens, Garrett, 79
Stigliani, Arnaud, 101, 263
Stitzel, Michael L., 256
Stitzel, Nathan O., 18
Stöber, Maja C., 52
Stockton, Joanne, 242
Stojanov, Petar, 264
Stolz, Joshua M., 74
Storer, Jessica M., 45
Stracquadanio, Giovanni, 268
Stratton, Michael, 46
Stringham, Heather M., 87
Strober, Benjamin, 202
Su, Anna, 148
Su, Chang, 22
Subramanian, Vidya, 7
Suderman, Keith, 54, 265
Suderman, Matthew, 279
Sudmant, Peter H., 6, 172, 184, 249, 261
Sullivan, Delaney, 24, 230
Sun, Jiawan, 103
Sun, Mohan, 30
Sun, Na, 266
Sundaram, Laksshman, 3, 161, 162
Sundares, Adithi, 14
Sung, Heejong, 63
Sunshine, Joel C., 236
Surapaneni, Likhitha, 64
Susanto, Teodorus T., 238
Sutherland, Chandler A., 267
Sutradhar, Anima, 268
Suva, Mario L., 244
Suvakov, Milovan, 212
Suzuki, Ayako, 134
Suzuki, Yutaka, 134, 151
Sykes, Nathan, 184
Symer, David E., 61
Szalay, Alexander S., 236
Szu-Tu, Chelsea, 29
Szymansky, Annabell, 52
Tabares, John A A., 149
Talkowski, Michael E., 135
Tan, Kar-Tong, 269
Tanigawa, Yosuke, 266

- Taube, Janis M., 236
Tay, Rebecca J., 262
Tay, Tristan, 166
Taylor, Chase, 28
Tegtmeyer, Matthew, 7
Teng, Jinyan, 273
Tervi, Anniina, 206
Tesio, Nicolò, 70
Tewhey, Ryan, 25, 32
Teysandier, Jean, 98
Thakore, Pratiksha, 16
Thapa, Kisan, 62
Thomas, James D., 215
Thompson, Mike, 192
Thybert, David, 64
Tian, Chi, 179
Timm, Laura, 184
Timp, Winston, 125
Timpson, Nicholas J., 276
Tingskov Pedersen, Casper Emil, 119
Tolkachov, Alexander, 9
Toneyan, Shushan, 270
Tong, Yihan, 179
Tonner, Peter, 111
Tonnies, Jackson, 130
Torres-Gonzalez, Edmundo, 271
Tovey, Nicholas, 242
Toyoshima, Yu, 245
Tralie, Christopher J., 296
Tran, Stella, 157
Treger, Taryn, 94
Trevanian, Stephen J., 64
Trevers, Katherine, 51
Trinh, Quang, 59
Tristani-Firouzi, Marti, 286
Trout, Diane, 10, 230, 231
Tsai, Ellen, 233, 281
Tsai, Pei-Chien, 279
Tsan, Yao-chang, 23
Tsiatsianis, Georgios G., 126
Tsuo, Kristin, 26
Tubbs, Colby, 272
Tuggle, Christopher K., 273
Tullius, Thomas W., 146
Tung, Jenny, 289
Tuomilehto, Jaakko, 23, 87
Tyndale, Selene, 297
Ucar, Duygu, 256
Udler, Miriam S., 144
Uhler, Caroline, 16
Ulirsch, Jacob, 25, 161
Ullrich, Sebastian, 219
Uminski, Michelle, 41
Ungar, Rachel, 173, 202
Uribe, Juber H., 273
Vaagenso, Christian, 112
Vaccarino, Flora M., 212
Vaddadi, Naga Sai Kavya, 60
Valencia, Cristian, 239, 274
Valentine, Clint, 249
Van Buren, Eric, 19
Van de Bunt, Martijn, 104
van der Wijst, Monique, 232, 281
van Heel, David, 276
Van Loo, Peter, 49
van Mierlo, Guido, 159
van Opijnen, Tim, 234
VandeBerg, John, 177, 203
Vandenbon, Alexis, 185
Vanderstichele, Thomas, 156
Vandomme, Audrey, 97, 98
Vannan, Annika, 275
Varshney, Arushi, 23, 87
Vazquez, Juan M., 6
Veiga, Raül G., 218
Vellarikkal, Shamsudheen K., 35
Velten, Lars, 29
Venkatesh, Samvida S., 276
Ventresca, Christa, 23
Verma, Anurag, 277
Vespasiani, Davide M., 92
Vicent, Guillermo P., 183
Vilella-Figuerola, Alba, 83
Vilgalys, Tauras P., 278
Villicaña, Sergio, 279
Viñuela, Ana, 104, 229, 280
Visel, Axel, 157, 255
Vitsios, Dimitrios, 122
Vo, Daniel, 190
Vochteloo, Martijn, 281
Voehringer, Harald, 86
Voight, Benjamin, 277
Voigtlaender, Rowina, 209
Voineagu, Irina, 169

- Völler, Mitchell R., 77
von Maydell, Kianna, 157
Võsa, Urmo, 281
Vromman, Amélie, 302
- Wacker, Sarah, 41
Wagner, Alex H., 67
Wagner, Justin, 111, 309
Wahl, Geoffrey M., 297
Walawalkar, Isha A., 43
Walker, Conor, 298
Walker, Mark, 104, 280
Wall, Jeff, 177, 203
Walsh, Thomas, 64
Wang, Jeffrey, 84
Wang, Joyce Y., 282
Wang, Kun, 143
Wang, Lily, 85
Wang, Meiyān, 99
Wang, Qingyang, 257
Wang, Quanli, 122
Wang, Shou-Wen, 13, 284
Wang, Shu, 207
Wang, Ting, 18, 139
Wang, Xian, 284, 285
Wang, Yichen, 46
Wangsa, Darawalee, 61
Ward, Alistair, 93, 286
Ward, Michelle C., 287
Warmerdam, Robert, 281
Wasik, Kaja A., 288
Watowich, Marina M., 96, 124, 241, 289
Wattenberg, Eve S., 290
Webb, Bryn D., 195
Weber, Christopher R., 308
Weedon, Michael N., 137
Wei, J., 228
Wei, Xiaoran, 291
Weimer, Annika K., 10
Weinberger, Daniel R., 74, 99, 237
Weinstock, Joshua S., 292
Weller, Julianne, 293
Wen, Cindy, 190
Wen, X., 228
Wessels, Hans-Hermann, 298
Westra, Harm-Jan, 281
- Whaling, Ian, 139
Wheeler, Matthew, 202
Wheeler, Vehia, 288
Whittaker, Allyson, 294
Wiebe, Victor, 209
Wigler, Michael, 171
Williams, Arianna L., 275
Williams, Brian, 10, 230, 231
Wilson, Michael, 98
Wirkler, Ivana, 9
Wirka, Robert, 162
Wittemans, Laura B.L., 276
Wittibschlager, Sandra, 220
Wittstruck, Nadine, 52
Wold, Barbara J., 10, 230, 231
Woldegebriel, Rosa, 14
Won, Hyejung, 167
Wong, Andrew, 279
Wong, Sandy, 30
Wong, Shuo, 189
Wong, Wilfred, 295
Wood, Andrew R., 137
Woodcock, Dan J., 153
Worth, Gemma, 149
Wray, Naomi R., 169
Wright, Caroline F., 137
Wright, Kevin, 223
Wrightsman, Travis, 130
Wu, Ting, 220
Wu, Yibing, 3
Wu, Yiming, 150
- Xi, Wang, 132
Xiao, Weihong, 61
Xie, Bingqing, 221, 308
Xiong, Kun, 148
Xiong, Xushen, 266
Xu, Boyan, 296
Xu, Hongxia, 30
Xu, Jinbo, 3
Xu, Jingwen, 308
Xu, Ke, 290
Xu, Zhichao, 297
Xu, Zichun, 22
Xue, Haoliang, 189
Xue, Xinhe, 298
- Yadav, Rachita, 135

- Yakymenko, Illya, 83
Yala, Feriel, 299
Yelloway, Gary, 97
Yamanaka, Yojiro, 149
Yan, Jielin, 295
Yang, Liu, 180
Yang, Rui, 300
Yang, Ruoyu, 140
Yang, Yanshen, 3
Yao, Fupan, 59
Yao, Jiayi, 101
Yasis, Jean, 297
Yates, Andrew, 64
Yazar, Seyhan, 239
Ye, Bin, 233
Yeates, Anna, 149
Yermakovich, Danat, 66
Yi Leung, Suet, 46
Yilmaz, Feyza, 301
Yin, Hongwei, 180
Yin, Melody, 89
Yonemitsu, Marisa A., 136
Yoo, DongAhn, 45
Yoon, Wan Hee, 212
Young, Sarah, 264
Yu, Bing, 302
Yu, Caroline, 31
Yu, Qi, 13
Yu, Shirong, 98
Yu, Steven, 224
Yu, Xuezhu, 290
Yu, Zhi, 302
Yuan, Long, 236
Yue, Feng, 139
Yung, Andrea, 16
Yunusov, Dinar, 303
- Zaitlen, Noah, 121, 192
Zajac, Cynthia K., 304
Zamanian, Jennifer L., 240
Zaurin, Roser, 29
Zeberg, Hugo, 56, 191, 305
Zeng, Jian, 169
Zeng, Lu, 306
Zeng, Tony, 35
Zhang, Jingfei, 22
Zhang, Nan, 295
Zhang, Pan, 190
- Zhang, Pengyue, 141
Zhang, Wenjin, 139
Zhang, Xiaolan, 7
Zhang, Yuntian, 179
Zhang, Zhe, 273
Zhang, Ziwei, 285
Zhao, Hongyu, 22
Zhao, Junhua, 254
Zhao, Nanxiang, 167
Zhao, Yixin, 181, 307
Zhao, Yu, 221, 308
Zhao, Zhongming, 178, 182
Zhou, Da, 143
Zhou, Ran, 221, 308
Zhou, Ronghao, 35
Zhou, Xingyan, 30
Zhou, Yiwen, 255
Zhu, Carrie, 5
Zhu, Qihui, 301
Zhu, Yiwen, 157
Zietz, Michael, 282
Zilioli, S., 228
Zinno, John, 207
Ziosi, Marcello, 39
Zook, Justin, 111, 309
Zou, James, 15

USING ANCIENT DNA TO DETECT AND UNDERSTAND RECENT NATURAL SELECTION IN HUMANS

Iain Mathieson

University of Pennsylvania, Philadelphia, PA

Detecting and understanding patterns of natural selection is one of the major goals of human evolutionary genomics. Historically, the most powerful approaches have been based on analysis of patterns of genetic variation in present day genomes. These approaches can detect natural selection over long but somewhat uncertain timescales. More recently, ancient DNA has enabled an orthogonal approach based on direct measurements of changes in allele frequency over relatively short but well-defined timescales. This allows precise estimates of the timing and nature of natural selection. In this study we focus on adaptation in Britain over the past 4500 years. We identify the strongest individual signals of selection and show that are all plausibly related to vitamin D or calcium metabolism.

Ancient DNA data can also be combined with information from present-day genome-wide association (GWAS) and expression quantitative trait loci (eQTL) studies to learn about the evolution of complex traits. Using the same British time transect, we identify several genes with rapid changes in predicted expression in the past 10,000 years, and predict changes in complex traits including skin pigmentation and stature. Because predicted genetic changes can be confounded by stratification or environmental effects, we validate predicted changes in stature by comparing to directly measured femur lengths. Finally, we test for evidence of selection on GWAS loci and find limited evidence of polygenic adaption on complex traits over this time period.

DISCOVERING EPISTASIS BETWEEN GERMLINE MUTATOR ALLELES IN MICE

Thomas A Sasani¹, Aaron R Quinlan^{1,2}, Kelley Harris³

¹Univ. of Utah, Dept. of Human Genetics, Salt Lake City, UT, ²Univ. of Utah, Dept. of Biomedical Informatics, Salt Lake City, UT, ³Univ. of Washington, Dept. of Genome Sciences, Seattle, WA

Maintaining genome integrity in the mammalian germline is essential and enormously complex. Hundreds of proteins comprise pathways involved in DNA replication, and hundreds more are mobilized to repair DNA damage. While loss-of-function mutations in any of the genes encoding these proteins might lead to elevated mutation rates, germline *mutator alleles* have largely eluded detection in mammals.

DNA replication and repair proteins often recognize particular sequence motifs or excise lesions at specific nucleotides. Thus, we might expect that the spectrum of *de novo* mutations -- the frequency of each individual mutation type (C>T, A>G, etc.) -- will differ between haplotypes that harbor either a mutator or wild-type allele at a given locus.

In 2022, we discovered a germline mutator allele in mice by analyzing whole-genome sequencing data from 152 recombinant inbred lines (RILs). These RILs, known as the BXDs, were derived from crosses of C57BL/6J and DBA/2J, two laboratory strains that exhibit significant differences in their germline mutation spectra. The BXD RILs were inbred for up to 180 generations, and each line therefore accumulated up to 2,000 germline *de novo* mutations. Using quantitative trait locus (QTL) mapping, we identified a locus on chromosome 4 that was strongly associated with the C>A germline mutation rate. This QTL overlapped *Mutyh*, which encodes a protein that normally prevents C>A mutations by repairing oxidative DNA damage.

In this study, we developed a new method to detect alleles that affect the mutation spectrum in biparental RILs. At every informative marker along the genome, we compute the aggregate *de novo* mutation spectrum in lines that inherited either parental allele. We then calculate the cosine distance between those aggregate spectra and identify loci at which that distance is greater than what we'd expect by random chance. By applying this method to mutation data from the BXDs, we discovered an additional C>A germline mutator locus that overlaps *Ogg1*, a key partner of *Mutyh* in base-excision repair of oxidative DNA damage.

Strikingly, BXDs with the mutator allele near *Ogg1* do not exhibit elevated rates of C>A germline mutation unless they also possess the mutator allele near *Mutyh*. However, BXDs with both alleles exhibit even higher C>A mutation rates than those with either one alone. To our knowledge, these new methods for analyzing mutation spectra reveal the first evidence of epistasis between mammalian germline mutator alleles, and may be applicable to mutation data from humans and other model organisms.

THE LANDSCAPE OF TOLERATED GENETIC VARIATION IN HUMANS AND PRIMATES

Hong Gao¹, Tobias Hamp¹, Jeffrey Ede¹, Joshua G Schraiber¹, Jeremy McRae¹, Moriel Singer-Berk², Yanshen Yang¹, Anastasia Dietrich¹, Petko Fiziev¹, Lukas Kuderna¹, Laksshman Sundaram¹, Yibing Wu¹, Aashish Adhikari¹, Yair Field¹, Jinbo Xu¹, Jeffrey Rogers³, Tomas Marques-Bonet⁴, Kyle Farh¹

¹Illumina, Inc., Genome Interpretation, Foster City, CA, ²Broad Institute of MIT and Harvard, Medical and Population Genetics, Boston, MA, ³Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX,

⁴Institute of Evolutionary Biology, PRBB, Barcelona, Spain

Personalized genome sequencing has revealed millions of genetic differences between individuals, but our understanding of their clinical relevance remains largely incomplete. To systematically decipher the effects of human genetic variants, we previously developed our variant pathogenicity predictor PrimateAI and achieved superior performance compared to all other classifiers. Nonetheless, earlier work was limited by the very small primate population sequencing datasets available. To expand upon this strategy, we obtained whole genome sequencing data for 809 individuals from 233 primate species, and identified 4.3 million common protein-altering variants with orthologs in human. We show that these variants can be inferred to have non-deleterious effects in human based on their presence at high allele frequencies in other primate populations. We use this resource to classify 6% of all possible human protein-altering variants as likely benign and impute the pathogenicity of the remaining 94% of variants with our deep learning method PrimateAI-3D. PrimateAI-3D is a semi-supervised 3-D convolutional neural network for variant pathogenicity prediction, which we trained using 4.5 million primate common missense variants with likely benign consequence. In a departure from prior deep learning architectures that operated on linear sequence, we voxelized the 3D structure of the protein at 2 Angstrom resolution and used 3D-convolutions to enable the network to recognize key structural regions that may not be apparent from sequence alone. We also leveraged language models of protein sequences and variational autoencoder on multiple sequence alignments by incorporating these models as additional loss functions. We evaluated PrimateAI-3D and 15 other published machine learning methods on their ability to distinguish between benign and pathogenic variants along six different axes, including deep mutational scan experimental assays across 9 genes and UK biobank clinical phenotypes, and demonstrated PrimateAI-3D achieves state-of-the-art accuracy for diagnosing pathogenic variants in patients with genetic diseases.

GENETIC AND EVOLUTIONARY BASIS OF POPULATION DIFFERENCES IN IMMUNE RESPONSE TO RESPIRATORY VIRUSES

Maxime Rotival¹, Yann Aquino^{1,2}, Aurélie Bisiaux¹, Zhi Li¹, Mary O'Neill¹, Javier Mendoza Revilla¹, Etienne Patin¹, Lluis Quintana-Murci^{1,3}

¹Human Evolutionary Genetics Unit, Institut Pasteur, Université Paris Cité, CNRS UMR2000, Paris, France, ²Sorbonne Université, Collège Doctoral, Paris, France, ³Human Genomics and Evolution, Collège de France, Paris, France

Humans display vast clinical variability upon infection by respiratory viruses, partly due to genetic factors. However, how natural selection and archaic introgression have contributed to current population differences in immune responses to respiratory viruses remains poorly appreciated. Here, we characterized the genetic basis of population variation in immune response to viruses by performing single cell RNA-sequencing on peripheral blood mononuclear cells from 222 healthy donors of African (n=80), European (n=80) and East-Asian (n=62) ancestries, stimulated with either SARS-CoV-2 or influenza A virus (IAV). Through the mapping of expression quantitative trait loci (eQTLs) across 22 different cell types and 5 immune lineages, we detected 12,753 cis-eQTL and 1,505 response eQTLs, and revealed cell-type- and virus-specific genetic control of immune responses. Using mediation analyses, we estimate that cellular heterogeneity, mostly environmentally induced, contributes to 16-62% of population differences in gene expression depending on the immune lineage considered. Conversely, common genetic variants displayed a weaker overall effect affecting 13-35% of differentially expressed genes across populations, yet they account for up to 58% of population differences among genes with an eQTL. Focusing on the impact of natural selection on the differentiation of immune response across populations, we found evidence of recurrent selection targeting effectors of the type I interferon response such as *IFITM2-3*, *IFIT5* or *ISG20*, as well as signals of polygenic adaptation ~25 kY ago targeting variants associated with SARS-CoV-2-specific responses in East Asians. Furthermore, our analyses revealed the cell-type-specific effects of Neanderthal introgression on immune functions and showed that introgressed regulatory alleles of myeloid cells were adaptive in Europeans after their split from East Asians. Finally, using colocalization analyses and TWAS, we report loci such as *DR1* and *OAS1*, where adaptive evolution targeting regulators of antiviral immunity has contributed to current disparities in COVID-19 risk. Collectively, these findings highlight the role of past natural selection in the population differentiation of immune responses to respiratory viruses and reveal the cellular and molecular mechanisms through which Neanderthal introgression has altered current immune functions.

SEX DIFFERENCES IN GENETIC EFFECTS ON COMPLEX TRAITS

Carrie Zhu^{1,2}, Matthew J Ming^{1,2}, Jared M Cole^{1,2}, Michael D Edge³, Mark Kirkpatrick², Arbel Harpak^{1,2}

¹The University of Texas at Austin, Department of Population Health, Austin, TX, ²The University of Texas at Austin, Department of Integrative Biology, Austin, TX, ³University of Southern California, Department of Quantitative and Computational Biology, Los Angeles, CA

Sexual dimorphism in complex traits is suspected to be in part due to widespread gene-by-sex interactions (GxSex), but empirical evidence has been elusive. Here, we infer the mixture of ways polygenic effects on physiological traits covary between males and females. Incorporating polygenic GxSex in polygenic score prediction significantly improves predictive performance in ~70% of traits considered. A key observation is that GxSex is pervasive but acts primarily through systematic sex differences in the magnitude of many genetic effects (“amplification”), rather than in the identity of causal variants. Amplification patterns account for sex differences in trait variance. In some cases, testosterone may mediate amplification. Finally, we develop a population-genetic test linking GxSex to contemporary natural selection and find evidence for sexually antagonistic selection on variants affecting testosterone levels. Taken together, our results suggest that the amplification of polygenic effects is a common mode of GxSex that may contribute to sex differences and fuel their evolution.

A PAN-PANGENOME CAPTURES THE FULL SPECTRUM OF
GENETIC VARIATION AND ANCIENT TRANS-SPECIES
STRUCTURAL POLYMORPHISM IN HUMANS, CHIMPANZEES AND
BONOBOS

Joana L Rocha¹, Juan M Vazquez¹, Alison Killilea², Runyang Nicolas Lou¹,
Stacy Li¹, Matthew W Mitchell³, Kendra Hoekzema⁴, Evan E Eichler⁵,
Peter H Sudmant¹

¹UC Berkeley, Department of Integrative Biology, Berkeley, CA, ²UC Berkeley, Department of Molecular and Cell Biology, Berkeley, CA, ³Coriell Institute for Medical Research, Camden, NJ, ⁴University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, ⁵University of Washington, Howard Hughes Medical Institute, Seattle, WA

Chimpanzees and bonobos (genus Pan) are our species closest living relatives, sharing more than 98% of our genetic makeup despite millions of years of divergence. The genetic background of these species provides evolutionary context for capturing and representing the extensive structural diversity of human haplotypes. However, while long-read sequencing consortium projects such as the T2T and HPRC are providing an unprecedented opportunity to decipher some of the most intractable and complex regions of the human genome, we have yet to fully understand the extent of genome structural diversity in the Pan genus. Here, we have assembled 17 reference-quality genomes (34 haplotypes) from a comprehensive sampling of a diverse set of 13 chimpanzees and 4 bonobos throughout their geographic range in Africa, to create a graph-based, telomere-to-telomere representation of Pan genomic diversity both within and between species. This Pan-pangenome effort comprises some of the most contiguous haplotype-resolved assemblies to-date for non-human primates at a population scale (averaging ~50x coverage, 2e7-10e7bp NG50). Our work complements and extends the recent human focussed efforts to obtain a near-complete representation of the full spectrum of global genetic variation. Using these resources, we are characterizing the structure, composition, function, and evolutionary trajectories of genomic variation that is shared between humans, chimpanzees and bonobos. This will allow us to interrogate the remarkable legacy of ancestral trans-species polymorphisms across the primate lineage, and test hypotheses on the adaptive significance of complex genomic structures that give way to shared genetic diversity.

SINGLE-CELL RNA-SEQ IN CELL VILLAGES ENABLES DISCOVERY OF STATE-DEPENDENT eQTLs

Elizabeth T Roberts^{*1}, Susie Song^{*1}, Matthew Tegtmeyer², Tommy Casolaro¹, Maddie Murphy¹, Xiaolan Zhang¹, Ralda Nehme², Eric Lander¹, Vidya Subramanian¹, Elisa Donnard^{*1}, Thouis R Jones^{*1}

¹Broad Institute of MIT and Harvard, Lander Lab, Cambridge, MA, ²Broad Institute of MIT and Harvard, Stanley Center for Psychiatric Research, Cambridge, MA

Over the past decades, genome-wide association studies have identified thousands of trait-associated genomic variants, largely within regulatory regions. However, only a small fraction of these have been functionally characterized, in part due to their cell type specific nature. To understand gene regulation, it is critical to determine how specific variants affect cellular programs and contribute to disease. Several questions remain unanswered in the field such as what cell types, stages of development or cellular contexts are impacted? What transcription factors and cellular pathways are involved? Answering these questions will guide clinical interpretation of known and new variants, and ultimately, how to treat and prevent diseases. To address this we applied single-cell RNA-sequencing (scRNA-seq) to multi-donor pools, or "cell villages", which enabled large-scale discovery of expression quantitative trait loci (eQTL) in multiple cell states. We differentiated human induced pluripotent stem cells (hiPSCs) from 273 donors into hepatic and neuronal precursors, and collected cells at several time points during differentiation. At each timepoint we detect on average 2,200 eGenes, out of which roughly 900 are universal and discovered in all cell states. From the 2,774 eGenes discovered in the undifferentiated iPSC state, roughly 200 show a complete reversal in eQTL direction at the hepatic day 7 time point, further highlighting the importance of evaluating variant effects at specific cellular contexts. To test whether disease heritability is enriched in eGenes specific to each point in our differentiation, we applied linkage disequilibrium (LD) score regression. We detect significant enrichments for disease relevant traits such as elevated levels of aspartate aminotransferase (AST) in day 7 hepatic differentiation over day 10 neuronal differentiation. Conversely, eGenes from day 10 neuronal differentiation show a significant enrichment for Schizophrenia (SCZ) over day 7 hepatic differentiation. This large scale dataset is not only critical for the identification of causal variants and their impact but also for developing and testing generalizable models of gene regulation which can then be applied to novel unprofiled cell states.

TRANSCRIPTIONAL AND EPIGENETIC IMPACT OF CIGARETTE SMOKING ACROSS HUMAN TISSUES

Jose Miguel Ramirez¹, Rogério Ribéiro^{2,3}, Oleksandra Soldatkina¹, Raquel Garcia¹, Pedro G Ferreira^{2,3}, Marta Mele¹

¹Barcelona Supercomputing Center, Life Sciences Department, Barcelona, Spain, ²University of Porto, Department of Computer Science, Porto, Portugal, ³INESC TEC, Laboratory of Artificial Intelligence and Decision Support, Porto, Portugal

Tobacco smoke is the main cause of preventable premature mortality worldwide. Smoking accelerates tissue aging and increases disease predisposition to many diseases including cancer. Yet, our understanding of the molecular mechanisms driving smoking-related health decline and tissue degeneration remain extremely limited. Here we characterize transcriptomic, epigenetic and tissue architectural alterations induced by cigarette smoking across human tissues combining gene expression, DNA methylation and histology image analysis from the Genotype-Tissue Expression Dataset. Differential expression analysis shows a widespread impact of smoking in the human body, with some genes being systematically upregulated across tissues. Notably, we observed concordant additive effects of smoking and aging in most tissues suggesting that smoking can have similar consequences than those of biological aging. The largest expression changes occur in the lung and are related to inflammation and macrophage infiltration. After lung, thyroid is the tissue most affected by cigarette smoking. We use convolutional neural networks on lung and thyroid histological images and confirm important architectural differences between never smokers and smokers. In thyroid, smoking-associated changes can be linked to the presence of larger follicles, a possible explanation to the observed association between tobacco smoking and hyperthyroidism. In general, gene expression patterns are reversible with ex-smokers exhibiting a profile between smokers and never smokers but generally closer to never smokers. In contrast, methylation effects in the lung, although reversible, are more persistent than gene expression effects, with most methylation changes at an intermediate state but closer to smokers. We do not observe large inter-tissue differences in reversibility rates but genes differ in their capacity to return to basal expression levels consistently across tissues. Finally, we observe overall low correlation between DNA methylation and expression although some loci located nearby immune response genes are highly correlated. Overall, our multi-tissue analysis of the effects of cigarette smoking provides an extensive characterization of the impact of tobacco smoke across tissues and will help gaining insights on the molecular changes driving smoking-driven tissue homeostasis decline.

THE FUNCTION AND DECLINE OF THE FEMALE REPRODUCTIVE TRACT AT SINGLE-CELL RESOLUTION

Ivana Winkler^{*1}, Alexander Tolkachov^{*2}, Duncan T Odom^{#2}, Angela Goncalves^{#1}

¹German Cancer Center, Somatic Evolution and Early Detection, Heidelberg, Germany, ²German Cancer Center, Regulatory Genomics and Cancer Evolution, Heidelberg, Germany

The study of female-specific biology has long been disregarded and discounted. Here I will present the large-scale generation and analysis of single-cell and spatial transcriptomics data from 50 cell types from 5 reproductive organs, across 5 phases of the estrous cycle and during pregnancy in young mice, and at 6 ageing time-points of the female reproductive tract (FRT). Our data represent a comprehensive atlas of the FRT over lifespan, and reveal pathological consequences of incomplete resolution of recurrent inflammation and tissue repair.

The FRT undergoes extensive remodelling during each reproductive cycle, regulated by systemic changes in sex hormones. Whether this recurrent remodelling influences a specific organ's aging trajectory is unknown. To address this, we systematically characterised at single-cell resolution the morphological and transcriptional changes that occur in ovary, oviduct, uterus, cervix, and vagina at each phase of the mouse estrous cycle, during decidualization, and into aging. Our analyses newly reveal how morphological differences between the upper and lower reproductive tract are closely mirrored by compositional and transcriptional differences. To explore whether the cyclic inflammation and remodelling that naturally occur during the reproductive lifespan of young mice result in age-related chronic inflammation and fibrosis, we extensively characterised the inflammatory status of fibroblasts and their cell-to-cell communication networks during normal cycling and aging. We determined that transcription factor and cell-to-cell communication networks active in fibroblasts during estrous cycling and aging are enriched for ECM remodelling and inflammation, and are conserved between humans and mouse uteruses. Our work directly links intensity of inflammation and ECM activity during the estrous cycle with the severity of chronic inflammation and fibrosis in old age. Our data supports a model wherein the incomplete resolution of inflammation and ECM remodelling during an increasing number of cycles leads to gradual development of fibrosis and chronic inflammation, predisposing organs to disease development. Using a mouse model of premature menopause, we decouple the effect of aging and cycling on fibrosis development and confirm the contribution from cycling. Our single-cell atlasing efforts implicate fibroblasts in maintaining a "memory" of past inflammation and propose a unifying mechanism for the epidemiological association between a variety of disparate cycle-number modifying factors and endometrial cancer risk in humans.

* authors contributed equally

corresponding authors

GENOMIC REGULATORY STRUCTURE OF THE ENCODE4 MOUSE POSTNATAL DEVELOPMENTAL TIME COURSE AT SINGLE-CELL RESOLUTION REVEALS HOMOLOGOUS REGULATORY TOPICS WITHIN AND ACROSS TISSUES

Elisabeth Rebboah^{1,2}, Narges Rezaie^{1,2}, Brian Williams³, Annika K Weimer⁴, Heidi Y Liang¹, Diane Trout³, Fairlie Reese^{1,2}, Jasmine Sakr², Michael Snyder⁴, Barbara Wold³, Ali Mortazavi^{1,2}

¹UC Irvine, Developmental and Cell Biology, Irvine, CA, ²UC Irvine, Center for Complex Biological Systems, Irvine, CA, ³Caltech, Biology, Pasadena, CA, ⁴Stanford, Genetics, Stanford, CA

Genomic regulation after birth contributes significantly to tissue and organ maturation, but is under-studied relative to existing genomic catalogs of prenatal development in mouse. As part of the final phase of the ENCODE consortium, we generated the first comprehensive single-nucleus atlas of postnatal development across seven postnatal time points in adrenal glands, heart, skeletal muscle, cerebral cortex, and hippocampus using a combination of single-nucleus RNA-seq using Split-seq from Parse Biosciences and 10X Multiome in matching samples. We identified at least one cell type in each of these tissues that shows changes either before or as the result of puberty.

We use a robust form of Latent Dirichlet Allocation (LDA) with a specific vocabulary of regulatory genes in Parse and 10X nuclei from each tissue to identify 111 significant topics in 420,376 nuclei. We find that this regulatory vocabulary is sufficient to recover manually annotated cell types in both platforms and that topics show further structure within cells, with the majority of cells participating in 2 to 3 major topics simultaneously. Using manual annotations in parallel, characterization of these topics reveals that they reflect cell type as well as cell state depending on the prevalence of these cells in the tissue. We find that transcription factors (TFs) and microRNA host genes are the most specific genes in topics followed by genes involved in chromatin binding and chromatin organization, which dominate a few topics related to global state changes such as cell cycle. Comparison of the topics learnt independently across the different tissues reveals homologous groups of topics for multiple cell types in two or more tissues, suggesting they share a common regulatory core. Motif analysis of the ATAC multiome data in cells based on RNA topics revealed enrichment of key topic TFs in sex specific adrenal cortical cell types. Using a restricted regulatory vocabulary, LDA is able to recover meaningful topics that are fingerprints of concurrent GRNs running within a cell, thus allowing a systematic characterization of cell types and states within and across tissues.

NO CELL LEFT BEHIND: DYNAMIC STUDIES OF GENE REGULATION IN HUMANS

Yoav Gilad

University of Chicago, Medicine, Chicago, IL

To effectively prevent and treat diseases, we need to understand how genes are dysregulated within disease-relevant contexts. One successful approach to do so is to investigate associations between genetic variation and regulatory phenotypes, as in expression quantitative trait locus (eQTL) analysis. Studies in humans have identified tens of thousands of eQTLs across different populations and, more recently, in multiple tissues. However, many disease-associated genes are not colocalized with known eQTLs and there is a great need for new approaches to characterize regulatory QTL mechanisms in diverse, disease-relevant cell types from the same genotype.

Existing relevant efforts, such as the ENCODE project, include only a few samples per cell type and do not enable direct investigation of genetic effects. Conversely, the GTEx project and other large-scale transcriptome and genotype datasets covering multiple tissues, currently provide only gene expression data and are not amenable to ongoing deep and dynamic molecular phenotyping in the same cohort, as the post-mortem frozen tissue samples are not renewable and are a challenging template for most functional genomic protocols.

To address this challenge, we established a new approach to differentiate iPSCs into dozens of different cell types. We call this approach ‘guided differentiation’, to contrast it with the more standard suite of ‘directed differentiation’ approaches. In guided differentiation, our goal is to establish cultures of differentiated cells that include multiple cell types rather than a single terminal cell type (as is the goal of standard directed differentiation). We then use single cell RNA sequencing to identify the cell types in each culture and characterize cell-type specific gene expression within and across individuals and species. For example, our cardiac lineage guided differentiation cultures include multiple related cell types, not just cardiomyocytes; our mesenchymal derived cultures include adipocytes, osteocytes and other related bone cells, and chondrocytes; our neuronal guided differentiated culture include more than a dozen different neuronal cell types.

We used our guided differentiation cultures to study how human gene regulation varies across different genetic backgrounds, cell types, and environmental exposures. By applying single-cell sequencing to the heterogenous differentiated cultures, we explored how a variety of relevant cell types respond to different external exposures that are known to affect function and/or disease risk. With these data, we mapped thousands of dynamic eQTLs and gene by environment interactions in more than a hundred cell types, and in response to a dozen exposure. We used our findings to interpret results from genome-wide association studies.

REGRESSION MODELING OF MULTIOME DATA IDENTIFIES FUNCTIONAL ENHancers AND ENABLES CHROMATIN POTENTIAL ANALYSIS

Sneha Mitra¹, Yuri Pritykin², Christina Leslie³

¹Research Scholar, Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, ²Assistant Professor, Lewis-Sigler Institute for Integrative Genomics and Computer Science Department, Princeton University, Princeton, NY, ³Member, Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY

Multiome single cell sequencing of chromatin accessibility and gene expression is where both scATAC-seq and scRNA-seq are read out from the same individual cells. This has paved the way for novel computational methods that attempt to link enhancers to genes. These approaches first identify a set of peaks from the scATAC-seq data and then for each individual peak, find genes such that the accessibility within the peak is strongly correlated with gene expression across all cells in the multiomic data set. While the approach can provide key insights into the gene-specific regulatory regions of the genome, it is limited by the peak set being used. Typical peak-calling approaches on scATAC-seq may miss peaks associated with rarer cell types. Therefore, the peak-to-gene linkage analysis using the peak set would provide little information on such cell types. To address this, we present SCARlink (Single-cell ATAC RNA linking), a novel gene-level regulatory model to link gene expression to tile-based chromatin accessibility within and flanking (+/- 250kb) the genic locus in single cell multiome (scRNA-seq and scATAC-seq co-assay) sequencing data. The approach uses regularized Poisson regression on tile-level accessibility data to jointly model all regulatory effects at the gene locus, avoiding the limitations of pairwise gene-peak correlations and dependence on a peak atlas. We find that Shapley value analysis on the trained models identifies cell-type-specific gene enhancers that are validated by promoter capture Hi-C. Using the Shapley values we show that the tissue-based fine-mapped eQTLs are cell type specific. We further show that SCARlink-predicted and observed gene expression vectors provide a robust way to compute a chromatin potential vector field to enable developmental trajectory analysis.

A MOUSE MODEL WITH HIGH CLONAL BARCODE DIVERSITY FOR JOINT LINEAGE, TRANSCRIPTOMIC, AND EPIGENOMIC PROFILING IN SINGLE CELLS

Li Li^{1,2}, Sarah Bowling^{1,2}, Qi Yu^{1,2}, Sean E McGeary³, Karel Alcedo^{1,2},
Bianca Lemke^{1,2}, Mark Ferreira^{1,2}, Allon M Klein³, Shou-Wen Wang⁴,
Fernando D Camargo^{1,2}

¹Boston Children's Hospital, Stem Cell Program, Boston, MA, ²Harvard University, Department of Stem Cell and Regenerative Biology, Boston, MA, ³Harvard Medical School, Department of Systems Biology, Blavatnik Institute, Boston, MA, ⁴Westlake University, School of Life Sciences, Hangzhou, China

Cellular lineage histories along with their molecular states encode fundamental principles of tissue development and homeostasis. Current lineage-recording mouse models have limited barcode diversity and poor single-cell lineage coverage, thus precluding their use in tissues composed of millions of cells. Here, we developed DARLIN, an improved Cas9 barcoding mouse line that utilizes terminal deoxynucleotidyl transferase (TdT) to enhance insertion events over 30 CRISPR target sites, stably integrated into 3 distinct genomic loci. DARLIN is inducible, has an estimated $\sim 10^{18}$ lineage barcodes across tissues, and enables detection of usable barcodes in ~60% of profiled single cells. Using DARLIN, we examined fate priming within developing hematopoietic stem cells (HSCs) and revealed unique features of HSC migration. Additionally, we adapted a method to jointly profile DNA methylation, chromatin accessibility, gene expression, and lineage information in single cells. DARLIN will enable widespread high-resolution study of lineage relationships and their molecular signatures in diverse tissues and physiological contexts.

SINGLE CELL ATLAS OF CORTICAL NEURON DEVELOPMENT *IN VITRO*

Adithi Sundares¹, Dimitri Meistermann¹, Riina Lampela¹, Rosa Woldegebriel¹, Pau Puigdevall Costa^{1,2}, Helena Kilpinen^{1,3}

¹University of Helsinki, Helsinki Institute of Life Science, Helsinki, Finland, ²University College London, Great Ormond Street Institute of Child Health, London, United Kingdom, ³University of Helsinki, Faculty of Biological and Environmental Sciences, Helsinki, Finland

Induced pluripotent stem cells (iPSCs) are an important *in vitro* model of human disease and development and present the advantage of studying donor-specific data in the relevant cellular context. However, remaining challenges of using iPSCs for disease modelling include characterizing and benchmarking the diversity of cell types differentiated *in vitro*. We differentiated iPSC lines from multiple donors to cortical neuron lineage using an established protocol, and, by profiling the transcriptomes of 60,000 single cells across three timepoints spanning 70 days, produced a single-cell atlas of cortical neuron development *in vitro*. We benchmarked our dataset against publicly available fetal and organoid references to assess how well *in vitro* differentiation recapitulates neural development *in vivo*. We tested various approaches to cell type annotation, highlighting how differing criteria in cell type definitions can influence biological findings. Overall, we found that over 60% of the cells mapped with high confidence to primary cell types, while an additional 30% aligned to broader cell types found in organoids than to any individual neuronal subtype. Assessment of the transcriptomic differences driving subtype resolution (or lack thereof) pointed to various metabolic processes such as oxidative phosphorylation or cell cycle effects. To complement this, we quantified morphological phenotypes of the differentiating cells across the same timepoints using Cell Painting, a high-content imaging assay. The imaging channels, which correspond to cellular components, were analysed to produce a matrix of image features per cell. We studied cell-type and donor-specific effects across these two modalities at single cell resolution by modelling the relationship between image features and gene expression. This model allowed us to annotate Cell Painting features with likely associated biological pathways, facilitating functional interpretation of potential donor-specific effects. The human-specific nature of the model enabled studying features distinctive to human development, such as the production of outer radial glia and cortical interneurons. Preliminary results highlight several mitochondrial processes enriched in inhibitory interneurons as compared to excitatory neurons across both modalities. Overall, this study represents a comprehensive molecular characterization of cortical neuron development *in vitro* that provides an important baseline for disease modelling.

DISCOVERING DISEASE-RELEVANT SPATIAL CELLULAR MOTIFS WITH GRAPH DEEP LEARNING ON SPATIAL OMICS.

James Zou

Stanford University, Stanford, CA

The recent development of highly multiplexed immunofluorescence imaging, such as CODEX, allows the rich characterization of protein abundances at subcellular resolution from patient-derived tissues. However, modeling and extracting biological insights from such rich spatial data is an open challenge. We developed SPAatial CEllular Graphical Modeling (SPACE-GM), a graph deep learning approach that models cellular neighborhoods as graphs in which the nodes are individual cells and edges encode physical proximity. Through training SPACE-GM to predict the prognosis and treatment response of patients, the model identified subgraphs that captured disease-relevant cellular structures. We apply SPAGE-GM to several diseases, including 658 head-and-neck and colorectal cancer tissue samples, to discover spatial cellular motifs that predict patient response to cancer treatments. Analysis of these motifs reveal biological insights into tumor-immune interactions that could affect patient outcomes. Through these studies, we will also present a flexible and general computational framework for analyzing spatial omics data.

PERTURBDECODE, A PROBABILISTIC ANALYSIS FRAMEWORK TO RECOVER REGULATORY CIRCUITS AND PREDICT GENETIC INTERACTIONS FROM LARGE-SCALE PERTURB-SEQ SCREENS

Basak Eraslan^{1,2}, Kathryn Geiger-Schuller^{1,2}, Olena Kuksenko², Pratiksha Thakore^{1,2}, Ozge Karayel¹, Andrea Yung¹, Anugraha Rajagopalan¹, Ana Meireles de Sous¹, Karren Dai Yang³, Liat Amir-Zilberstein², Toni Delorey², Devan Phillips², Rakimra Raychowdhury², Christine Moussion¹, Nir Hacohen², Caroline Uhler³, Orit Rozenblatt-Rosen^{1,2}, Aviv Regev^{1,2,4}

¹Genentech, Genentech Research and Early Development, South San Francisco, CA, ²Broad Institute of MIT and Harvard, Klarman Cell Observatory, Cambridge, MA, ³Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, ⁴Massachusetts Institute of Technology, Department of Biology, Cambridge, MA

Pooled genetic perturbation screens with single cell RNA-seq readouts (Perturb-Seq) open up new avenues in dissecting the function of genes and deciphering the gene regulatory networks. However, it is challenging to analyze large-scale screens with thousands of perturbed genes and millions of profiled cells due to noise from varying degrees of efficiency of the CRISPR-based perturbations, the sheer scale of the data, and the need to predict effects of combinations of perturbations that have not been observed experimentally. Here, we present PerturbDecode, a framework for the automated analysis of such screens, including ComBVAE, a probabilistic deep generative model to identify effective CRISPR guides and significantly perturbed cells, as well as to predict the outcome of the unseen combinations of perturbations. To test PerturbDecode we performed and analyzed a new large scale Perturb-Seq screen, spanning 3,390 perturbations of 1,130 E3 ligase and family members across 838,201 primary immune dendritic cells, including 660,330 single and 177,871 combinatorial perturbations. Applying PerturbDecode to these data, we demonstrate that it is powerful in filtering out the ineffective guides as well as the unperturbed cells, thus increasing the signal to noise ratio. PerturbDecode grouped the perturbed E3 ligase family members into co-functional modules that were enriched for physical interactions and impacted specific programs through substrate transcription factors. ComBVAE, leveraged the modular organization of the regulatory network to predict the outcome of unseen combinations of perturbations from sparse, randomly sampled combinatorial observations. ComBVAE-generated profiles were close to experimentally observed profiles (of withheld test data), outperformed additive models in predicting the combinatorial perturbation responses, and revealed principles of combinatorial regulation in cell circuits. PerturbDecode provides a computationally efficient probabilistic statistical analysis framework to recover causal regulatory circuits from large-scale Perturb-Seq screens.

LEARNING MULTIMODAL CELL TRAJECTORIES IN DIFFERENTIATING SYSTEMS USING SINGLE CELL MULTOMIC DATA

Alireza Karbalayghareh¹, Christopher Chin^{2,3}, Darko Barisic³, Martin Rivas³ Ari Melnick³, Christina Leslie¹

¹Memorial Sloan Kettering Cancer Center, Computational and Systems Biology Program, New York, NY, ²Weill Cornell Medicine, Department of Physiology and Biophysics, New York, NY, ³Weill Cornell Medicine, Division of Hematology and Oncology, Department of Medicine and Meyer Cancer Center, New York, NY

The epigenome and transcriptome influence each other in differentiation trajectories: epigenomic features affect the transcription of nearby genes, while expression of signaling proteins and transcription factors (TFs) enable downstream changes in chromatin state. We use single cell multiomic data, which provides RNA expression and chromatin accessibility (ATAC) readouts in each cell, to model the interplay of epigenome and transcriptome in evolving systems. More specifically, we ask if we can learn the joint dynamics of cells in both the epigenomic and transcriptomic spaces from RNA expression and TF binding motif accessibility across single cells. Although RNA velocities can provide us some hint of cell trajectories in RNA space, there is no analogous velocity information captured in the epigenomic space. We adopt techniques from variational autoencoders and dynamical systems and leverage RNA expression and velocity as well as TF motif accessibility to learn the joint dynamics of cells in both spaces. We further define a notion of TF motif accessibility velocity and latent time for cells in the evolving system. Finally, we apply these models to the germinal center (GC) B cells to learn the dynamics of wildtype cells as well as cells with mutations in important epigenetic regulators like ARID1A and CTCF, modeling somatic alterations seen in B cell lymphomas. These models are able to capture the transition of B cells from the GC dark zone to the light zone as well as cells that recycle back to the dark zone. The learned latent times demonstrate plausible starting and end points. In particular, memory B cells and plasma cells, the ultimate phenotypes of GC B cells, have the highest latent times. In silico perturbation of these models provide predictions for the impact of experimental interventions in either the epigenomic or transcriptomic spaces on both. We perform in silico perturbations of critical B cell TFs and observe the predicted velocities in both spaces, leading to further insights of the roles of these TFs in GC reaction and providing potential therapeutic targets for B cell lymphomas.

COMPLETE GENOMES OF A MULTI-GENERATIONAL PEDIGREE TO EXPAND STUDIES OF GENETIC AND EPIGENETIC INHERITANCE

Monika Cechova¹, Sergey Koren², Julian K Lucas¹, Rebecca Serra Mari³, Mobin Asri¹, David Porubsky⁴, Andrey Bzikadze⁵, Christopher Markovic⁶, Tamara Potapova⁷, Jennifer L Gerton⁷, Evan E Eichler⁴, Benedict Paten¹, Adam M Phillippy², Ting Wang⁸, Nathan O Stitzel⁸, Robert S Fulton⁶, Tobias Marschall³, Karen H Miga¹

¹UC Santa Cruz Genomics Institute, Santa Cruz, CA, ²National Human Genome Research Institute, Bethesda, MD, ³Heinrich Heine University, Medical Faculty, Düsseldorf, Germany, ⁴University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, ⁵University of California San Diego, San Diego, CA, ⁶Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, ⁷Stowers Institute for Medical Research, Kansas City, MO, ⁸Washington University School of Medicine, St. Louis, MO

Multi-generational pedigrees offer an opportunity to study transgenerational inheritance, including in biologically critical regions that are highly repetitive and are copy number variable within the population. Here, we present work aimed to release complete, T2T diploid assemblies representing a three-generational pedigree of African-American ancestry. To reach this goal, we have utilized a combination of high-coverage, long-read, and short-read paired technologies for haplotype assembly and phasing. The use of two iterative, graph-based methods with ONT-UL integration—Verkko (Rautiainen et al. 2023) and hifiasm-UL (Cheng et al. 2022)—revealed a large number (e.g. 22/46) of automated telomere-to-telomere (T2T) chromosome assemblies, with additional improvements in finishing with the two assembly methods combined. As expected, breakpoints in automated assemblies were enriched for highly repetitive regions. A study of these assemblies allowed us to predict shared haplotypes and sites of meiotic exchange. We identified a subset of inconsistent breakpoints across shared haplotypes between family members, allowing us to explore methods to further automate T2T genomes using pedigree-pangenomes.

The complete genomes provide an opportunity for a new biological discovery in repetitive parts of the genomes that were frequently missing, incomplete, or misrepresented in the past. We provide evidence for the inheritance of the multi-megabase centromere array on the X chromosome transmitted without any variation. Moreover, the methylation patterns, specifically large dips in methylation in the portion of centromeric array on the X chromosome, remained consistent across generations. Additionally, our assemblies fully span several rDNA arrays on chromosomes 21 and 22. All together these results provide a high-quality multi-generational pedigree that serves as a community resource for tracing of transgenerational inheritance of centromeres, satellite DNA, and rDNA arrays, and includes their genetic and epigenetic variation.

CELLSTAAR: INCORPORATING SINGLE-CELL BASED CELL-TYPE SPECIFIC FUNCTIONAL DATA IN RARE VARIANT ASSOCIATION TESTING OF NON-CODING REGIONS OF WHOLE GENOME SEQUENCING STUDIES

Eric Van Buren¹, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Xihong Lin^{1,2}

¹Harvard TH Chan School of Public Health, Dept. of Biostatistics, Boston, MA, ²Harvard University, Dept. of Statistics, Cambridge, MA

Whole-Genome sequencing (WGS) studies, such as the Trans-Omics for Precision Medicine Program (TOPMed) of the NHLBI, include hundreds of millions of genetic variants from hundreds of thousands of individuals. The vast majority of variants are rare, non-coding, and with unknown functional roles. It is of substantial interest to identify associations of such rare variants (RVs) with a variety of phenotypes. Because single variant tests have limited power for association tests of RVs, a number of methods to perform set-based RV Association Tests (RVATs) have been developed, including SKAT, burden, their combinations, and STAAR. For example, STAAR upweights potentially functional variants using multi-faceted functional annotations to boost power of RVATs. However, in view of massive non-coding variants whose regulatory functions are cell-specific, it is of substantial interest to incorporate single-cell sequencing data to capture the cell-type-specific functional variability that exists in the non-coding genome. In this work, we propose cellSTAAR to address two opportunities for methodological improvement of STAAR as applied to non-coding genetic variants in candidate cis-regulatory elements (cCREs) in whole genome sequencing RVATs by incorporating cell-type specific functional annotations using single cell data. First, cellSTAAR links cCREs to their target genes using an omnibus framework that aggregates results from a variety of linking approaches of mapping regulatory elements to genes, each of which uses differing kinds of genomic data. Specifically, cellSTAAR includes cCRE-gene links from distance-based intervals, ABC (enhancer-promoter contacts), EpiMap (correlation of gene expression and epigenetic activity), and SCREEN (both eQTL and 3D-based). Second, cellSTAAR integrates single-cell ATAC-seq data to capture cell-type-specific chromatin accessibility via the construction of cell-type-specific variant sets and cell-type-specific functional annotations. We perform extensive simulations to show that cellSTAAR boosts power over alternative RVAT methods in a variety of settings for analyzing non-coding variants in WGS. We further demonstrate cellSTAAR with an application to lipids phenotypes (LDL and HDL cholesterol and triglyceride levels) from Freeze 8 of TOPMed. Cell types which are the most relevant to lipids show increased discoveries and improved power as compared to less relevant cell types.

GENOMICS AT SCALE WITH THE NHGRI ANVIL

Michael Schatz¹, Anthony Philippakis²

¹Johns Hopkins University, Computer Science and Biology, Baltimore, MD,

²Broad Institute of MIT and Harvard, Data Sciences Platform, Boston, MA

Modern genomics often require very large numbers of samples to detect any subtle patterns that may be present. For example, the NHGRI Centers for Common Disease Genomics (CCDG) and Centers for Mendelian Genomics (CMG) seek to identify the genetic components of common and rare diseases through the sequencing of several hundred thousand genomes. The scale of these projects opens new opportunities for discovery; however, this scale also introduces major new technical challenges that require overhauling how genomics and genomics data science are performed.

Addressing this, the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space, or AnVIL

(<https://anvilproject.org/>), provides a secure cloud environment for the storage and 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 provides elastic, shared computing resources on demand. AnVIL currently provides harmonized access to >600,000 genomes, with many more on the horizon. We also provide access to thousands of software tools, plus several options for interactive and batch analysis.

In this presentation, we describe how AnVIL has been used in several major studies of human genomics. First, we discuss how AnVIL supports the Telomere-to-Telomere (T2T) consortium through a large-scale reanalysis of the 1000 Genomes cohort orchestrated through the Workflow Description Language (WDL) on Terra. This enabled us to identify over one million variants in the newly resolved regions of the human genome, including within the recently completed chromosome Y sequence. Next, we present new results detecting single- and multi-tissue SV-eQTLs by genotyping SVs discovered with long-reads within the GTEx short-read sequencing data. This work was quickly and securely executed through WDLs, Jupyter notebooks, and R/Bioconductor, and lead to the discovery of 5,580 SV-eQTLs where the SV has the highest CAVIAR score (a metric of causality) over other nearby SNVs. Finally, we discuss how to analyze pangenomes and haplotype diversity using Galaxy within Terra. This is critically important to diversity and disease studies, especially to capture and analyze variation not found in any single reference genome.

* The full list of contributors is available at:
<https://anvilproject.org/about/team>.

A NOVEL METHOD TO ACCOUNT FOR FINE-SCALE POPULATION STRUCTURE IN LARGE-SCALE GENOMIC ANALYSIS

Ruhollah Shemirani¹, Sinead Cullina¹, Gillian M Belbin¹, Christopher R Gignoux², Noah A Zaitlen³, Eimear E Kenny¹

¹Icahn School of Medicine at Mount Sinai, Institute for Genomic Health, New York, NY, ²University of Colorado Anschutz Medical Campus, Colorado Center for Personalized Medicine, Aurora, CO, ³University of California Los Angeles, David Geffen School of Medicine, Los Angeles, CA

Principal Components (PCs) play a prominent role in model calibration in genomic analyses, but have limited efficacy for rare variants, which often have discrete or population restricted patterns of structure. Here we propose to use graphical models derived from pairwise haplotypes shared identical-by-descent (IBD) to better capture fine-scale population structure of rare variants to optimize model calibration in rare variant analysis.

First, we detected pairwise IBD sharing for haplotypes longer than 3 cM across 350,110 self-reported White UK-born participants in the UK Biobank. We demonstrated that the eigenvectors of the Laplacian matrix of the global IBD network, called Spectral Representations (SRs), optimally extract population structure. To evaluate SRs' performance, we used easting and northing birth coordinates of the UK Biobank participants under the assumption that these vectors have solely environmental origins. SRs yield an increase of 66.7%[66.5-66.8] and 27.9%[27.6-28.0] in the proportion of variance explained compared to PCs as covariates in a linear model of easting and northing birth coordinates, respectively. In GWAS analyses, the combination of both SRs and PCs as covariates decreases genomic inflation of p-values from 1.369 to 1.097 for associations with easting birth coordinates compared to PC-only models. When estimating the heritability of easting and northing birth coordinates, adjusting for SRs reduced the estimates from 58% to 5%, 26% to 2%, respectively.

Next, we analyzed 10 phenotype and 5 Polygenic Risk Scores (PRSSs), which were previously shown to have a significant degree of spatial autocorrelation with birth coordinates due to population structure. Using Moran's Index, a measure of spatial autocorrelation, we demonstrated SRs reduce their spatial autocorrelation to a greater degree than PCs in 14 out of 15 tested phenotypes and PRSSs. In 33% (5 out of 15) for PCs, and 66% (10 out of 15) for SRs, spatial autocorrelation was no longer significantly observed after correction. Our results show that SRs have better performance in reducing biases derived from recent population structure, and have the potential to improve calibration of genomic discovery and genomic prediction models that include rare variants.

CELL-TYPE-SPECIFIC CO-EXPRESSION INFERENCE FROM SINGLE CELL RNA-SEQUENCING DATA

Chang Su¹, Zichun Xu¹, Xinning Shan¹, Biao Cai¹, Hongyu Zhao^{*1}, Jingfei Zhang^{*2}

¹Yale University, Department of Biostatistics, New Haven, CT, ²Emory University, Information Systems and Operations Management, Atlanta, GA

The advancement of single cell RNA-sequencing (scRNA-seq) technology has enabled the direct inference of co-expressions in specific cell types, facilitating our understanding of cell-type-specific biological functions. However, the high sequencing depth variations and measurement errors in scRNA-seq data present significant challenges in inferring cell-type-specific gene co-expressions, and these issues have not been adequately addressed in the existing methods. We propose a statistical approach, CS-CORE, for estimating and testing cell-type-specific co-expressions, built on a general expression-measurement model that explicitly accounts for sequencing depth variations and measurement errors in the observed single cell data. Systematic evaluations show that most existing methods suffer from inflated false positives and biased co-expression estimates and clustering analysis, whereas CS-CORE has appropriate false positive control, unbiased co-expression estimates, good statistical power and satisfactory performance in downstream co-expression analysis. When applied to analyze scRNA-seq data from postmortem brain samples from Alzheimer's disease patients and controls and blood samples from COVID-19 patients and controls, CS-CORE identified cell-type-specific co-expressions and differential co-expressions that were more reproducible and/or more enriched for relevant biological pathways than those inferred from other methods.

DISCOVERING STIMULATORY STATE SPECIFIC T2D GWAS MECHANISMS WITH SINGLE CELL MULTI-OMICS ON iPSC-DERIVED FAP VILLAGES

Christa Ventresca^{1,2}, Arushi Varshney², Peter Orchard², Yao-chang Tsan¹, Andre Monteiro da Rocha³, Markku Laakso^{4,5}, Jaakko Tuomilehto⁶, Timo A Lakka⁴, Karen L Mohlke⁷, Michael Boehnke⁸, Laura Scott⁸, Heikki A Koistinen⁶, Francis S Collins⁹, Todd Herron³, Stephanie Bielas¹, Stephen Parker^{1,2,8}

¹University of Michigan, Department of Human Genetics, Ann Arbor, MI,

²University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, ³University of Michigan, Department of Molecular & Integrative Physiology, Ann Arbor, MI, ⁴University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland, ⁵Kuopio

University, Department of Medicine, Kuopio, Finland, ⁶Finnish Institute for Health and Welfare, Department of Public Health and Welfare, Helsinki, Finland, ⁷University of North Carolina, Department of Genetics, Chapel Hill, NC, ⁸University of Michigan, Department of Biostatistics, Ann Arbor, MI, ⁹National Institutes of Health, NHGRI, Bethesda, MD

Type 2 diabetes (T2D) and related trait genome wide association study (GWAS) signals likely impact T2D risk via skeletal muscle, a primary insulin responsive tissue. We previously generated *in vivo* fibro-adipogenic progenitor (FAP) data through single nucleus (sn-)multi-omics (RNA+ATAC) profiling of 286 skeletal muscle biopsies and found that a subset of GWAS signals colocalize with cell-type specific e/caQTL in skeletal muscle FAPs. For 50 of these individuals, we derived induced pluripotent stem cell (iPSC) lines from fibroblasts. To investigate GWAS signal mechanisms, we differentiated iPSC lines to FAPs, allowing us greater flexibility to dissect the underlying signals. We hypothesize that FAPs can be differentiated from iPSC lines for the purpose of investigating the impact of FAP-specific molecular mechanisms on T2D and related traits. Here, we demonstrate that FAPs can be derived from iPSC lines for deep molecular phenotyping. These iPSC-derived FAPs display FAP morphology, expression of FAP marker genes, and loss of expression of pluripotency markers. Over the course of the differentiation, *Tra-1-60*, a pluripotency marker, drops from 99% of cells expressing this to 0.2%. Meanwhile, *CD73*, a FAP marker, increases from 0.1% of cells expressing up to 99%, based on flow cytometry analysis. We have performed a time course analysis with ten independent iPSC lines multiplexed into a single cell village to explore the trajectory from iPSCs to FAPs. Additionally we will compare the *in vitro* sn-multi-ome signatures to those in matched *in vivo* FAPs. We will upscale the experimental design to a cohort of 50 iPSC-derived FAP samples to investigate stimulatory state specific genetic regulatory effects (e/caQTL) within T2D pathways and will validate our results on a subset of loci.

EFFICIENT AND ACCURATE DETECTION OF VIRAL SEQUENCES IN BULK AND SINGLE-CELL TRANSCRIPTOMICS DATA

Laura Luebbert¹, Delaney K Sullivan¹, Lior Pachter^{1,2}

¹California Institute of Technology, Biology and Biological Engineering, Pasadena, CA, ²California Institute of Technology, Computing and Mathematical Sciences, Pasadena, CA

More than 300,000 mammalian virus species are estimated to cause infectious disease in humans. They inhabit a wide range of human tissues, including the lungs, blood, and brain, and often remain undetected. Efficient and accurate detection of viral infection is vital to understand its impact on human health, and to make accurate predictions to limit negative effects, including the prevention of future epidemics. The increasing use of high-throughput sequencing methods in research, agriculture, and healthcare provides an opportunity for the cost-effective surveillance of viral diversity and investigation of virus-disease correlation. However, there are no existing workflows for accurate real-time detection of viral infection from sequencing data. We introduce a method that accurately and rapidly detects viral sequences in bulk and single-cell transcriptomics data, enabling the detection of ongoing infection by RNA viruses covering up to 10^{12} virus species.

FUNCTIONAL DISSECTION OF COMPLEX AND MOLECULAR TRAIT VARIANTS AT SINGLE NUCLEOTIDE RESOLUTION

Layla Siraj¹, Hannah Dewey², Susan Kales², Steven Reilly³, Hilary Finucane¹, Jacob Ulirsch^{1,4}, Ryan Tewhey²

¹The Broad Institute of Harvard and MIT, Cambridge, MA, ²The Jackson Laboratory, Bar Harbor, ME, ³Yale University, Genetics, New Haven, CT,

⁴Illumina, Inc, San Diego, CA

Up to 90% of causal SNPs from genome-wide association studies reside in accessible chromatin in trait-relevant cell types, pointing to the importance of regulatory element function in human health and disease. Yet, our understanding of how single nucleotide changes affect the activity of regulatory elements is incomplete. We aimed to uncover genomic mechanisms governing regulatory element function by testing the activity of 221,747 fine-mapped variants in a Massively Parallel Reporter Assay (MPRA), the largest assessment of trait-associated variant effects by a high-throughput reporter assay to date.

We selected variants from our recent large-scale fine-mapping study across 3 biobanks, representing 124,793 credible sets including 39,868 variants with a high posterior inclusion probability (PIP) of being causal > 0.5 , as well as carefully matched location, genomic annotation, and null controls. We assayed variants across 5 diverse cell lines and observed that 92,656 (30.4%) variants exhibited baseline transcriptional activity, but only 37,284 of these modulated expression in an allele-specific manner (emVars).

We evaluated our assay's ability to identify causal regulatory variants using likely causal variants (PIP > 0.9 for at least one trait) as a gold standard. Overall, we found that variants that were emVars and within accessible chromatin outperformed all other measures for identifying likely causal noncoding variants (81% precision, 18% recall). We observed that MPRA activity correlated strongly with chromatin accessibility and transcription factor (TF) occupancy across baseline ($r = 0.61$) and allele-specific measurements ($r = 0.44$ and 0.54).

Having validated our assay, we nominated molecular mechanisms for trait-associated variants. We determined that 39% of likely causal emVAs in CREs disrupted a canonical TF binding motif but that more flexible sequence-based models of TF function could identify up to 54% of emVars with a lower background, suggesting that many complex trait variants act through non-canonical mechanisms. To this end, we performed saturation mutagenesis on 177 emVars, finding that 71 of these (40%) disrupt a clear sequence motif, many of which are unannotated.

Finally, we investigated whether multiple causal variants underlie complex-trait loci, finding both epistatic effects and multiple functional variants in single haplotypes. In conclusion, MPRA provides a powerful approach to identify and dissect non-coding regulatory variants underlying human health.

DIVERSITY IMPROVES ALL ASPECTS OF GENOMIC RESEARCH: LESSONS FROM 700,000 HUMAN EXOMES, GENOMES, AND GENOTYPES

Julia Sealock^{1,2}, Rahul Gupta^{1,2}, Katherine Chao^{1,2}, Masahiro Kanai^{1,2}, Siwei Chen^{1,2}, Kristin Tsuo^{1,2}, Wenhan Lu^{1,2}, gnomAD Consortium^{1,2}, Pan-UKB Team^{1,2}, Benjamin Neale^{1,2}, Heidi Rehm^{1,2}, Kaitlin Samocha^{1,2}, Mark Daly^{1,2}, Alicia Martin^{1,2}, Konrad Karczewski^{1,2}

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

Diversifying genomic studies is important for social justice, but often overlooked are critical scientific advances that benefit all individuals. Here, we highlight the benefits of building diverse cohorts to identify more variants across the entire frequency spectrum, improving variant interpretation for rare disease diagnosis, as well as gene discovery for common diseases.

In the gnomAD project, we have aggregated increasingly large datasets of human genomes and exomes to produce publicly available allele frequency resources. Here, we describe a new gnomAD release that comprises over 700,000 exomes, including over 30,000 and 27,000 individuals with African (AFR) and Amerindigenous (AMR) genetic ancestries, respectively. The breadth of variation discovered in gnomAD improves rare disease diagnosis by decreasing the number of putative pathogenic variants by an order of magnitude. Additionally, rare variant phasing improves within ancestries, better identifying compound heterozygosity, reaching accuracies near 1, compared to cosmopolitan estimates with accuracies ~ 0.5. Finally, we show that metrics of mutational intolerance built from diverse cohorts perform better in discerning genes and genomic regions undergoing natural selection, increasing AUC for identifying non-coding regions with fine-mapped GWAS variants from 0.66 in NFE to 0.682 in a multi-ancestry cohort of the same size.

Further, large cohorts with matched genotype and phenotype data such as the UK Biobank have been invaluable for describing etiology of common diseases. However, limited diversity hinders these efforts: in gnomAD, we find that of the 3.9 million variants accessible in rare variant association studies (max population frequency 0.01–1%), 78% of these are rare (<0.01%) and thus unassayed in individuals with European genetic ancestries (EUR). Further, we identify 12,715 genes where any genetic ancestry reaches a cumulative loss-of-function frequency > 0.01%, compared to 8,506 in EUR (a ~50% increase), increasing the power for burden testing. We extend rare variant testing in the Genebass framework to multiple genetic ancestries, as well as data from the *All of Us* project using ancestry-aware association methods.

Finally, in the Pan-UK Biobank project, we comprehensively analyzed GWAS of 452 high-quality phenotypes in multiple ancestries: GWAS quality control substantially improves with multiple ancestry groups and we discover 5,551 novel significant associations compared to Open Targets despite lower sample sizes. Further, we observe increased polygenic score accuracy for some phenotypes with lower polygeneticities for which multi-ancestry GWAS improved accuracy over EUR-only analyses, such as LDL (1.4-fold increase in AFR).

Use of data from multiple ancestries results in direct scientific advances in addition to societal benefits. We better understand the landscape of genetic variation, yielding insights into rare disease diagnosis and common disease etiology.

MILO2.0 UNLOCKS POPULATION GENETIC ANALYSES OF CELL STATE ABUNDANCE USING A COUNTS-BASED MIXED MODEL

Alice Kluzer^{1,2}, John C Marioni², Michael D Morgan³

¹ETH Zurich, Department of Biosystems Science and Engineering, Zurich, Switzerland, ²European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom, ³Institute of Medical Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, United Kingdom

Genetic demultiplexing of single-cell data has empowered the acquisition of population and patient cohort-scale data sets. Large sample sizes are needed for quantitative trait locus (QTL) analyses of gene expression and other biomolecular measurements due to the relatively small individual impact of common genetic variants on these traits. The integration of population-scale single-cell gene expression profiling with statistical genetic analysis can overcome the pitfalls of bulk RNA-sequencing that averages across cell types and which can obscure the interpretation of results from such studies. Single-cell eQTL studies have tackled this problem by first identifying constituent cell types through iterative rounds of clustering, sub-clustering and merging to generate a consistent cell type annotation across donor samples, before performing eQTL analyses within each cell type. This necessarily leads to the notion of identifying genetic variation that alters cell-type composition, which has previously only been possible using cytometry-based approaches. Therefore, single-cell population scale genetic analyses have the capacity to identify both eQTLs and cell-type QTLs within the same study.

We recently developed a powerful computational method, Milo, to identify differentially abundant cell states in single-cell experiments that dispenses with the need to define clusters and cell types. Here, we combine this clustering-free approach with the power of generalised linear mixed models to unlock the genetic analysis of cell states in population-scale single cell studies. Our approach uses a combination of approximations to rapidly and robustly analyse millions of cells from hundreds of samples, which we illustrate by identifying cell-state QTLs from population-scale cohorts of peripheral blood immune cells.

CELL TYPE-SPECIFIC AND DISEASE-ASSOCIATED eQTL IN THE HUMAN LUNG

Heini M Nattri¹, Christina Del Azodi², Lance Peter¹, Chase Taylor³, Sagrika Chugh², Robert Kendle¹, Jonathan Kropski³, Davis McCarthy², Nicholas E Banovich¹

¹Translational Genomics Research Institute (TGen), Integrated Cancer Genomics, Phoenix, AZ, ²St. Vincent's Institute of Medical Research, Bioinformatics and Cellular Genomics, Melbourne, Australia, ³Vanderbilt University Medical Center, Division of Allergy, Pulmonary and Critical Care Medicine, Nashville, TN

Understanding the genetic architecture of gene regulation and disease is a central goal of functional genomics; however, it has become increasingly clear that to fully realize the promise of these approaches, it is necessary to interrogate the genetic control of gene regulation within the proper context and the right cell type. We have applied these principles to map regulatory loci (expression quantitative trait loci, eQTL) across major cell types in the human lung. Specifically, our efforts are focused on understanding the genetic and regulatory underpinnings of interstitial lung disease (ILD). ILD is a chronic, progressive lung disease characterized by the scarring of lung tissue through epithelial remodeling and accumulation of extracellular matrix (ECM). Pulmonary Fibrosis (PF) is a clinical phenotype that exhibits the end stage of ILD. To enable cell-type specific mapping of eQTL from primary lung tissue, we have employed single-cell RNA-sequencing (scRNA-seq) to generate expression profiles from more than 475,000 cells of 116 donors (49 healthy and 67 ILD, including 40 PF). Per previous work from our group defining best practices to map single-cell eQTL (sc-eQTL), we have employed a pseudo-bulk approach, aggregating reads across cells for a given donor and cell type. With this approach, we identify thousands of sc-eQTLs across 38 cell types. Given the differences in power to detect eQTL between cell types, we leveraged multivariate adaptive shrinkage to identify shared and cell type-specific effects. Using this approach, we detect clear cell population, lineage, and cell-type-specific signals. Further, using interaction models, we identify disease-interaction sc-eQTL (int-eQTL) that exhibit differences in effects between ILD and unaffected donors. We have intersected our eQTL results with recent genome-wide association studies (GWAS) of ILD and other lung diseases, identifying sc-eQTLs that colocalize with PF associated loci. These include well-characterized risk loci regulating MUC5B and DSP. Finally, we identify subsets of sc-eQTL and int-eQTL associated with changes in cell type proportions, connecting regulatory loci to wider alveolar remodeling. This work aids in determining the cell types and contexts in which PF-associated genetic variants function and highlights the importance of cellular context in gene regulation in health and disease.

TARGETED SINGLE CELL METHYLOME PROFILING REVEALS EPIGENETIC ENCODING OF HEMATOPOIETIC STEM CELL FATE

Michael Scherer¹, Indranil Singh², Agostina Bianchi¹, Chelsea Szu-Tu¹,
Roser Zaurin¹, Renée Beekman¹, Alejo Rodriguez-Fraticelli², Lars Velten¹

¹Centre for Genomic Regulation (CRG), Barcelona, Spain, ²Institute for Research in Biomedicine (IRB), Barcelona Institute of Science and Technology, Barcelona, Spain

Isogenic Hematopoietic stem cells (HSCs) are functionally heterogeneous: They display different lineage biases, output, and propensity to transform to leukemia. These properties are stable and cell-intrinsic, that is, they are maintained over several rounds of stem cell transplantation. Surprisingly, HSCs with different function do not differ substantially in terms of gene expression.

Here, we propose that long-term HSC function is encoded at the level of DNA methylation. To investigate this hypothesis, we developed scTAM-seq, a targeted single cell methylome method based on the Mission Bio Tapestri platform. scTAM-seq profiles up to 800 CpGs in up to 10,000 cells per experiment, with a dropout rate as low as 7% per CpG, and is compatible with multiplex readout of surface proteins as well as the readout of mutations, gRNAs or lineage tracing barcodes.

We applied scTAMseq to mouse HSCs carrying lentiviral lineage tracing barcodes. We thereby obtain high-resolution maps of the early steps of HSC lineage commitment. We demonstrate that dynamic changes of CpG methylation near transcription factor motifs allows for an estimation of factor activity, and a marker- and reference-free annotation of single cell methylome data. In the final part of my talk, I will show data on the association between CpG methylation heterogeneity at the level of long term HSCs, and their functional output measured by lineage tracing.

SINGLE CELL SEQUENCING AS A UNIVERSAL VARIANT INTERPRETATION ASSAY

Dan Cao*, Ken Jean-Baptiste*, Mohan Sun*, Xingyan Zhou*, Sandy Wong*, Ling Chen, Hongxia Xu#, Francois Aguet#, Kyle Farh#

Illumina, Inc., Artificial Intelligence Laboratory, Foster City, CA

*Contributed equally, #Contributed equally

The vast majority of all possible ~70 million protein-altering variants in the human genome are of uncertain significance, with only a small fraction annotated in clinical variant databases. Closing this gap is essential to identify clinically relevant variants and understand their mechanism of action. Toward this goal, we transduced all possible coding variants in the TP53, CDKN2A (p16^{INK4a}), and SOD1 genes underlying pan-cancer, melanoma, and amyotrophic lateral sclerosis, respectively, and developed an approach based on single-cell RNA sequencing to read out the functional effects of each variant in the global expression signature. We sequenced ~215,000 cells expressing TP53 variants, ~140,000 cells expressing CDKN2A variants, and ~450,000 cells expressing SOD1 variants, with 89.3% (972), 98.5% (2774), and 100% (1113) of all possible amino acid variants observed in at least 20 cells for the three genes, respectively. We used variant enrichment and expression profiles to identify the gene expression programs that are perturbed by pathogenic variants and quantify proliferation differences. Using both supervised and unsupervised classification approaches to quantify the pathogenicity of each variant (e.g., with respect to synonymous variants), we show strong concordance with prior experimental measurements of TP53 variant effects as well as computational predictions, and accuracies of >90%–100% for pathogenic ClinVar variants. In summary, we show that it is possible to quantify the effects of almost all variants of a gene with single cell sequencing, demonstrating its use as a universal variant interpretation assay.

A CHROMOSOME-SCALE CRISPR SCREEN TO IDENTIFY ESSENTIAL ELEMENTS IN THE HUMAN NONCODING GENOME

Xinyi Guo^{1,2}, Chirag Lakhani^{1,3}, Noa Liscovitch-Brauer^{1,2}, Qiyao Lin^{1,2}, Christina Caragine^{1,2}, Giuseppe Narzisi¹, Edward Paik^{1,2}, Caroline Yu^{1,2}, David Knowles^{1,3}, Neville E Sanjana^{1,2}

¹New York Genome Center, New York, NY, ²New York University, Department of Biology, New York, NY, ³Columbia University, Department of Computer Science, New York, NY

More than 98% of human genome does not code for proteins, and it is unclear which regions of the noncoding genome may be indispensable for cell fitness. To date, CRISPR-based functional genomic screens have focused on protein-coding genes or on specific *cis*-regulatory regions. To identify noncoding essential elements at chromosome scale, we developed a pooled CRISPR-Cas12a screen to systematically tile the entire human chromosome 20 with overlapping deletions. In total, we engineered ~120,000 5kb deletions in a human haploid cell line (HAP1) spanning 64 Mb in total. Overall, ~6% of noncoding deletions negatively impact HAP1 cell growth, and only a small fraction (0.04%) stimulates cell growth. Using this unique dataset, we assembled the first whole-chromosome map of essential elements — achieving a final resolution of 1kb by taking advantage of designed overlap between neighboring deletions (robust rank aggregation). Top-ranking essential noncoding regions tend to be near gene-dense regions. About half of them reside within gene introns (especially the first and second introns), and the genes that harbor them are often essential genes. The other half are intergenic, and on average are 8 kb away from the nearest gene. We find that 63% of essential intergenic elements are upstream of the closest TSS, reflecting the known 5' bias of *cis*-regulatory elements. We individually tested and validated 22 essential regions and found excellent correlation with the initial screen ($r = 0.9$), demonstrating the high quantitative accuracy of the chromosome-wide screen. To pinpoint functional elements with nucleotide resolution, we designed a saturation mutagenesis screen using ~29,000 CRISPR-Cas9 constructs that targets 141 regions of chromosome 20 that ranked as highly-essential in the chromosome-wide screen. Then, using a deep learning approach, we show that these regions are predicted to be enriched for specific transcription factor binding sites, which are notably absent from control noncoding regions or regions with mutations in healthy controls from GnomAD ($n = 76,156$ whole genomes).

Altogether, we have developed a series of approaches to functionally screen entire human chromosomes with base pair resolution. This work brings us a major step closer to the tantalizing possibility of truly genome-wide CRISPR screens that are capable of discovery in all parts of the human genome.

MACHINE-GUIDED DESIGN OF *CIS*-REGULATORY ELEMENTS WITH CELL-TYPE SPECIFICITY

Rodrigo I Castro^{*1}, Sager J Gosai^{*2,3}, Natalia Fuentes^{2,3}, Kousuke Mouri¹, Pardis C Sabeti^{2,3,4,5}, Steven K Reilly⁶, Ryan Tewhey¹

¹The Jackson Laboratory, Bar Harbor, ME, ²The Broad Institute, Cambridge, MA, ³Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA, ⁴Howard Hughes Medical Institute, Chevy Chase, MD, ⁵Harvard T.H. Chan School of Public Health, Department of Immunology and Infectious Disease, Boston, MA, ⁶Yale University, Department of Genetics, New Haven, CT

*Authors contributed equally

In recent years, the generation of genomic functional datasets together with machine-learning advances have enabled the development of sequence models that implicitly learn much of the logic driving transcription and gene expression. These advances provide the opportunity to comprehensively inspect the landscape of regulatory syntax and potentially design novel sequences with defined transcriptional objectives. Here, we present an experimentally validated generative framework of cell-specific synthetic enhancers using Malinois, a multi-task convolutional neural network that accurately predicts *cis*-regulatory activity in three human cell lines by leveraging a large high-quality set of massively parallel reporter assays (MPRA). Malinois precisely reproduces reporter assays (Pearson's $r = 0.88$), tiling and saturation mutagenesis screens, and its predictions are tightly associated with DHS and H3K27ac signals. We coupled Malinois with evolutionary and gradient-based algorithms to guide the design of novel sequences incentivized by the activity separation between target and non-target cell types, and proposed a library of ~70K synthetic enhancers with high cell specificity to be tested by MPRA. We also selected 24K sequences naturally occurring in the genome either based on their high cell-specific DHS signal or Malinois predictions. Experimental validation shows that the specificity hit rate of the synthetic sequences ranged between 95% and 99%, while the rates of the natural sequences proposed by Malinois or DHS were 81% and 53% respectively. The synthetic group exhibited distinct combinations of transcriptional programs not observed in natural sequences, underscoring our ability to design novel regulatory syntax. Finally, we evaluated the ability of these synthetic sequences to drive specific expressions *in vivo*. One of two neuron-specific synthetic sequences demonstrated tissue specific expression in the brains of 15-day-old mouse embryos, while three of four liver-specific synthetic sequences showed high liver-specific expression in zebrafish. This work provides a generalizable framework to rationally design *cis*-regulatory elements that can jointly refine transgene expression across multiple cell types.

EXTENSIVE NATURAL VARIATION AND SILENCING OF lncRNAs IN ARABIDOPSIS THALIANA

Aleksandra E Kornienko, Viktoria Nizhynska, Magnus Nordborg
Gregor Mendel Institute, Norborg group, Vienna, Austria

Background

Long non-coding RNAs (lncRNAs) are under-studied and under-annotated in plants. In mammals, lncRNA expression has been shown to be reaching the extent of protein-coding expression and be highly variable between individuals of the same species. Using *A. thaliana* as a model plant organism, we aimed to understand the true scope of lncRNA transcription across plants from different regions, characterize lncRNA natural expression variability, and study the causes of this variability.

Results

Using RNA-seq data spanning 499 natural lines and 4 different developmental stages to create a more comprehensive annotation of lncRNAs in *A. thaliana*, we found over 10,000 novel loci — three times as many as in the current public annotation. We showed that, while lncRNA loci are ubiquitous in the genome, most appear to be actively silenced and their expression and repressive chromatin levels are extremely variable between natural lines. It was particularly prominent in intergenic lncRNAs, where TE-like sequences present in 50% of the loci are associated with increased silencing and variation and such lincRNAs tend to be targeted by TE silencing machinery. Analyzing full genomes of multiple natural accessions showed that lincRNAs show very high structural variation that is responsible for much of the expression variation we observed.

Conclusion

lncRNAs are ubiquitous in the *A. thaliana* genome, but their expression is highly variable between different lines and tissues. This high expression variability is largely caused by high structural and epigenetic variability of non-coding loci, especially those containing pieces of transposable elements. We create the most comprehensive *A. thaliana* lncRNA annotation to date and improve our understanding of plant lncRNA biology.

DRAGONNNFRUIT: LEARNING CIS- AND TRANS-REGULATION OF CHROMATIN ACCESSIBILITY AT SINGLE BASE AND SINGLE CELL RESOLUTION

Jacob Schreiber, Surag Nair, Anshul Kundaje

Stanford University, Genetics, Stanford, CA

Cellular differentiation and reprogramming involve multiple trajectories of continuous cell state transitions, which can be characterized with scATAC-seq and scRNA-seq experiments. However, deciphering the cis- and trans-regulatory drivers of cell state at the single-cell level is challenging due to sparse and noisy measurements.

We introduce DragoNNFruit, a first-of-its-kind approach that jointly models cis- and trans-regulatory factors of genome-wide chromatin accessibility from scATAC-seq experiments at single-cell and base-pair resolution. At a high level, DragoNNFruit models cis-regulatory sequence using a convolutional neural network whose parameters are dynamically generated from a second network that models trans-regulatory state (from ATAC-seq, RNA-seq, spatial coordinates, etc). Through the inclusion of an explicit model of Tn5 bias, DragoNNFruit's de-noised predictions reveal TF footprints that cannot be observed in the original, biased, data. Taken together, DragoNNFruit's capabilities enable the identification of cell type-specific motifs and their higher-order syntax, the prediction of variant effect and footprinting genome-wide, and the tracking of all these across entire single-cell experiments without the need for cell clustering and annotation.

By explicitly modeling both cis- and trans-regulatory factors, DragoNNFruit departs from current regulatory modeling approaches such as Enformer, scBasset, and ChromVAR. Neither Enformer nor scBasset explicitly model trans-regulatory factors and so cannot generalize past observed cell states, and neither operate at basepair resolution, limiting their interpretability and predictive power. ChromVAR explicitly tracks global TF motif activity across cells, but lacks a cis component and so cannot reveal how local syntax affects TF binding at individual loci.

We showcase DragoNNFruit's capabilities by modeling single cell chromatin dynamics across a fibroblast to iPSC reprogramming timecourse. DragoNNFruit's predictions allow precise timestamping of enhancer activation/repression along the reprogramming trajectory. Interpreting DragoNNFruit reveals regulatory motifs and their cell-specific activities. Further, DragoNNFruit reveals that scATAC-seq encodes differences in TF footprint depths that correlate with TF stoichiometry and motif affinity. We infer cis- and trans-regulatory drivers of on- and off-target trajectories. These analyses highlight the locus-specific, temporal, and quantitative interplay between cis- and trans-regulatory factors, and demonstrate that DragoNNFruit offers a powerful new paradigm for understanding dynamic regulation of cell fate decisions at single cell resolution.

MAPPING THE CONVERGENCE OF GENES FOR CORONARY ARTERY DISEASE ONTO ENDOTHELIAL CELL PROGRAMS

Gavin R Schnitzler^{*1}, Helen Kang^{*2,3}, Vivian S Lee-Kim^{1,4}, Rosa X Ma^{2,3}, Tony Zeng^{2,3}, Ramcharan S Angom⁵, Shi Fang^{1,4}, Shamsudheen K Vellarikkal^{1,4}, Ronghao Zhou^{2,3}, Katherine Guo^{2,3}, Oscar Sias-Garcia^{1,4}, Alex Bloemendal¹, Glen Munson¹, Debabrata Mukhopadhyay⁵, Eric S Lander^{1,6,7}, Hilary K Finucane^{1,8}, Rajat M Gupta^{†1,4}, Jesse M Engreitz^{†1,2,3}

¹Broad Institute, Cambridge, MA, ²Stanford University, Dept of Genetics, Stanford, CA, ³Betty Irene Moore Children's Heart Center, BASE Initiative, Stanford, CA, ⁴Brigham and Women's Hospital, Div of Genetics and Cardiology, Boston, MA, ⁵Mayo Clinic College of Medicine and Science, Dept of Biochemistry + Molecular Biology, Jacksonville, FL, ⁶MIT, Dept of Biology, Cambridge, MA, ⁷Harvard Medical School, Dept of Systems Biology, Cambridge, MA, ⁸Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA

*Equal contribution.

†Equal contribution.

Genome-wide association studies (GWAS) have discovered thousands of risk loci for common, complex diseases, each of which could point to genes and gene programs that influence disease. For some diseases, it has been observed that GWAS signals converge on a smaller number of biological programs, thereby helping to identify causal genes. However, identifying such convergence remains challenging: each GWAS locus can have many candidate genes, each gene might act in one or more possible programs, and it remains unclear which programs might influence disease risk. Here, we developed a new approach to address this challenge, by creating unbiased maps to link disease variants to genes to programs (V2G2P) in a given cell type. We applied this approach to study the role of endothelial cells in the genetics of coronary artery disease (CAD). To link variants to genes, we constructed enhancer-gene maps using the Activity-by-Contact model. To link genes to programs, we applied CRISPRI-Perturb-seq to knock down all expressed genes near 306 CAD GWAS signals and identify their effects using single-cell RNA-sequencing. By combining these variant-to-gene and gene-to-program maps, we find that 43 of 306 CAD GWAS signals converge onto 5 gene programs linked to the cerebral cavernous malformations (CCM) pathway—which is known to coordinate transcriptional responses in endothelial cells, but has not been previously linked to CAD risk. The strongest regulator of these programs is TLNRD1, which we show is a new CAD gene and novel regulator of the CCM pathway. Together, our study identifies convergence of CAD risk loci into prioritized gene programs in endothelial cells, nominates new genes of potential therapeutic relevance for CAD, and demonstrates a generalizable strategy to connect disease variants to functions.

MUTATE EVERYTHING: GLOBALLY MAPPING ALLOSTERIC COMMUNICATION IN PROTEINS

Ben Lehner^{1,2}

¹Wellcome Sanger Institute, Human Genetics, Cambridge, United Kingdom, ²CRG, Systems and Synthetic Biology, Barcelona, Spain

Thousands of proteins have now been genetically-validated as therapeutic targets in hundreds of human diseases. However, very few have actually been successfully targeted and many are considered ‘undruggable’. This is particularly true for proteins that function via protein-protein interactions: direct inhibition of binding interfaces is difficult, requiring the identification of allosteric sites. However, most proteins have no known allosteric sites and a comprehensive allosteric map does not exist for any protein. Here we address this shortcoming by charting multiple global atlases of inhibitory allosteric communication in KRAS, a protein mutated in 1 in 10 human cancers. We quantified the impact of >26,000 mutations on the folding of KRAS and its binding to six interaction partners. Genetic interactions in double mutants allowed us to perform biophysical measurements at scale, inferring >22,000 causal free energy changes, a similar number of measurements as the total made for proteins to date. These energy landscapes quantify how mutations tune the binding specificity of a signalling protein and map the inhibitory allosteric sites for an important therapeutic target. Allosteric propagation is particularly effective across the central beta sheet of KRAS and multiple surface pockets are genetically-validated as allosterically active, including a distal pocket in the C-terminal lobe of the protein. Allosteric mutations typically inhibit binding to all tested effectors but they can also change the binding specificity, revealing the regulatory, evolutionary and therapeutic potential to tune pathway activation. Using the approach described here it should be possible to rapidly and comprehensively identify allosteric target sites in many important proteins.

UNDERSTANDING COMPLEX GENOTYPE-PHENOTYPE MAPS

Carlos Martí-Gómez, David McCandlish

Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY

Multiplex assays of variant effect now allow the functional characterization of an unprecedented number of sequence variants in both gene regulatory regions and protein coding sequences. This has enabled the study of full combinatorial libraries containing hundreds of thousands to millions of genotypes, and revealed the widespread influence of higher-order genetic interactions that arise when multiple mutations are combined. However, the lack of appropriate tools for exploratory analysis of this high-dimensional data limits our overall understanding of the main qualitative properties of complex sequence-function relationships. To fill this gap, we have developed gpmap-tools (<https://gpmap-tools.readthedocs.io>), a python library that integrates (1) Gaussian process-based methods for inference, phenotypic imputation, and error estimation from incomplete and noisy data, and (2) methods for non-linear dimensionality reduction and visualization of genotype-phenotype maps containing millions of genotypes.

To demonstrate the power of this framework, we applied our software to two case studies. In the first case study, we examined how translation efficiency in *E.coli* depends on the Shine-Dalgarno sequence, a 5'UTR motif that modulates binding to the 16S rRNA through base pair complementarity during translation initiation. Visualization of the inferred genotype-phenotype map (containing ~250,000 genotypes) shows that the set of functional Shine-Dalgarno sequences has a complex structure consisting of both extended ridges of highly functional sequences and isolated functional peaks. This pattern results from the combination of the ability of the 16S rRNA to bind at variable distances from the start codon and the quasi-repetitive nature of the canonical Shine-Dalgarno motif, which is a near match to its own 3 nucleotide shift. In the second case study, we explore a 4-codon genotype-phenotype map containing ~16 million sequences based on data from a deep mutational scanning study of Protein G, a bacterial immunoglobulin binding protein. We find that the set of functional sequences is connected in sequence space, but only via long winding paths along which mutations must be accumulated in a highly specific order. We show that these constrained paths result from a combination of epistatic interactions at the protein sequence level and the structure of the genetic code.

PROFILING THE NEUROBIOLOGY UNDERLYING BRAIN STRUCTURE IN LIVING HUMAN SUBJECTS

Anina N Lund, Noam D Beckmann, Alexander W Charney

Icahn School of Medicine at Mount Sinai, Charles Bronfman Institute for Personalized Medicine, New York, NY

A singular goal of neuroscience is to advance knowledge of the processes whereby neurobiology gives rise to human brain function. The anatomical structure of the human brain is believed to be a critical component of these processes, but surprisingly to date there have been no large-scale studies of the relationship between neurobiology and brain structure in living humans. Such studies have not been performed due to the inability to obtain brain tissue samples from living human cohorts. The Living Brain Project (LBP) was designed to overcome this limitation of human brain research by obtaining samples from the dorsolateral prefrontal cortex (dlPFC) for molecular research in a large cohort of living subjects who undergo multimodal neuroimaging. RNA sequencing was performed on dlPFC samples from 171 individuals, and integration of this data with structural MRI data from the same individuals identified genes whose levels of expression associated with imaging metrics of the dlPFC including cortical thickness, volume and area. This study represents the first attempt to connect genomics with brain structure in living people, and marks an important step towards gaining a more complete picture of the molecular processes underlying human brain function.

PRECISE MODULATION OF TRANSCRIPTION FACTOR LEVELS
REVEALS NON-LINEAR DOSAGE RESPONSES WITHIN
TRANSCRIPTIONAL NETWORKS

Julia Domingo¹, Mariia Minaeva², Marcello Ziosi¹, John A Morris^{1,3}, Tuuli Lappalainen^{1,2}

¹New York Genome Center, Computational Genomics, New York, NY,

²Science for Life Laboratory, KTH Royal Institute of Technology,

Department of Gene Technology, Stockholm, Sweden, ³New York

University, Department of Biology, New York, NY

Genomic loci associated with common traits and diseases are typically non-coding and likely impact gene expression, sometimes coinciding with rare loss-of-function variants in the target gene. However, our understanding of how gradual changes in gene dosage affect molecular, cellular, and organismal traits is currently limited. To address this gap, we induced gradual changes in gene expression of three master trans-regulators associated with blood cell traits (GFI1B, NFE2, and MYB) using CRISPR activation and inactivation, and examined the downstream consequences in their transcriptional pathways using targeted single-cell multimodal sequencing. We showed that guide tiling around the TSS is the most effective way to modulate cis-gene expression across a wide range of fold-changes, with further effects from chromatin accessibility and histone marks that differ between the inactivation and activation systems. Our single-cell data allowed us to precisely detect subtle to large gene expression changes in dozens of trans genes, revealing that many responses to dosage changes of these three transcription factors are non-linear, including non-monotonic behaviors, even when constraining the fold-changes of the master regulators to a copy number gain or loss. We found that some of these non-linear responses are enriched for disease and GWAS genes. Additionally, we observed that trans-genes with a large magnitude of expression response are enriched for less central roles in TF-target networks, indicating a lack of constraint. Overall, our study provides a straightforward and scalable method to precisely modulate gene expression and gain insights into its downstream consequences at high resolution.

DYNAMIC GENE CONTENT EVOLUTION ON *DROSOPHILA* Y CHROMOSOMES

Matthew J Nalley, Doris Bachtrog

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

Y chromosomes in *Drosophila* are typically viewed as gene deserts, and theory and empirical data suggest that nucleotide diversity in protein coding genes is low. Here, we show that the Y chromosomes of *D. pseudoobscura* and *D. miranda* contains hundreds or thousands of genes, many of which are multi-copy. Most surprisingly, the gene content of different Y chromosomes is highly variable within a species, and polymorphic Y's contain dozens of unique single-copy and amplified gene families. Thus, counter to the current dogma, the Y chromosome is in fact one of the most variable chromosomal environments within the genome of a species. Many Y-linked genes are expressed in testis and enriched for spermatogenesis functions, suggesting an important role in males. Y-genes are also enriched for chromatin organization and histone exchange, processes that are known to be targeted by meiotic drivers, and several highly amplified Y-genes produce small RNAs and have orthologs that co-amplified on the X chromosome. Polymorphic Y's could thus allow for male-specific phenotypic variation during spermatogenesis or be involved in meiotic conflicts with their homolog.

GENETIC CAUSES AND PHENOTYPIC CONSEQUENCES OF NEWLY EVOLVED ADRENAL CELL TYPE

Natalie Niepoth^{1,2}, Jennifer Merritt^{1,2}, Michelle Uminski^{1,2}, Sarah Wacker³,
Stefano Lutzu⁴, Stephanie Rudolph⁴, Andres Bendesky^{1,2}

¹Columbia University, Department of Ecology, Evolution, and Environmental Biology, New York, NY, ²Columbia University, Zuckerman Mind Brain Behavior Institute, New York, NY, ³Manhattan College, Department of Chemistry & Biochemistry, New York, NY, ⁴Albert Einstein College of Medicine, Department of Neuroscience, New York, NY

Animal behavior is fundamentally regulated by cell types with specialized functions; however, the genetic mechanisms underlying the emergence of novel cell types and their consequences for behavior are not well understood. Here, we show that the monogamous oldfield mouse (*Peromyscus polionotus*) has evolved a novel cell type in the adrenal cortex that is characterized by the expression of an enzyme that reduces progesterone to 20 α -hydroxyprogesterone (20 α -OHP). We then demonstrate that 20 α -OHP is more abundant in oldfield mice than in the closely-related promiscuous deer mouse (*P. maniculatus*), and that it induces monogamous-typical parental behaviors when administered to promiscuous mice. We ultimately discover genetic variation between deer mice and oldfield mice which drives expression of the glycoprotein Tenascin-N and underlies the existence of this cell type. Our results provide an example by which the recent evolution of a new cell type in a gland outside the brain contributes to the evolution of social behavior.

FORGING NEW REGULATORY ELEMENTS DURING THE 500 MILLION YEARS OF EVOLUTION LEADING TO HUMANS

Riley J Mangan, Christiana Fauci, Yanting Luo, Craig B Lowe

Duke University School of Medicine, Molecular Genetics and Microbiology, Durham, NC

We are interested in understanding the genetic changes that encode the phenotypic diversity observed across vertebrate species. Differences in gene regulation are likely to be important, but the relative contribution of forging new regulatory elements, versus modifying existing ones, is not well understood. We are working to understand the origination of regulatory elements across evolutionary time scales: a macroevolutionary analysis to understand the creation of regulatory elements during the last 500 million years of vertebrate evolution, and a microevolutionary analysis on the last 6 million years of human evolution where new regulatory elements separate us from the other great apes. In our macroevolutionary analysis we infer the branch of origin for all open chromatin regions across hundreds of human cell types by identifying the most distantly related vertebrate ortholog. Skeletal muscle has the strongest enrichments for ancient regulatory elements shared across all vertebrates, while cell types associated with the immune system have the strongest enrichments for recently gained regulatory elements in the last 100 million years. We also observe that cell types originating within vertebrates show a burst of regulatory elements being born on the branch when the cell type originated. Our microevolutionary analysis to identify regulatory regions born on the branch from the human-chimpanzee ancestor to present-day humans recently used extreme divergence between the inferred ortholog in the human-chimpanzee ancestor and the sequence in present-day humans, without restricting to regions of cross-species conservation. We termed these regions HAQERs (Human Ancestor Quickly Evolved Regions). Based on derived allele frequencies in human populations, HAQERs show a signature of rapid divergence and positive selection, followed by constraint. Based on epigenetic datasets, many HAQERs are likely to be tightly-controlled regulatory elements that have preferentially influenced gene expression in the developing brain, immune system, and gastrointestinal tract of humans. We are also investigating what could be considered an even more extreme version of genomic divergence: insertions in the human genome where the other great apes lack an orthologous sequence. We term these regions UHIs (Unique Human Insertions). To verify that the sets of HAQERs and UHIs contain hominin-specific enhancers, we developed a multiplex single-cell enhancer assay in the developing mouse brain. Based on testing extant and ancestral sequences, we have identified six HAQERs that act as hominin-specific enhancers. We propose that forging functional elements from previously non-functional regions is likely to play an outsized role in regulatory differences among species by circumventing pleiotropic constraints that reduce the evolvability of many highly conserved developmental enhancers.

RESOLUTION OF STRUCTURAL VARIATION IN DIVERSE MOUSE GENOMES REVEALS CHROMATIN REMODELING DUE TO TRANSPOSABLE ELEMENTS

Ardian Ferraj^{1,2}, Peter A Audano², Parithi Balachandran², Anne Czechanski³, Jacob I Flores², Alexander A Radecki^{1,2}, Varun Mosur², David S Gordon⁴, Isha A Walawalkar^{1,2}, Evan E Eichler⁴, Christine R Beck^{1,2}

¹University of Connecticut Health Center, Genetics and Genome Sciences, Farmington, CT, ²The Jackson Laboratory for Genomic Medicine, Beck Lab, Farmington, CT, ³The Jackson Laboratory, Reinholdt Lab, Bar Harbor, ME, ⁴University of Washington School of Medicine, Genome Sciences, Seattle, WA

Diverse inbred mouse strains are important biomedical research models, yet genome characterization of many strains is fundamentally lacking in comparison to humans. In particular, catalogs of structural variants are incomplete, limiting the discovery of causative alleles for phenotypic variation. Here, we utilize long-read sequencing to resolve genome-wide structural variants (SVs, variants ≥ 50 bp) in 20 genetically distinct inbred mice. We report 413,758 site-specific SVs that affect 13% (356 Mbp) of the mouse reference assembly, including 510 previously unannotated coding sequence variants. We find that 39% of SVs are attributed to transposable elements (TEs), accounting for 75% of bases altered by SV. We utilized this callset to investigate the impact of TE heterogeneity on mouse embryonic stem cells (mESCs) and find multiple TE classes that influence chromatin accessibility. Our work provides a comprehensive analysis of SVs found in diverse mouse genomes and illustrates that previously unresolved TEs underlie epigenetic differences in mESCs.

INVESTIGATING THE ROLE OF INSERTIONS IN THE HUMAN GENOME

Yanting Luo, Craig Lowe

Duke University Medical Center, Department of Molecular Genetics and Microbiology, Durham, NC

Since humans diverged from chimpanzees, we evolved unique phenotypes, such as advanced cognitive capabilities and susceptibilities to psychiatric disorders. The search for the genetic bases of these traits, which are likely to be changes in gene regulation, has previously focused on regulatory elements that have been modified by substitutions, small insertions, and/or small deletions. The study of new regulatory sequences that arose from larger insertions has been uniquely challenging due to their length (which often requires long reads), repetitiveness and lack of orthologs in other organisms. However, new technologies and resources have allowed me to study how novel sequences regulate gene expression to contribute to human-unique phenotypes. I have written software to identify 23,660 Unique Human Insertions (UHIs) that are present in humans but absent in non-human primates. UHIs are 98% noncoding and enriched in population variants. Based on existing epigenetic data sets, many UHIs are putative gene regulatory elements in a variety of tissues, such as the developing brain. I am investigating the functional consequences of UHIs not only as distal enhancers but also as proximal regulatory elements. I am taking advantage of the *in vivo* multiplex STARR-seq assay to quantify the enhancer activity of 41 UHIs in the cerebral cortex of embryonic mice. Additionally, I developed a multiplex promoter assay to test UHIs. My pilot experiment found one UHI that caused higher transcriptional activity in a human promoter compared to the corresponding chimpanzee sequence. Overall, these previously understudied novel sequences may have shaped human adaptation, and their variants may contribute to differences in disease susceptibilities between human populations.

TELOMERE-TO-TELOMERE APE SEX CHROMOSOME
ASSEMBLIES UNRAVEL RAPID EVOLUTION ON THE Y AND
CONSERVATIVE EVOLUTION ON THE X

Kateryna D Makova¹, Brandon D Pickett², Sergey Nurk², DongAhn Yoo³,
Hyeonsoo Jeong³, Barbara McGrath¹, Robert S Harris¹, Monika Cechova⁴,
Gabrielle A Hartley⁵, Jessica M Storer⁶, Patrick Grady⁵, Tamara Potapova⁷,
Matthew Borchers⁷, Sergey Koren², Jennifer L Gerton⁷, Rachel O'Neill⁵, Evan E
Eichler³, Adam M Phillippe²

¹Penn State University, Department of Biology, University Park, PA, ²NHGRI,
Bethesda, MD, ³University of Washington, Department of Genome Sciences,
Seattle, WA, ⁴UCSC, Department of Biomolecular Engineering, Santa Cruz,
CA, ⁵University of Connecticut, Department of Cellular and Molecular Biology,
Storrs, CO, ⁶ISB, Seattle, WA, ⁷Stowers Institute, Kansas City, MO

Evolution of ape sex chromosomes has remained enigmatic due to their highly repetitive nature and incomplete reference assemblies (particularly for the Y). Here we generated gapless, telomere-to-telomere (T2T) assemblies of the X and Y chromosomes for all extant great ape species—chimpanzee, bonobo, gorilla, Bornean and Sumatran orangutans—and for an outgroup gibbon species, the siamang. To achieve this, we utilized state-of-the-art experimental and computational methods developed for deciphering the human T2T genome. These assemblies completely resolved ampliconic and satellite sequences, and allowed us to untangle ape sex chromosome evolution in unprecedented detail, leading to the following results. First, despite the divergence time of <18 million years, only 12-26% of non-human ape Y sequences align to the human Y, compared to 84-97% of non-human ape X sequences aligning to the human X. Second, large interspecies structural rearrangements are strongly enriched on the Y vs. the X. The siamang Y acquired the highest number of structural rearrangements, consistent with rapid evolution of gibbon genomes. Third, depending on species, segmental duplications represent 21-53% of ape Y assemblies, compared to only 3.9-6.0% of ape X assemblies. The chimpanzee and bonobo Y chromosomes have two times more segmental duplications than the other Ys, which might be explained by differences in mating patterns among apes. Fourth, ampliconic sequences constitute 8.7-46% of ape Y assemblies, compared to only 0.15-1.1% of ape X assemblies. Most of such sequences on the Y are species-specific, except for those in closely related species (e.g. chimpanzee and bonobo). Fifth, per RepeatMasker, repeats account for 71-85% of ape Y assemblies compared to 57-64% of ape X assemblies. One-third of the gorilla and siamang Ys consists of satellites, with the former enriched in subterminal pCht repeats and the latter in alpha-satellites. Sixth, ribosomal DNA arrays were identified on the Y chromosomes of only the siamang and Sumatran orangutan and never on the X chromosomes. Seventh, whereas most apes harbor one pseudoautosomal region (PAR) on their sex chromosomes, which is homologous among them, bonobo and human each independently acquired a second PAR. Our analyses indicate a remarkably dynamic evolution on the largely non-recombining Y chromosome, in contrast to a more stable evolution on the X chromosome. As the Y harbors regions important for fertility, our research will inform future studies of conservation genetics of non-human apes, all of which are endangered species.

THE MUTATIONAL LANDSCAPE OF NORMAL GASTRIC EPITHELIUM

Tim Coorens^{1,2}, Grace Collord¹, Hyunchul Jung¹, Yichen Wang¹, Suet Yi Leung³, Michael Stratton¹

¹Wellcome Sanger Institute, Cancer, Ageing and Somatic Mutation, Hinxton, United Kingdom, ²Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, ³Hong Kong University, Department of Pathology, Hong Kong, Hong Kong

The gastric mucosa is the thin lining of the stomach and contains an epithelial layer, which harbors the glands that produce digestive enzymes and acid. Gastric cancer arises from this epithelial lining and is one of the most prevalent cancer types in humans. Major risk factors for developing gastric cancer include infection with the bacterium Helicobacter pylori as well as smoking. However, the background mutational landscape in normal gastric epithelium and the first genomic steps towards the formation of gastric cancer remain poorly understood.

Here, we used whole-genome sequencing of microdissected gastric glands ($n=275$) from 34 donors, 18 of whom had gastric cancer. We show that gastric glands are clonal units derived from single stem cells and accrue approximately 27 base substitutions per year. The mutagenic processes active in normal gastric epithelium differ profoundly from those found in gastric cancers. While the mutational signatures in most normal glands reflect age-related mutagenesis, gastric glands sampled close to a tumor showed exposure to a mutagenic process highly enriched in tumors (COSMIC reference mutational signature SBS17). This suggests the acquisition of SBS17 substitutions is closely linked to overt malignant transformation in the stomach.

We also observe widespread trisomies of specific chromosomes (13, 18 and 20), which are recurrently and independently acquired in many gastric glands in a small subset of patients. These trisomies were often confined to a single biopsy per patient and acquired around the same time, with different alleles of the same chromosome duplicated within patients. These findings indicate the trisomies likely conveyed an advantage to a transient selection pressure, possibly through a gene dosage effect.

We find that mutations in genes encoding epigenetic modifiers and chromatin remodelers were under positive selection for and highly enriched in some patients. This was confirmed by targeted sequencing of cancer genes in a further 1,008 gastric glands. Strikingly, glands that exhibit driver mutations, a recurrent trisomy or elevated mutation loads only overlap minimally.

Taken together, these results suggest a highly variable and patient-specific genomic landscape in the normal stomach. Our findings portray the gastric epithelium with dynamic selective pressures, actively molding the somatic evolution of these cells in healthy, precancerous and malignant states.

GERMLINE-MEDIATED IMMUNOEDITING SCULPTS BREAST CANCER SUBTYPES

Kathleen E Houlahan¹, Aziz Khan¹, Noah F Greenwald^{2,3}, Michael Angelo³, Christina Curtis^{1,4,5,6}

¹Stanford University School of Medicine, Stanford Cancer Institute, Stanford, CA, ²Stanford University School of Medicine, Cancer Biology Program, Stanford, CA, ³Stanford University School of Medicine, Department of Pathology, Stanford, CA, ⁴Stanford University School of Medicine, Department of Medicine (Oncology), Stanford, CA, ⁵Stanford University School of Medicine, Department of Genetics, Stanford, CA, ⁶Stanford University School of Medicine, Department of Biomedical Data Science, Stanford, CA

Cancer represents a wide spectrum of molecularly and morphologically diverse diseases. Individuals with the same histopathological classification can have tumors with drastically different molecular profiles and clinical responses to treatment. It remains unclear as to when during the disease course these differences arise and why some tumors are addicted to one oncogenic pathway over another. Somatic genomic aberrations occur within the context of an individual's germline genome, which can vary across millions of polymorphic sites. It is unknown the extent to which germline differences influence the somatic evolution of a tumor. Germline variation is linked to differences in the tumor immune landscape, but, canonically, immune response to epitopes derived from germline variation are damped by central tolerance. Select germline-derived epitopes, however, can escape central tolerance and illicit immune responses under specific circumstances. Interrogating 3,855 breast cancer lesions, spanning pre-invasive to metastatic disease, we demonstrate that germline variants in highly expressed and amplified genes influence somatic evolution by modulating immunoediting at early stages of tumor development. Specifically, we show that the burden of germline-derived epitopes in recurrently amplified genes selects against somatic gene amplification in pre-invasive and invasive breast cancer. As a representative example, individuals with high burden of germline-derived epitopes in ERBB2, encoding HER2, are significantly less likely to develop HER2+ breast cancer compared to other breast cancer subtypes. The same negative association between epitope burden and somatic amplification is observed for four recurrent amplicons observed in high risk of relapse ER+ breast cancers. No association was observed between somatic amplifications and epitope burden in metastatic breast cancer suggesting a subset of tumors are able to overcome immune-mediated negative selection. Tumors that do are more aggressive and demonstrate an "immune cold" phenotype. Taken together, these data show the germline genome plays a previously unappreciated role in dictating somatic evolution in breast cancer. Exploiting germline-mediated immunoediting may inform the development of biomarkers that refine risk stratification within breast cancer subtypes.

ONCOGENES OUTSIDE CHROMOSOMES

King L Hung¹, Howard Y Chang^{1,2}

¹Stanford University School of Medicine, Center for Personal Dynamic Regulomes, Stanford, CA, ²Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA

Cancer genomes undergo extensive alterations during the course of tumor evolution. These alterations often include copy number amplification of oncogenes, subsequently leading to increased dosages of tumor-driving gene products. Oncogenes are frequently amplified on extrachromosomal DNA (ecDNA), which consists of circular DNA molecules millions of bases in size that are located outside chromosomes. ecDNA molecules enable high levels of oncogene expression even when accounting for copy numbers, and their presence in tumors is associated with poor survival of patients with cancer. Despite their prevalence and functional importance in tumors, we have a limited understanding of how the genetic sequences and chromatin features of ecDNAs enable such high levels of oncogene overexpression. Using epigenomic profiling methods, imaging and multiplexed genetic perturbations, we identified unique molecular characteristics of ecDNA molecules that allow massive oncogene upregulation. We discovered that clusters of ecDNAs in the cancer nucleus enable cooperative, intermolecular oncogene activation by promoting enhancer-gene interactions. Furthermore, we successfully adapted a method, termed CRISPR-CATCH, to isolate and profile ecDNAs. This method allowed us to identify oncogene mutations specifically selected on ecDNA. We also performed haplotype variant analysis, showing evidence for excision and circularization as a mechanism of ecDNA formation in cancer. CRISPR-CATCH followed by nanopore sequencing revealed hypomethylation of gene promoters on ecDNAs, suggesting decreased DNA methylation as a potential mechanism of gene activation on ecDNAs. Finally, mapping and reconstruction of diverse ecDNA species revealed molecules containing select enhancers and those containing oncogene-coding sequences, suggesting functional specialization. Together, our studies provide insights into the origin, structural diversity and mechanism of gene activation of extrachromosomally amplified oncogenes in cancer.

MOLECULAR ARCHEOLOGY OF CANCER

Peter Van Loo^{1,2,3}

¹The University of Texas MD Anderson Cancer Center, Department of Genetics, Houston, TX, ²The University of Texas MD Anderson Cancer Center, Department of Genomic Medicine, Houston, TX, ³The Francis Crick Institute, Cancer Genomics Group, London, United Kingdom

Tumor development is driven by changes to the genome and epigenome leading to fitness advantages underlying successive clonal expansions. As somatic genetic changes occur across most or all cell cycles, the cancer genome carries an archeological record of its past. Over the past years, we have developed several approaches to mine that archeological record from the cancer genome, which we collectively call 'molecular archeology of cancer'. Using these approaches, we are able to infer the subclonal architecture of tumors, and gain key insights into the order and timing of the genomic changes that occurred over their evolutionary history. We have applied these approaches in a large-scale pan-cancer setting, showing that intra-tumor heterogeneity is pervasive across cancers, and that the timelines of tumor evolution span multiple years to decades, with typically similar key driver events occurring early.

We recently developed GRITIC (Gain Route Identification and Timing In Cancer), a novel molecular archeology of cancer approach able to time any complex copy number gains, including higher copy number states in whole-genome doubled tumors. By applying GRITIC to 6,010 primary and metastatic tumor samples from the Pan-Cancer Analysis of Whole Genomes and Hartwig Medical Foundation datasets, we find that the principle of maximum parsimony is violated in 35% of all copy number gains in whole genome duplicated tumors, with gains occurring both much earlier and later than thought under this assumption. We also find evidence for punctuated bursts of gains in WGD tumors, independent of the duplication itself. GRITIC allows for a more accurate and complete inference of evolutionary histories in different cancer types and better insights into the early copy number events in genetically unstable tumors.

HUMAN GENETICS OF ENDOCRINE-RELATED BRAIN ANATOMY USING PHENOTYPES FROM LARGE-SCALE BIOMEDICAL IMAGING

Hannah Currant^{*1}, Christoph Arthofer^{*2}, Stephen Smith², Teresa Ferreira³, Christoffer Nellaker³, Gwenaëlle Douaud², Andreas Bartsch⁴, Jesper Andersson², Margaret F Lippincott⁵, Yee-Ming Chan⁶, Stephanie B Seminara⁵, Thomas E Nichols^{2,3}, Søren Brunak¹, Frederik J Lange^{[†]2}, Cecilia M Lindgren^{[†]3,7}

¹University of Copenhagen, Novo Nordisk Foundation Center for Protein Research, Copenhagen, Denmark, ²University of Oxford, FMRIB (Oxford Centre for Functional MRI of the Brain), Nuffield Department of Clinical Neuroscience, Oxford, United Kingdom, ³University of Oxford, Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, Oxford, United Kingdom, ⁴University of Heidelberg, Department of Neuroradiology, Heidelberg, Germany, ⁵Massachusetts General Hospital, Reproductive Endocrine Unit, MGH-Harvard Center for Reproductive Medicine, Boston, MA, ⁶Boston Children's Hospital, Division of Endocrinology, Department of Pediatrics, Boston, MA, ⁷University of Oxford, Wellcome Centre for Human Genetics, Nuffield Department of Medicine, Oxford, United Kingdom

*[†] authors contributed equally

Several brain structures participate in regulation and thus dysfunction of the endocrine system. This is reflected in their morphology: from common variation such as pituitary gland volume correlating with sex-steroid concentrations; to rare variation such as olfactory bulb hypoplasia, a phenotype of Kallmann syndrome which causes infertility. Magnetic resonance imaging (MRI) can provide rich, high-dimensional data describing variation in brain morphology along the spectrum of severity.

We derived volumetric and intensity measures of the hypothalamus, pituitary gland, and olfactory bulbs from UK Biobank MRI (n=34,834). We conducted the largest genome- and exome-wide association studies (GWAS and EWAS) to date of these measures. To harness the high-dimensionality of the MRI images and assess genetic effect on fine-grain morphology, models of each structure were made for the population stratified by genotype at discovered loci.

We discovered many loci associated with volume of hypothalamus (24 loci), pituitary gland (16), and left (4) and right (6) olfactory bulbs ($p < 5E-8$). Moreover, 13 loci were associated with one or more intensity measures of the structures. Interestingly, 8 of the identified loci were sex-dimorphic ($p < 5.5E-4$) when comparing effect size in males and females. In the EWAS, 23 genes were associated with one or more brain structure volumes ($p < 5E-5$). Annotating loci found prior associations to endocrine phenotypes including testosterone levels, age at menopause and menarche.

Empowered by its unparalleled size and high-dimensional phenotype source, our study furthers understanding of genetics underlying brain structure and shows links with endocrine biology.

EVOLUTIONARY TRAJECTORIES OF COMPLEX GENOME REARRANGEMENTS IN CANCER

Jose Espejo Valle-Inclan^{*1}, Solange de Noon^{*2,3}, Katherine Trevers^{2,3}, Hillary Elrick¹, Adrienne M Flanagan#^{2,3}, Isidro Cortes-Ciriano#¹

¹European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom, ²University College London Cancer Institute, Research Department of Pathology, London, United Kingdom, ³Royal National Orthopaedic Hospital, Department of Histopathology, Stanmore, United Kingdom

* These authors contributed equally

Corresponding authors

Whole-genome sequencing studies of human tumors have revealed that cancer genomes are riddled by remarkably intricate forms of structural variants (SV), collectively known as complex genomic rearrangements (CGR). CGR occur at high frequency in some of the most aggressive cancers and are associated with drug resistance and poor prognosis. Yet, most CGR remain unexplained, hinting at the possibility of undiscovered mechanisms that might lead to novel therapeutic strategies. To elucidate the mechanisms underpinning the formation of CGR and their downstream consequences during tumor evolution, we performed high-depth ($>120x$) multi-regional short- and long-read whole-genome sequencing of hundreds of osteosarcomas, which exhibit some of the highest rates of CGR across human cancers. Our analysis revealed that whole-genome doubling and diverse types of CGR, such as chromothripsis and double minutes, are frequent clonal events that occur in most tumors. Through the integration of multi-regional WGS data we show that subclonal CGRs are also frequent events, triggering clonal diversification and rapid tumor growth suggestive of punctuated evolution. Indeed, clonal expansions triggered by the acquisition of subclonal CGR often colonize distant tumor regions, indicating that CGR act as subclonal driver events. Notably, metastatic clones often arise very early during tumor evolution (years before diagnosis) and often expand after the occurrence of CGR. Finally, we find that derivative chromosomes generated by chromothripsis events in the early stages of tumor evolution, including double minutes, acquire hundreds of subclonal SV, which increases intra-tumor heterogeneity and cancer cell plasticity. Thus, our results indicate that chromothripsis is a process that occurs throughout tumor evolution and primes the cancer genome for ongoing genomic instability, leading to SV accumulation, rapid karyotype evolution, and clonal diversification. These results have implications for intra-tumor heterogeneity and drug resistance development in cancer types driven by genomic instability.

INTERCELLULAR EXTRACHROMOSOMAL DNA COPY NUMBER HETEROGENEITY DRIVES CANCER CELL STATE DIVERSITY

Maja C Stöber^{1,2,3}, Rocío Chamorro González⁵, Lotte Brückner⁵, Thomas Conrad¹, Nadine Wittstruck^{1,5}, Annabell Szymansky⁵, Angelika Eggert⁵, Johannes H Schulte⁵, Richard P Koch⁹, Anton G Henssen^{*1,5,8,10}, Roland F Schwarz^{*4,6,1}, Kerstin Haase^{*5,7}

¹Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany, ²Institute of Pathology, Charité – Universitätsmedizin Berlin, Berlin, Germany, ³Faculty of Life Science, Humboldt-Universität zu Berlin, Berlin, Germany, ⁴Institute for Computational Cancer Biology, Faculty of Medicine and University Hospital Cologne, Cologne, Germany, ⁵Department of Pediatric Oncology/Hematology, Charité-Universitätsmedizin Berlin, Berlin, Germany, ⁶BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany, ⁷German Cancer Consortium (DKTK), partner site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany, ⁸BIH, Berlin Institute of Health, Berlin, Germany, ⁹Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, NY, ¹⁰Experimental and Clinical Research Center (ECRC), MDC and Charité Berlin, Berlin, Germany

Neuroblastoma is a paediatric cancer characterised by extensive inter- and intra-tumour genetic heterogeneity and varying clinical outcomes. One possible driver for this heterogeneity is extrachromosomal DNA (ecDNA), which segregates independently to the daughter cells during cell division and can lead to rapid amplification of oncogenes. While ecDNA-mediated oncogene amplification has been shown to be associated with poor prognosis in many cancer entities, the effects of ecDNA copy number heterogeneity on intermediate phenotypes are still poorly understood.

We leverage DNA and RNA sequencing data from the same single cells in cell lines and neuroblastoma patients to quantify ecDNA copy number variability and investigate the transcriptional effects of ecDNA-mediated oncogene amplifications including MYCN and its downstream targets. We utilise ecDNA amplicon structures to determine precise ecDNA copy number and reveal extensive intercellular ecDNA copy number heterogeneity.

Results show substantial heterogeneity of ecDNA across cells and highlight the strong dosis effect of ecDNA copy number on gene expression levels for ecDNA cargo genes. We compared high and low MYCN expressing cells within individual tumours and uncovered diverse transcriptomic profiles that affect MYCN target gene expression. Unsupervised gene set enrichment analysis of this intra-tumoural ecDNA heterogeneity uncovers a variety of enriched terms, including increased ribosome biogenesis activity in cells with high MYCN expression.

These results highlight the potential for rapid adaptability of cellular states within a tumour cell population mediated by ecDNA copy number, emphasising the need for ecDNA-specific treatment strategies to tackle tumour formation and adaptation.

PERSONALIZED TRANSCRIPTION FACTOR BINDING FROM DEEP LEARNING ALGORITHMS OFFERS A NEW FRAMEWORK TO IDENTIFY DISEASE ASSOCIATIONS

Temidayo Adeluwa¹, Saideep Gona¹, Boxiang Liu⁴, Ravi Madduri^{5,6},
Tiffany Amariuta^{2,3}, Hae Kyung Im¹

¹The University of Chicago, Genetics, Genomics and System Biology, Chicago, IL, ²University of California San Diego, Halıcıoğlu Data Science Institute, La Jolla, CA, ³University of California San Diego, Department of Medicine, La Jolla, CA, ⁴National University of Singapore, Department of Pharmacy, Singapore, Singapore, ⁵The University of Chicago, Consortium for Advanced Science and Engineering, Chicago, IL, ⁶Argonne National Laboratory, Data Science and Learning Division, Lemont, IL

Transcription factors (TFs) are important in gene regulation and in the expression of phenotypes. Altered TF binding may be the cause of human diseases (Reshef *et al.*, 2018). To find TFs that are associated with diseases, our goal is to develop personalized genetic predictors of TF binding (which is often only measured on a small number of samples) so that we can perform population-level association analysis using genome-wide analysis study (GWAS) data. Previous studies have measured the androgen receptor (AR) and forkhead box A1 (FOXA1) across prostate tumor samples of over 100 individuals (Baca *et al.*, 2022); however, this sample size is not sufficient to perform population-level disease analysis. Our aim is to generate individual-level genetic predictions for many TFs across many cell types. The paucity of population-level TF ChIP-seq data has prevented the association of genome-wide variation in TF binding across individuals with disease-associated genetic variation. Given the cost of TF ChIP-seq experiments, this would be infeasible. Here, we propose leveraging recent methods that predict epigenetic features directly from DNA sequences, such as ENFORMER, to build a model for personalized TF binding prediction. While currently available ENFORMER models are limited to only a few TFs and cell types, we utilize the idea that TF binding can be accurately characterized as a function of the genome-wide patterns of thousands of epigenetic key features as previously shown with the IMPACT method (Amariuta *et al.*, 2019). We then use ENFORMER to predict the 5,313 features used in the IMPACT framework as a function of genetic variation. Next, using IMPACT's logistic regression model we can predict genome-wide TF binding per individual as a function of individualized epigenetic features. A key advantage of our approach is that it relies on the same set of 5,313 ENFORMER models for the training of each new TF binding dataset. We validated the prediction accuracy of our strategy by comparing predictions from Baca *et al.*, who trained linear predictors with *in vivo* measured TF binding across 80 individuals, and found correlations up to 75% with a median of 45%. We plan to use our individual-level predictions of TF binding activity to interrogate the genetic basis of heritable polygenic diseases and complex traits.

HOW TO ESTIMATE CLOUD COSTS FOR GENOMICS ANALYSES

Enis Afgan¹, Keith Suderman¹, Nuwan Goonasekera², Michele Savage¹, Michael Schatz¹

¹Johns Hopkins university, Biology, Baltimore, MD, ²University of Melbourne, Melbourne bioinformatics, Melbourne, Australia

As the NIH continues to invest in establishing domain-specific cloud platforms - NHGRI AnVIL, NHLBI BioData Catalyst, NCI Cancer Research Data Commons, to name a few - there is a strong expectation that future workloads will increasingly execute on commercial cloud infrastructure providers. A critical consideration for researchers choosing to adopt such platforms is a better understanding of costs associated for running analyses, especially when running with very large sample sizes. Here, we present data and tools to help researchers analyze and monitor their cloud usage with the aim of optimizing their cloud spending.

As part of this tool suite, we provide an interactive Cloud Cost Dashboard that displays detailed benchmark data for several popular bioinformatics tools as well as historic usage of hundreds of tools from the public usegalaxy.org server. The benchmark runs were pragmatically chosen to test various tool execution options, including processor and memory scalability, different cloud providers, impact of size and type of input data on runtime and consequently cost. The data can be used to infer one's own resource requirements and associated costs with running workloads on the cloud. The dashboard also allows researchers to explore usage trends of tools on usegalaxy.org. This helps identify "staple" tools as well as "emerging" tools and help inform a decision about which tool to use for a given analysis.

In addition to insights about individual tools, the Cloud Cost Dashboard supports collection and analysis of runtime data from the execution of complete pipelines. To date, we have focused on collecting data for RNA-Seq analyses as those workloads are one of the most widely used in the community. The dashboard allows researchers to inspect resource utilization and associated costs of entire pipelines as well as individual steps within. Reasoning at the level of a complete pipeline is useful for inferring analysis costs for larger experiments that consist of many samples that need to be analyzed using the same pipeline.

Beyond the data we collected that is showcased on the dashboard, we have automated the process as open source software for running the benchmarks and collecting the run metrics. This makes it possible for researchers and consortia to reuse this framework and perform their own benchmarks in case the available data does not capture their data of interest.

Together, we will present the Cloud Cost Dashboard and the underlying framework while showcasing example use cases that translate biological experiments into estimates of costs required for analyzing produced data on the cloud.

ROLE OF CHROMATIN ARCHITECTURE OF HUMAN CELLS IN THE RESPONSE TO INFECTION BY SARS-COV-2

Saumya Agrawal¹, Kosuke Miyauchi^{1,2}, Kokoro Ozaki¹, Saera Fujiki¹, Prashanti Jeyamohan¹, Hidehiro Fukuyama^{1,3}, Fumihiko Ishikawa¹, Masato Kubo¹, Michiel de Hoon¹

¹RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan,

²National Institute of Infectious Diseases, AIDS Research Center, Tokyo, Japan, ³Kansai Medical University, Near InfraRed Photo-Immuno Research Institute, Osaka, Japan

Viral infection can induce changes in the chromatin conformation of the infected cells, resulting in rewiring of the promoter-enhancer network. In our work, we identify changes in the transcriptome and chromatin conformation in human lung epithelial cells in the first 12 hours following infection by Severe acute respiratory syndrome coronavirus 2 (Sars-CoV-2), and compare it with the response after infection by the influenza-A virus. Using CAGE expression profiling, we identify the commonalities and variations in the early cellular response of infected human cells by comparing the response after infection by five different Sars-CoV-2 strains (Original, Alpha, Beta, Delta, Gamma), revealing thousands of novel promoters and enhancers specific to Sars-CoV-2 strains. Next, we analyzed changes in the chromatin conformation of human cells to identify the targets of novel enhancers, as well as changes in the regulatory network of differentially expressed genes specific to viral infections. Based on these analyses, we will discuss the role of chromatin architecture restructuring due to viral infection progression and its effect on targets of regulatory elements.

ENRICHMENT OF NATIVE AMERICAN AND AFRICAN HAPLOTYPES FOLLOWING THE COLUMBIAN INTERCHANGE

Richard Ågren¹, Hugo Zeberg^{1,2}

¹Karolinska Institutet, Department of Physiology and Pharmacology, Stockholm, Sweden, ²Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany

The Columbian interchange brought several severe infectious diseases to the Americas from the Old World. Given the impact of the introduced communicable diseases, it has been suggested that there was a negative selection of genetic variants of Native Americans. Here, we analyze 7,612 present-day genomes of Admixed Americans and evaluate selection of Native American and African haplotypes following the Columbian interchange. We identify genomic regions, up to 10 megabases long, showing significant enrichment of Native American or African haplotypes. Moreover, we find regions depleted of European ancestry, but no regions with European ancestry enrichment. To investigate the biological associations of these enriched haplotypes, we analyze the effects on blood proteomics using two databases with 42,772 individuals. We find two independent African haplotypes enriched in the present-day genomes of Admixed Americans which are strongly associated with reduced levels of three proteins involved in T-cell function. These findings suggest that one genetic legacy of the post-Columbian admixture was enrichments of Native American and African haplotypes and depletion of less locally adapted European haplotypes.

TIGERFISH AND FISHTANK: AN OPEN-SOURCE TOOLKIT FOR OLIGONUCLEOTIDE PROBE DESIGN AGAINST REPETITIVE DNA IN EMERGING GENOMES.

Robin Aguilar¹, Conor K Camplisson¹, Chelsey Lin¹, Karen H Miga^{3,4}, William S Noble^{1,2}, Brian J Beliveau^{1,5}

¹University of Washington, Department of Genome Sciences, Seattle, WA,

²University of Washington, Paul G. Allen School of Computer Science and Engineering, Seattle, WA, ³University of California Santa Cruz, Department of Biomolecular Engineering, Santa Cruz, CA, ⁴University of California Santa Cruz, UC Santa Cruz Genomics Institute, Santa Cruz, CA,

⁵University of Washington, Brotman Baty Institute for Precision Medicine, Seattle, WA

Fluorescent *in situ* hybridization (FISH) is a powerful technique to visualize nucleic acids and learn from their spatial position in 3D. Recent technological advances have enabled the computational design of oligonucleotide (oligo) probes to interrogate >100 distinct targets within a single sample which has allowed for a greater understanding of how genome organization plays key roles in the maintenance of nuclear stability. Despite this, repetitive DNA intervals, including large arrays of satellite DNA, are typically not included as targets due to the computational complexity of ensuring that oligo probes uniquely target repeat arrays of interest. Consequently, many open questions remain unanswered about the organization and function of highly repetitive sequences. Here, we introduce Tigerfish, which is a software tool that allows for the genome-scale design of oligo probes against repetitive DNA intervals. We showcase Tigerfish by designing a panel of 24 interval-specific oligo probes spanning all 24 human chromosomes in the fully assembled CHM13 genome reference and imaging this panel on metaphase spreads and in interphase nuclei. Tigerfish also includes open-source Read the Docs with user tutorials. To maximize accessibility of future datasets, we also introduce FISHtank, which will serve as a community resource containing datasets of repetitive DNA specific oligo probes in emerging and diverse genome assemblies. This database may be used for a variety of research applications to create a more seamless workflow for identifying valuable repeat-specific oligos to plan FISH experiments. FISHtank may be used to query available genomes, learn about each probe's predicted *in silico* performance, visualize *in situ* results for select candidate probes, identify relevant literature resources for select repeat DNA intervals, and build experiments by appending desired oligo probes to a file to be downloaded. Through the development of these tools, Tigerfish and FISHtank will serve as valuable resources for performing diverse biological experiments to better understand repetitive DNA across genome builds.

MODELING THE STRUCTURE AND FUNCTION OF GENE REGULATORY NETWORKS: FROM GRAPH PROPERTIES TO EXPRESSION VARIATION

Matthew Aguirre¹, Guy Sella^{2,3}, Jonathan K Pritchard^{4,5}

¹Stanford University, Department of Biomedical Data Science, Stanford, CA, ²Columbia University, Department of Biological Sciences, New York, NY, ³Columbia University, Program for Mathematical Genomics, New York, NY, ⁴Stanford University, Department, Stanford, CA, ⁵Stanford University, Department of Genetics, Stanford, CA

Gene regulatory networks (GRNs) govern many of the core developmental and biological processes underlying human complex traits. Accordingly, major lines of scientific inquiry have been directed towards understanding the structure and function of GRNs, including broad scale experiments to characterize the effects of gene perturbations and genome-wide studies to learn the genetic architecture of gene expression. However, patterns in such measures of gene expression are influenced by key structural hallmarks of biological networks including sparsity, modular organization, and regulatory feedback loops. To better situate these data within the mechanistic context of GRNs, we propose a simple approach to simulate the structure and model the function of GRNs using techniques from small world network theory and systems biology. We generate realistic GRN structures using a generating algorithm whose parameters tune network properties of interest: sparsity, modularity, and regulatory architecture. We model gene expression regulation using a stochastic differential equation, where the steady-state expression of each gene is formulated to easily model the effect of a molecular perturbation (i.e., gene knock-down or knock-out) or genetic effects from expression quantitative trait loci (eQTLs). We use these techniques to generate synthetic gene expression data and systematically characterize the effects of gene knockouts and eQTLs, finding that structural hallmarks of GRNs tend to make networks more rather than less susceptible to perturbation. We also extend our approach to perform parameter inference, making use of simulated data and techniques from Approximate Bayesian Computation. We conclude by discussing implications of our work towards mapping the architecture of gene regulatory networks, gene expression, and other complex traits.

ROLE OF ALTERNATIVE POLYADENYLATION IN DRIVING NORADRENERGIC-TO-MESENCHYMAL TRANSITION IN NEUROBLASTOMA

Rhea Ahluwalia^{1,2,3}, Quang Trinh³, Fupan Yao², Gabrielle Persad^{1,3}, Brent Derry^{1,2}, Lincoln Stein^{1,3}

¹University of Toronto, Department of Molecular Genetics, Toronto, Canada, ²The Hospital for Sick Children, Developmental and Stem Cell Biology Program, Toronto, Canada, ³Ontario Institute for Cancer Research, Adaptive Oncology, Toronto, Canada

Neuroblastoma is the most common extracranial tumor in children, contributing to an estimated 15% of all pediatric cancer-related deaths¹. Compared to adult cancers, neuroblastoma has a distinctly lower number of somatic mutations, with only a few known drivers including MYCN amplifications, NRAS, and ALK activating mutations. There are two distinct cell states in neuroblastoma, adrenergic (ADRN) and mesenchymal (MES). These can dynamically interconvert and are thought to play a role in neuroblastoma pathogenicity as MES cells are more migratory and resistant to therapy, while ADRN cells are more proliferative. The transition between ADRN and MES cells is known as noradrenergic-to-mesenchymal transition (NMT). Despite being relatively 'quiet' tumours genetically, neuroblastoma exhibits a high level of clinical heterogeneity, ranging from a rapidly progressive disease to complete, spontaneous regression. Epigenetic mechanisms, including methylation and histone modification-based silencing, have been shown to play an important role in neuroblastoma, and a recent study has linked alternative polyadenylation (APA) to proliferation and neuronal differentiation in neuroblastoma.

Using an integrated computational and experimental approach, we are exploring whether changes in APA affects NMT in neuroblastoma. Using five established neuroblastoma scRNA-seq cell lines with ALK, NRAS, and MYCN driver mutations, we have identified distinct ADRN and MES cell populations and compared the usage of 3'UTR polyadenylation sites between these two cell populations. Additionally, we are establishing an *in vitro* model to study APA in neuroblastoma by biasing the UTR usage to either preferentially shorter or longer extremes. By doing so, we will be able to observe if globally truncated or extended UTRs can affect neuroblastoma pathogenicity by performing migration/invasion assays and performing scRNA-seq with globally shortened and lengthened UTRs. This work aims to establish a role for APA in neuroblastoma pathogenicity and progression, and explore potential targetable vulnerabilities in the APA landscape in neuroblastoma.

COMPRESSED LINEAR PANGENOME INDEXES FOR TAXONOMIC CLASSIFICATION AND GENOTYPING

Omar Ahmed, Naga Sai Kavya Vaddadi, Taher Mun, Ben Langmead

Johns Hopkins University, Computer Science, Baltimore, MD

Assembly of high-quality genomes is faster and more automated than ever before. This has spurred interest in pangenome representations that avoid "reference bias," i.e. spurious alignment penalties due to non-reference alleles. One useful pangenome representation is a graph, where variants are included as alternate paths. A disadvantage of this approach is increased sequence ambiguity, particularly when small windows contain many variants without information about how their combinations are constrained. Constructing a graph also requires an expensive multiple sequence alignment that can struggle to include all structural variation. As we look to a future with an abundance of assemblies, we need pangenome indexes that encompass all genetic variation, including structural variants, while respecting linkage disequilibrium.

We highlight a series of advances in pangenome indexing for read classification and genotyping that address these issues. These methods build on the r-index, which improves on the popular FM-index by enabling it to scale sublinearly, i.e. it grows with the amount of distinct sequence in the pangenome. Further, the r-index stores the genomes as linear strings, fully respecting linkage disequilibrium and structural variation while avoiding multiple sequence alignment. SPUMONI uses matching statistics to rapidly classify sequencing reads and it utilizes minimizer schemes similar to mdBG to "digest" and shrink the index. When analyzing a mock community, SPUMONI was 15x faster with an index 68x smaller compared to using minimap2. In the field of genotyping, recent work has documented the issues that can occur when using alignments with respect to large genomic databases to genotype. To address these issues, we proposed the new "rowbowt" software tool that rapidly genotypes sequencing datasets with respect to a pangenome panel. We show rowbowt can infer genotypes with respect to the 1000 Genomes Project panel faster and using less memory compared to graph-based approaches.

Finally, we report recent improvements that widen the applicability of these tools. Taxonomic classification is computationally difficult given the large number of taxa and high similarity among subsets of them. We developed a new data structure ("document array profiles") that allows SPUMONI to list the taxa hit by an exact match while maintaining sublinear space complexity. We can now compress this data structure by a factor of over 800x when classifying ribosomal RNA genes from the SILVA database. The compression yielded by the r-index combined with the detailed document array provide a promising new direction in classification in an era with an abundance of assemblies.

INTRATUMORAL HETEROGENEITY AND CLONAL EVOLUTION INDUCED BY HPV INTEGRATION

Keiko Akagi¹, David E Symer², Medhat Mahmoud³, Bo Jiang¹, Sara Goodwin⁴, Darawalee Wangsa⁵, Zhengke Li¹, Weihong Xiao¹, Joe D Dunn¹, Thomas Ried⁵, Kevin R Coombes⁶, Fritz J Sedlazeck^{3,7}, Maura L Gillison¹

¹MD Anderson Cancer Center, Department of Thoracic-Head & Neck Med Onc, Houston, TX, ²MD Anderson Cancer Center, Department of Lymphoma & Myeloma, Houston, TX, ³Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, ⁴Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ⁵National Cancer Institute, Bethesda, MD, ⁶The Ohio State University, Columbus, OH, ⁷Rice University, Department of Computer Science, Houston, TX

The human papillomavirus (HPV) genome is integrated into host DNA in most HPV-positive cancers, but the consequences for chromosomal integrity are unknown. Continuous long-read sequencing of oropharyngeal cancers and cancer cell lines identified an undescribed form of structural variation, “heterocateny,” characterized by diverse, interrelated, and repetitive patterns of concatemerized virus and host DNA segments within a cancer. Unique breakpoints shared across structural variants facilitated stepwise reconstruction of their evolution from a common molecular ancestor. This analysis revealed that virus and virus-host concatemers are unstable and, upon insertion into and excision from chromosomes, facilitate capture, amplification, and recombination of host DNA and chromosomal rearrangements. Evidence of heterocateny was detected in extrachromosomal and intrachromosomal DNA. These findings indicate that heterocateny is driven by the dynamic, aberrant replication and recombination of an oncogenic DNA virus, thereby extending known consequences of HPV integration to include promotion of intratumoral heterogeneity and clonal evolution.

SIGNIFICANCE

Long-read sequencing of HPV-positive cancers revealed “heterocateny,” a previously unreported form of genomic structural variation characterized by heterogeneous, interrelated, and repetitive genomic rearrangements within a tumor. Heterocateny is driven by unstable concatemerized HPV genomes, which facilitate capture, rearrangement, and amplification of host DNA, and promotes intratumoral heterogeneity and clonal evolution.

THE GENETIC REGULATION OF PROTEINS AND POST-TRANSLATIONAL MODIFICATIONS ACROSS TISSUES AND CANCER

Yo Akiyama^{*1}, Yifat Geffen^{*1}, Shankara Anand¹, Özgün Babur⁵, Meric Kinali⁵, Kisan Thapa⁵, Clinical Proteomic Tumor Analysis Consortium (CPTAC) ⁶, François Aguet^{#1,4}, Gad Getz^{#1,2,3}

¹Broad Institute of MIT and Harvard, Cambridge, MA, ²MGH, Cancer Center and Dept. of Pathology, Boston, MA, ³Harvard Medical School, Boston, MA, ⁴Present address: Illumina AI Laboratory, Illumina, Inc., Foster City, CA, ⁵UMass Boston, Boston, MA, ⁶CPTAC, Bethesda, MD

*Co-first

#Co-corresponding

The vast majority of common variants associated with complex traits and diseases, including cancer, lie in non-coding regions of the genome. While a subset of these trait-associated variants have been linked to molecular quantitative trait loci (QTLs) ranging from DNA methylation to metabolite levels, little is known about the genetic control of proteins and post-translational modifications (PTMs), particularly across tissues and in cancer. Advances in tandem mass spectrometry (MS/MS) enable comprehensive and sensitive measurements of proteins and PTMs such as phosphorylation and acetylation. Leveraging proteogenomic data of 770 tumor and 462 normal adjacent tissue (NAT) samples across 7 cancer cohorts from the Clinical Proteomics Tumor Analysis Consortium (CPTAC), we mapped the effects of common germline variants in cis and trans on proteins (pQTLs) and PTMs (ptmQTLs; for phosphorylation and acetylation) in both cancer and NAT contexts.

After strict quality controls to exclude potential artifacts in the MS/MS quantifications, we identified almost 2000 cis-pQTLs, with a majority (81%) colocalizing with an expression or splicing QTL identified in CPTAC or GTEx. Among the remaining cis-pQTLs, we identified examples that likely involve other transcriptional regulatory mechanisms, including at the 3' UTR.

Since PTM sites are modified by direct interactions with PTM regulators, we sought to identify whether cis-effects on regulators can be detected on their targets in trans. We used a combination of trans-QTL mapping and colocalization to link around 6,000 trans-ptmQTLs (~11% of measured PTM sites) to cis regulatory effects on distal genes ($\geq 5\text{ Mb}$). Only 8% of these cis-genes were known PTM regulators, emphasizing the complexity of PTM regulation. To characterize indirect effects, we leveraged protein–protein interaction and gene regulatory databases to identify the likely underlying molecular pathways. Overall, we present the first characterization of pQTLs and ptmQTLs across diverse tissue and tumor types based on MS/MS data. Moreover, we demonstrate the value of these proteomic QTLs for elucidating how genetic effects propagate across the regulatory cascade and linking these effects to complex traits and diseases.

ALLELIC SPECIFIC EXPRESSION IN BRAIN OF GENES INVOLVED IN PSYCHIATRIC DISORDER HERITABILITY AND CORTICAL THICKNESS

Nirmala Akula¹, Siyuan Liu², Shrey Shah¹, Teja N Peddada¹, Heejong Sung¹, Armin Raznahan², Francis J McMahon¹

¹Genetic Basis of Mood and Anxiety Disorders Section, Human Genetics Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, ²Section on Developmental Neurogenomics, Human Genetics Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD

Allele-specific expression (ASE) is an important mechanism by which genetic variants exert dynamic control over gene expression but has been little studied in major adult psychiatric disorders. Here we used deep RNA sequencing of postmortem brain tissue to explore the role of ASE in gene expression and genetic risk for major psychiatric disorders.

We performed a genome-wide analysis of ASE in brain by comparing deep RNA-sequencing read ratios of parental alleles in subgenual anterior cingulate cortex (sgACC) donated by 185 individuals diagnosed with bipolar disorder, schizophrenia, major depression, or no psychiatric disorder. ASE was detected by read-back phasing using both genotype and RNA sequencing information (phASER). Differential ASE between cases and controls was performed using a mixed model that takes into account sample information and correlation among single nucleotide polymorphisms (ASEP). Heritability enrichment was calculated using linkage disequilibrium score regression (LDSC). Allen Human Brain Atlas (AHBA) was used to map regional expression of dASE genes.

Of the 11,234 genes scored, 3,057 (27%) showed ASE at FDR<0.05. Partitioned heritability analysis showed that SNPs in these genes accounted for 5-7% of the heritability of BD, MDD, and SCZ (enrichment p<0.001). Between 8-13% of genes showed differential ASE (dASE), evident in some but not other diagnostic groups, including genes where the direction of ASE shifted between diagnostic groups. dASE genes clustered in genomic regions overlapping known neuropsychiatric copy number variants (CNVs); were functionally enriched in postsynaptic density, synapse, and cell junction pathways; and were most highly expressed in AHBA regions known to show thinning of the brain cortex in psychiatric patients.

This is the first study of ASE in sgACC from adults with major psychiatric disorders.

These results show that cis-regulatory variants are widespread in brain, shape gene expression in a diagnosis-related manner, and play a significant role in the heritability, neurobiology, and cortical brain thickness differences characteristic of major psychiatric disorders.

ENABLING VARIANT ANNOTATION AND DISPLAYS ACROSS MULTIPLE GENE ANNOTATED HUMAN ASSEMBLIES IN ENSEMBL

Jamie Allen, Olanrewaju Austine-Orimoloye, Gurpreet Ghattaoraya, S. Nakib Hossain, Diana Lemos, Diego Marques-Coelho, Anne Parker, Nuno Saraiva-Agostinho, Likhittha Surapaneni, Thomas Walsh, Leanne Haggerty, Stephen J Trevanion, David Thybert, Sarah E Hunt, Andrew Yates, Fiona Cunningham, Fergal J Martin

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

The creation of a reference human genome changed what was possible, enabling greater understanding of evolution, population genetics, genetic disease and drug response. The current reference is a composite created from a single individual at each location, which limits the discovery of variants in global populations. The Human Pangenome Reference Consortium (HPRC) aims to create a genome reference which better represents human diversity by combining high quality haplotype assemblies from hundreds of individuals of different ancestries. As members of HPRC, we are updating our tools and infrastructure to annotate and display variants across multiple human assemblies.

To support their adoption, we now display 95 gene annotations for the new human assemblies of which 90 carry variant annotation, including the recently published T2T-CHM13 genome, via the Ensembl Rapid Release browser. The individual annotations were constructed using a scalable and accurate approach, centred on the use of local alignments of genes and clusters of genes between the reference and target genomes.

We have lifted over the Genome Aggregation Database (gnomAD) genomic and ClinVar variant sets to display in the context of the new assemblies. We also show study-specific data sets such as the recently published chromosome Y report. The Ensembl Variant Effect Predictor (VEP) enables the annotation and prioritisation of genomic variants. We have created and released resources to enable efficient annotation of variants called against the new assemblies, including molecular consequence prediction and reporting of allele frequency and phenotype from key resources.

Here we will describe the data available in our browser, and how to run Ensembl VEP to annotate variants called against the T2T-CHM13 and HPRC genome assemblies.

SIGNALS OF EPISTATIC INTERACTIONS IN TIME SERIES GENOMIC DATA

Nathan W Anderson, Carol E Lee, Aaron P Ragsdale

UW - Madison, Integrative Biology, Madison, WI

The role of epistasis in driving adaptation has remained an unresolved problem dating back to the Evolutionary Synthesis. In particular, the question of whether epistatic interactions among genes could promote rapid, parallel evolution remains unanswered. This question has important implications in human health because epistasis is thought to be common among carcinogenic mutations. Understanding how these interactions affect speed and predictability of cancer evolution can have important implications in personalized treatments. In a recent “evolve and re-sequence” (E&R) experiment, adaptation to declining salinity in the copepod *Eurytemora affinis* exhibited surprisingly high genetic parallelism across replicate lines. Of alleles identified as targets of selection, a mean of 67% and 78% were shared between populations at generations 6 and 10, exceeding predictions under the traditional population genetic framework of a multiplicative fitness function. This result was highly discordant with a previous study on adaptation to heat tolerance in *Drosophila simulans* which, in agreement with predictions from quantitative genetic theory, found little genetic parallelism underlying adaptation to temperature change. We sought to understand why adaptation may exhibit such variable levels of genetic parallelism. Here, we employed extensive computer simulations, where we modelled various genetic trait architectures undergoing selection. We show that polygenic parallelism is highly dependent on the presence of epistasis underlying the trait. Furthermore, we find that high parallelism is consistent with positive synergistic epistasis. Our results provide theoretical support for a novel mechanism promoting rapid polygenic adaptation to a novel environmental stressor. Epistasis promoting rapid evolution may be a common phenomenon during complex traits adaptation, such as a cancer’s response to new drug regimes

POSITIVE SELECTION IN THE GENOMES OF TWO PAPUA NEW GUINEAN POPULATIONS AT DISTINCT ALTITUDE LEVELS

Mathilde André¹, Nicolas Brucato², Georgi Hudjasov¹, Vasili Pankratov¹, Danat Yermakovich¹, Rita Kreevan¹, Jason Kariwiga^{3,4}, John Muke⁵, Anne Boland⁶, Jean-François Deleuze⁶, Nicholas Evans⁷, Murray P Cox⁸, Matthew Leavesley^{9,10}, Michael Dannemann¹, Tõnis Org¹, Mait Metspalu¹, Mayukh Mondal^{*1,11}, François-Xavier Ricaut^{*2}

¹University of Tartu, Institute of Genomics, Tartu, Estonia, ²Université de Toulouse Midi-Pyrénées, Laboratoire Évolution et Diversité Biologique, Toulouse, France, ³University of Papua New Guinea, School of Humanities and Social Sciences, Port Moresby, Papua New Guinea, ⁴University of Queensland, School of Social Science, St Lucia, Australia, ⁵Social Research Institute Ltd, Port Moresby, Papua New Guinea, ⁶Université Paris-Saclay, Centre National de Recherche en Génomique Humaine, Evry, France, ⁷Australian National University, ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia, ⁸Massey University, School of Natural Sciences, Palmerston North, New Zealand, ⁹James Cook University, College of Arts, Society and EducationCairns, Australia, ¹⁰University of Wollongong, ARC Centre of Excellence for Australian Biodiversity and Heritage, Wollongong, Australia, ¹¹Christian-Albrechts-Universität, Institute of Clinical Molecular Biology, Kiel, Germany

*These authors contributed equally

Papuan highlanders have been exposed to low oxygen levels for 20,000 years, while Papuan lowlanders have encountered a unique pathogenic environment. We hypothesized that these Papuan populations carry specific signatures of positive selection to their living environment. In this study, we sequenced 128 new whole genomes and measured phenotypic data for two Papua New Guinean populations: one highlander group living between 2300 to 2700 meters above sea level (n=54) and one lowlander group living at sea level (n=74). We first applied PBS and XP-EHH scores to identify genomic regions under selection. We investigated 21 and 23 genomic regions under selection specific to Papuan highlanders or to Papuan lowlanders respectively. Next, we used CLUES to detect the SNP that most likely drives selection in each of these genomic regions (i.e. candidate SNP). Finally, we explored which phenotypes measured in our Papuan dataset are associated with the candidate SNPs. One candidate SNP from highlanders lowers the heart rate, whereas one SNP from lowlanders increases it. Moreover, both SNPs are associated with blood composition phenotypes in the UK biobank. Our results suggest that selection acted on hematological components in Papuan highlanders and lowlanders. We also found introgressed haplotypes in the genomic regions under selection. The selection signal might be driven by the introgressed archaic haplotypes in four of these regions. We suggest that archaic admixture might have acted significantly in local adaptation in Papuan populations.

MAPPING VARIANTS FROM MULTIPLEX ASSAYS OF VARIANT EFFECT (MAVEs) TO HUMAN REFERENCE SEQUENCES

Jeremy A Arbesfeld^{1,2}, Kori Kuzma², Kevin Riehle³, Julia Foreman⁴, Sumaiya Iqbal⁵, Melissa Cline⁶, Alan F Rubin^{7,8}, Alex H Wagner^{1,2}

¹The Ohio State University, Biomedical Informatics, Columbus, OH,
²Nationwide Children's Hospital, The Steve and Cindy Rasmussen Institute for Genomic Medicine, Columbus, OH, ³Baylor College of Medicine, Molecular and Human Genetics, Houston, TX, ⁴European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom, ⁵The Broad Institute of MIT and Harvard, The Center for the Development of Therapeutics, Cambridge, MA, ⁶University of California, Santa Cruz, BRCA Exchange, Santa Cruz, CA, ⁷WEHI, Bioinformatics, Parkville, Australia, ⁸University of Melbourne, Medical Biology, Parkville, Australia

The accurate clinical assessment of variant pathogenicity requires structured, computable evidence. A critical type of evidence used in this assessment is the functional impact of genetic variants derived from functional assays, but effects for individual variants may not be easily retrievable, hindering pathogenicity assessment. The Atlas of Variant Effects Alliance is working to address this problem by standardizing the structure and dissemination of multiplex assays of variant effect (MAVEs) data, but the synthetic sequences used to generate these data make translations to human genomics databases challenging. We addressed this challenge using the GA4GH Variation Representation Specification (VRS) as a standard framework for representing variants in the MaveDB database with their mapped homologs on human reference sequences.

Mapping from MAVE synthetic sequences to human endogenous sequences was accomplished via a three-step mapping process. We applied this process to 209 human score sets in MaveDB, across approximately 2.5 million protein and genomic variants. Variants in these score sets were mapped to their homologous human reference sequences and assigned unique VRS allele digests. Our method successfully mapped 99.82% of examined variants in MaveDB and highlighted key points of consideration for the remaining 0.18% including the representation of variants that span intron-exon boundaries.

Our effort enables the dissemination of MaveDB data to downstream resources. Planned integrations include the UCSC Genome Browser, ClinGen Linked Data Hub, Ensembl Variant Effect Predictor, and Broad Institute Genomics 2 Proteins Portal, demonstrating a wide range of use cases for the mapped MaveDB data. The mappings and the associated analysis notebook for this study are publicly available at https://github.com/ave-dcd/dcd_mapping.

THE DISTRIBUTION OF DISTANCES BETWEEN HETEROZYGOUS SITES IN DIPLOID SPECIES ALLOWS TO EFFICIENTLY INFER DEMOGRAPHIC HISTORY

Peter F Arndt¹, Florian Massip², Michael Sheinman³

¹Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Berlin, Germany, ²MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Paris, France, ³Sevastopol State University, Institute of Advanced Studies, Sevastopol, Russia

Heterozygous sites along a diploid genome are not uniformly distributed. On the contrary, their density varies as a consequence of recombination events, and their local density reflects the time to the last common ancestor of the maternal and paternal copy of a genomic region. The distribution of the density of the heterozygous sites therefore carries information about the history of the population size. Despite previous efforts, an exact derivation of the heterozygous sites distribution is still missing. As a consequence, estimating population size variations is difficult and requires several simplifying assumptions. Using a novel theoretical framework we are able to deduce an analytical formula for distribution of distances between heterozygous sites. Our theory can account for arbitrary demographic histories including bottlenecks and more general scenarios where population size is temporally constant during several epochs. In case the population size is constant throughout, the distribution follows a simple function and exhibits a power-law tail $1/r^\alpha$ with $\alpha=3$ where r is the distance between heterozygous sites. This prediction is nicely validated when considering heterozygous sites in individuals from African populations. Other populations migrated out of Africa and underwent at least one bottleneck which left a distinctive mark in their interval distribution between heterozygous sites, i.e. an over-representation of intervals of length from 10 to 100 kbp. Our analytical theory for non-constant population sizes reproduces this behavior and can be used to study historic changes in population sizes with high accuracy. The simplicity of our approach makes it easier to analyse demographic histories for diploid species including human, great apes, rodents and flies, requiring only a single unphased genome.

LONG-READ STRUCTURAL VARIANT BREAKPOINTS ARE ALTERED BY SMALL POLYMORPHISMS

Peter A Audano¹, Christine R Beck^{1,2,3}

¹The Jackson Laboratory, Genomic Medicine, Farmington, CT, ²University of Connecticut Health Center, Genetics and Genome Sciences, Farmington, CT, ³University of Connecticut, Institute for Systems Genomics, Storrs, CT

Structural Variants (SVs) called from read or assembly alignments against a reference genome typically rely on the alignment for breakpoint placement. Alignment methods choose optimal breakpoint locations by minimizing indels and mismatches, and small polymorphisms inside SVs or near breakpoints heavily influence the alignment. Across the 64 phased haplotypes recently released by the Human Genome Structural Variant Consortium (HGSVC), we find inconsistencies in the breakpoint locations for 882 SV insertions and 180 SV deletions not anchored in tandem repeats (TRs) or segmental duplications (SDs). We examined read-based callsets from the same sequencing data and found 1,566 insertions and 986 deletions with inconsistent breakpoints also not anchored in TRs or SDs. Using the assembled haplotypes, we find no enrichment for breakpoint differences between HiFi and CLR; however, there is a strong effect when segregating by ancestry suggesting that allelic polymorphisms drive differential breakpoint placement. We confirm that mismatches and small indels are enriched at shifted breakpoints, and because they become nested inside SV sequences, these small variants are effectively removed from the callset. We see the largest effects for variants with long tracts of breakpoint homology, such as TE-Mediated Rearrangements (TEMRs), which account for 9% of shifted SV insertions and 20% of shifted SV deletions. Tandem Duplication (TD) breakpoints are also heavily affected with 14% of TDs placed at different locations across haplotypes. When TDs are shifted, the inserted sequence is a chimera of both duplication copies and the true breakpoint is lost. Differences in breakpoint location for the same SV create difficulties for merging variants into nonredundant callsets and for annotating signatures of SV formation such as microhomology. The breakpoint inconsistencies we characterize collectively affect ~5% of the SVs called in a human genome and underscore a need for algorithm development to improve SV databases, mitigate the impact of ancestry on breakpoint placement, and increase the value of callsets for investigating mutational processes.

COPY-NUMBER VARIANTS AS MODULATORS OF COMMON DISEASE SUSCEPTIBILITY

Chiara Auwerx^{1,2}, Maarja Jõeloo³, Nicolò Tesio¹, Alexandre Reymond¹, Zoltán Kutalik²

¹University of Lausanne, CIG, Lausanne, Switzerland, ²University of Lausanne, DBC, Lausanne, Switzerland, ³University of Tartu, EGC, Tartu, Estonia

Copy number variations (CNVs) cause rare genomic syndromes but their impact on complex traits remains understudied. We called CNVs in 331'522 white, unrelated UK Biobank participants, followed by genome-wide association scans (GWASs) between the copy-number of CNV-proxy probes and 57 continuous traits and detected 131 associations. Next, we interrogated susceptibility to 60 clinical diagnoses based on four modes of CNV action and identified 70 signals involving 40 diseases. Logistic regression results were confirmed by Fisher test (40%), residual regression (32%), and time-to-event analysis (100%), the latter suggesting that CNVs cause earlier disease onset. We replicated 4 of 33 testable associations in the Estonian Biobank and observed an enrichment for nominally significant signals ($OR=5.5$; $p=2.5e-5$).

Genes encompassed by disease-associated CNVs were under stronger evolutionary constraint than genes with similar CNV frequency but no disease association ($p_{PLI}=1.0e-4$; $p_{LOEUF}=1.9e-7$). The extent of pleiotropy was dependent on the number of affected genes ($\beta=0.2$ associations/gene; $p=1.5e-5$), with the most pleiotropic region (16p11.2 BP4-5) associating with 26 continuous traits and 15 diseases. Disease associations were driven by the region's deletion (*e.g.*, renal and pulmonary disorders), except for psychiatric conditions, including schizophrenia ($OR>9.2$; $p=1.5e-8$), that were driven by its duplication.

We recapitulated known associations, *e.g.*, between the deletion of *BRCA1* and ovarian cancer ($OR>24.6$; $p=6.1e-6$) or the LDL-binding domain of *LDLR* and ischemic heart disease (IHD; $OR>7.1$; $p=5.6e-6$). Other associations were supported by colocalization with single nucleotide polymorphism (SNP)-GWAS signals (39%), overlap with related OMIM genes (21%), or CNV-biomarker associations (52%), *e.g.*, altered dosage of 17q12 (*HNF1B*), which we previously found to alter renal biomarkers, increased chronic kidney disease risk ($OR>3.4$; $p=5.9e-9$); deletion of 16p13.11, harboring the gene causing the autosomal recessive calcification disorder pseudoxanthoma elasticum (*ABCC6*), increased kidney stone risk ($OR>2.9$; $p=7.3e-5$); CNVs at 22q11.2, whose deletion is linked to congenital heart diseases, increased IHD ($OR>1.6$; $p=1.5e-7$) and aneurysm ($OR>10.0$; $p=3.2e-7$) risk.

Finally, even after correcting for GWAS signals, a high CNV load increased risk for 18 disorders, mainly through disproportionate effect of the number of deleted genes ($p=1.3e-6$). Together, these results shed light on the prominent role of CNVs in common diseases within the general population.

QUANTITATIVE TRAIT GENE DISCOVERY BY GENOME-WIDE RECIPROCAL HEMIZYGOTE SCANNING

Randi R Avery, Sheila Lutz, Frank W Albert

University of Minnesota, University of Minnesota, Minneapolis, MN

Genetic variation among individuals influences many important traits, including common human disease. Quantitative trait locus (QTL) mapping in model organisms has revealed that most quantitative traits are affected by multiple QTLs throughout the genome. However, identifying the causal genes within QTLs (quantitative trait genes; QTGs) remains challenging because most QTLs are wide and can contain dozens of genes. Experimental fine-mapping approaches typically test causality one gene at a time. This process is both laborious and potentially biased towards genes previously shown to affect the trait. An unbiased, systematic approach for direct QTG identification is advantageous.

To systematically identify QTGs for a model complex trait, we applied genome-wide reciprocal hemizygote (RH) scanning to the growth of *Saccharomyces cerevisiae* in culture. In an RH test, two genetically different strains are crossed to form a diploid hybrid. Knocking out one allele of a given gene creates a hemizygous genotype. This strain is compared to the reciprocal strain, in which the corresponding allele on the homologous chromosome is knocked out. A phenotypic difference between the reciprocal strains reveals the gene to be a causal QTG.

To apply the RH test genome-wide, we follow recent advances in interspecies hybrids by using the piggyBac transposon to mutagenize a hybrid between two genetically different *S. cerevisiae* strains to yield a large reciprocal hemizygote pool. We used Illumina sequencing of transposon insertion sites to count insertions at each open reading frame (ORF) in the pool. Out of the 4,784 ORFs that carry DNA variants between the two parental strains of the hybrid, 4,440 contained at least one insertion, with 4,260 ORFs containing at least one insertion in both alleles. This comprises ~65% of all yeast ORFs.

We grew replicates of the hemizygote pool in nutrient-rich media at 30°C for ~70 cell divisions and tracked insertion frequencies at multiple timepoints. Genes with a significant allelic difference in change in insertion frequency over time were considered candidate QTGs. Using a custom computational pipeline and linear modeling, we identified 265 genes with at least a nominally significant ($p < 0.05$) allelic effect on growth. However, validation experiments using individually engineered strains hemizygous for each of the most significant genes did not recapitulate the results of the scan. Currently, we are improving the sequencing library preparation to increase the number of insertions we can count, thereby increasing statistical power. We will also perform the scan in environments that will have a stronger selection (e.g. growth in 37°C), which could lead to stronger allelic effects that will be easier to detect. A successful approach revealing QTGs will aid in understanding how genetic variation affects important cellular traits such as growth.

AUTOMATED CANCER CELL LINE IDENTIFICATION FROM RNA-SEQ DATA

Milad Alasady, Elizabeth T Bartom

Northwestern University, Biochemistry and Molecular Genetics, Chicago, IL

Mislabeling the biological source of a sequenced sample leads to systematic errors in downstream analyses, and wastes time and resources¹. Two identical samples labeled as coming from different patients may lead to inaccurate assumptions about disease heterogeneity, while unrelated samples labeled as originating from the same patient (tumor and normal, for example) increase the variance in the dataset, making it harder to identify real signal. The standard approach to authenticating cell lines is to submit them for short tandem repeat (STR) genotyping, but this adds additional cost and is done on a separate sample. Recently, methods have been developed to authenticate cell lines using the sequencing data that is being gathered for other experimental purposes, as an internal control. These include NGScheckmate², which uses genotypes intrinsic in sequencing data to compare all samples in a data set to each other, and CeL-ID³, which compares RNA-seq genotypes from a new sample to a reference set of genotypes from the Cancer Cell Line Encyclopedia. We propose to combine aspects of both these tools, to extract genotypes rapidly from raw, unaligned human sequencing data and quickly identify matches between samples within a data set and to a reference pool of cancer cell line genotypes. The goal of this project is to quickly flag suspicious fastq files before they impact downstream sequence analyses, to ensure rigor and reproducibility in cancer studies

References

1. Lorsch JR, et al. Cell Biology. Fixing problems with cell lines. *Science*. 2014 Dec 19;346(6216):1452-3.
2. Lee S, et al. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. *Nucleic Acids Res*. 2017 Jun 20;45(11):e103.
3. Mohammad TA, et al. CeL-ID: cell line identification using RNA-seq data. *BMC Genomics*. 2019;20(Suppl 1):81. Published 2019 Feb 4. doi:10.1186/s12864-018-5371-9.

GENOME ORGANIZATION AND NONCODING RNAs
SYNERGISTICALLY CONTROL THE TIMING OF HOX GENE
TRANSCRIPTION DURING DEVELOPMENT

Philippe J Batut, Michael S Levine

Princeton University, Lewis-Sigler Institute for Integrative Genomics,
Princeton, NJ

Metazoan genes are often regulated by constellations of enhancers scattered over vast regulatory landscapes. While 3D genome organization is thought to shape long-range interactions between enhancers and promoters, it is unclear whether and how it actually determines gene expression. In addition, many regulatory sequences are transcribed into long noncoding RNAs – but how lncRNAs might modulate long-range regulation largely remains an open question. Using the *Drosophila* embryo as a model system, we showed through Micro-C analysis that genome organization is governed by two classes of regulatory sequences with opposing functions: tethering elements and boundaries. Live single-cell analysis of transcription at *Hox* gene clusters established that tethering elements foster long-range enhancer-promoter interactions, and are key to fast transcriptional activation kinetics. Conversely, boundaries prevent spurious interactions across neighboring topologically associating domains, or TADs (Batut *et al.*, *Science* 2022). Intriguingly, both a long-range enhancer and a tethering element for the fly *Hox5* ortholog, *Scr*, generate deeply conserved lncRNAs. Direct single-cell visualization of lncRNA transcription in living embryos established that enhancer transcription begins long before gene activation. Disruption of enhancer-associated lncRNA (eRNA) synthesis causes markedly precocious activation of the *Scr* gene, 35 kb away – revealing that enhancer transcription antagonizes enhancer function to control the timing of gene expression. Epistasis analysis shows that the tethering element is essential for this process, pointing to an unexpected interplay between genome organization and lncRNA function. Furthermore, many tethering elements are associated with components of the Polycomb/Trithorax epigenetic memory systems. Simultaneous disruption of the *Scr* eRNA and tethering elements is lethal, suggesting that early misexpression might lead to severe defects in the priming of chromatin-based maintenance systems. Taken together, our findings indicate that tethering elements constitute a nexus for the interplay between genome organization, noncoding transcription, and epigenetic regulation – and play a key role in determining the temporal dynamics of gene regulation.

GENETIC AND ENVIRONMENTAL CONTRIBUTIONS TO ANCESTRY DIFFERENCES IN GENE EXPRESSION IN THE HUMAN BRAIN

Kynon J M Benjamin^{1,2,3}, Qiang Chen¹, Nicholas J Eagles¹, Louise A Huuki-Myers¹, Leonardo Collado-Torres^{1,4}, Joshua M Stoltz¹, Joo Heon Shin¹, Apuā C M Paquola^{1,2}, Thomas M Hyde^{1,2,5}, Joel E Kleinman^{1,3}, Andrew E Jaffe⁶, Shizhong Han^{1,3}, Daniel R Weinberger^{1,3,5,7}

¹Lieber Institute for Brain Development, NA, Baltimore, MD, ²Johns Hopkins University School of Medicine, Department of Neurology, Baltimore, MD, ³Johns Hopkins University School of Medicine, Department of Psychiatry and Behavioral Sciences, Baltimore, MD, ⁴Johns Hopkins University, Center for Computational Biology, Baltimore, MD, ⁵Johns Hopkins University School of Medicine, Department of Neuroscience, Baltimore, MD, ⁶Neumora Inc, NA, Watertown, MA, ⁷Johns Hopkins University School of Medicine, Department of Genetic Medicine, Baltimore, MD

Introduction: Health disparities have endured for centuries. In neuroscience and genomics, individuals with recent African ancestry account for less than 5% of large-scale research cohorts for brain disorders but are 20% more likely to experience a major mental health crisis. Insights gained from genome-wide association studies (GWAS) about disease risk are promising for clinical applications. However, the lack of diversity in GWAS limits the accuracy of genetic risk prediction and hinders the development of effective and equitable neurotherapeutics.

Rationale: While diversity in GWAS has increased in recent years, population-based genetic association studies do not directly elucidate the biological mechanisms of risk variants. Recent efforts to investigate the biological impact of genetic variation on molecular traits of diverse populations have focused on improved fine mapping. These prior studies implicitly assume that the genetic mechanisms of risk and resilience are mostly conserved across ancestries; however, differences in the pathogenic role of ApoE in Alzheimer's disease between individuals of African and European genetic ancestries breaks this assumption.

Results: We examined the impact of genetic ancestry on gene expression and DNA methylation (DNAm) in admixed Black American neurotypical individuals to reduce confounding effects of ancestry-related environmental factors. Ancestry-associated differentially expressed genes (DEGs) and gene networks, while notably not implicating neurons, are enriched for genes related to immune response, the direction of effect varying by brain region. These DEGs are less evolutionarily constrained; genetic variations explain nearly 60% of expression differences. We also compared Black and White Americans, confirming most of these ancestry-associated DEGs and identifying potentially environmentally associated DEGs and differentially methylated regions. Findings at the transcript level implicate splicing and isoform processing as a major biological mechanism of ancestry variation in the brain.

Conclusion: Our results highlight how environment and genetic background affect genetic ancestry differences in gene expression in the human brain; opening new avenues to the development of ancestry-aware therapeutics and paving the way for equitable, personalized medicine.

UTILIZING ULTRA-DEEP WGS TO INVESTIGATE SOMATIC MOSAICISM IN HEALTHY TISSUE OF A BRCA VARIANT CARRIER

Gage Black, Andrew Farrell, Xiaomeng Huang, Gabor Marth

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

Somatic mosaicism is a phenomenon where genetically distinct populations of cells exist within a single organism. While this occurs naturally as individuals age, some mutations can confer an evolutionary advantage to a new cell, leading to cancer development. BRCA1 and BRCA2 are well-known tumor suppressor genes that play a crucial role in double-stranded break repair, and mutations in these genes increase the risk of developing breast and ovarian cancer. Despite extensive research investigating somatic mutations in BRCA-related cancers, little is known about the prevalence and impact of somatic mutations in the healthy tissues of individuals with germline BRCA variants.

We aim to determine whether individuals with germline BRCA variants have an increased prevalence of somatic mosaicism in healthy tissue on a whole-genome level. We collected noncancerous breast tissue, blood, and breast tumor tissue from a BRCA2 carrier undergoing a mastectomy. We sequenced each of these tissues using Illumina and PacBio HiFi whole-genome sequencing (WGS). The blood and tumor samples were sequenced to a depth of 120X using Illumina sequencing and 20X using PacBio HiFi sequencing. The noncancerous breast tissue was sequenced to a depth of 1800X using Illumina sequencing and 200X using PacBio HiFi sequencing. This sequencing data is far deeper than any publicly available WGS dataset from either technology.

Here, we illustrate many of the advantages and disadvantages of using WGS data at this depth to investigate somatic mosaicism. One such advantage is obtaining a depth necessary to differentiate real somatic variants from sequencing artifacts, which can be a significant issue when analyzing low-frequency somatic mutations. Additionally, it is possible to generate high-confidence calls by incorporating ultra-deep sequencing data from both Illumina and PacBio platforms. However, the datasets that result from sequencing to these depths are very large, making data storage and processing very difficult. Furthermore, we found that current variant calling tools are insufficient for somatic mutation calling in PacBio HiFi long-read sequencing data. We are continuing to develop new approaches to identify somatic mutations in PacBio sequencing data by leveraging the haplotype information that long-read sequencing can provide. Our approach can improve the accuracy of somatic mutation detection and contribute to a better understanding of somatic mosaicism in BRCA carriers.

m6A PATTERNS ARE CONSISTENT ACROSS DIFFERENT DROSOPHILA DATASETS

George Boateng-Sarfo¹, Sarah Signor¹, Lijuan Kan², Eric Lai²

¹North Dakota State University, Biological Sciences, Fargo, ND, ²North Dakota State University, Biological Sciences, Fargo, ND, ³Memorial Sloan Kettering Cancer Center, Developmental Biology, New York, NY,

⁴Memorial Sloan Kettering Cancer Center, Developmental Biology, New York, NY

Title: m6A patterns are consistent across different Drosophila datasets
Authors: George Boateng-Sarfo, Sarah Signor, Lijuan Kan, and Eric Lai

Abstract

Methylation of adenosine at the N-6 position (m6A) is the most common internal RNA modification in eukaryotes. It has been hypothesized to play significant biological roles including alternate splicing, RNA decay, neural function, and sex determination. However, there are many artifacts in m6A data that potentially preclude these conclusions. Here, we are developing an atlas of m6A from publicly available data. We mapped m6A genomic regions in wild-type and knocked-out samples of Whole fly, Head, Neuron, Neuroblast, Schneider cells (S2 cells), and Embryo samples. Each dataset is assessed for quality and excluded from the meta-analysis if the data collection was flawed. The included datasets are cross-referenced to identify artifacts affecting m6a calling. Using this approach we have identified a number of different patterns in m6A data that alter some of the existing conclusions about RNA modifications. First, we have found that in Drosophila m6A modifications are primarily enriched in the 5' Untranslated Regions (5' UTR). Second, we showed that m6A modification patterns do not vary significantly across samples or tissue although biological roles and mechanisms vary. We finally provide an atlas that represents the distribution of m6A across the epitranscriptome of varying tissues. In conclusion, we show that m6A is significantly enriched in 5'UTR of flies compared to mammalian cells although the methylome pattern is consistent across different datasets.

TARGETED DEEP COVERAGE EPIGENETIC PROFILES WITH SINGLE-MOLECULE AND SINGLE-NUCLEOTIDE PRECISION

Stephanie C Bohaczuk^{1,2}, Morgan O Hamm², Chang Li¹, Mitchell R Vollger^{1,2}, Anupama Jha², Benjamin J Mallory^{1,2}, Jane Ranchalis¹, Katherine M Munson², Andre Lieber¹, Andrew B Stergachis^{1,2}

¹Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA, ²Department of Genome Sciences, University of Washington, Seattle, WA

The chromatin accessibility of regulatory elements along individual chromatin fibers is essential for patterning gene expression. Recent advances in long-read sequencing allow for the precise identification of single-molecule chromatin architectures via non-specific N^6 -adenine methyltransferases (m6A-MTases) that selectively mark regions of chromatin accessibility and protein occupancy along individual DNA molecules via m6A-modified bases (Fiber-seq). We leverage Fiber-seq in combination with recent advances in machine learning and targeted high-molecular weight DNA enrichment to generate deep coverage haplotype-resolved epigenetic profiles with single-molecule and single-nucleotide precision within targeted loci of interest (Targeted Fiber-seq). Specifically, we present a targeted chromatin enrichment strategy using Cas9/pulsed-field electrophoresis to enrich 100-250kb genomic loci - allowing multiplexing of multiple samples into a single sequencing run for deep coverage epigenetic profiles with >20-fold enrichment over standard whole-genome sequencing approaches. In addition, we pair this with a convolutional neural network to identify m6A-marked bases using HiFi-kinetic information - enabling single-molecule m6A-calling with a 1,000x improvement in speed and the ability to simultaneously identify CpG methylation on the same DNA molecule. Using these comprehensive deep coverage epigenetic profiles, we dissect CTCF occupancy within cells at single-molecule and single-nucleotide resolution, demonstrating that single-molecule CTCF occupancy is heterogenous and typically marked by DNA binding of only a limited number of CTCF's 11 zinc fingers (ZF). Notably, DNA binding via ZFs 4-7 appears to be a requirement for CTCF occupancy within cells, with DNA binding via ZFs 8-11 and 1-3 occurring almost exclusively on fibers co-bound by ZFs 4-7. Finally, we combine Targeted Fiber-seq with adenine base editing of the gamma-globin promoters in primary human CD34+-derived erythroid cells - a therapeutic target for treating beta hemoglobinopathies. We find that mutations that abrogate a BCL11A binding element that is identical in both the *HBG1* and *HBG2* promoters selectively alter the chromatin accessibility of the *HBG1* and not *HBG2* promoter. Overall, Targeted Fiber-seq enables comprehensive deep coverage haplotype-resolved maps of both the genetic and epigenetic architecture of diploid organisms at single-molecule and single-nucleotide resolution.

BEYOND STAR ALLELES - GENETIC VARIANT BURDEN SCORING FRAMEWORK FOR PHARMACOGENOMICS

Malgorzata Borczyk, Jacek Hajto, Marcin Piechota, Michal Korostynski

Maj Institute of Pharmacology Polish Academy of Sciences, Laboratory of Pharmacogenomics, Krakow, Poland

Inter-individual heterogeneity in drug responses has a substantial genetic component and involves nonresponse to treatment and adverse drug reactions (ADRs). In the UK, ADRs account for around 6% of adult hospital admissions. Pharmacogenomics (PGx) aims to personalize drug treatment by including individuals' genomic information in clinical decisions. However, current PGx guidelines are limited to selected pharmacogenes and variants with experimentally confirmed phenotypic consequences. Today, as clinical sequencing becomes prevalent, a need arises for PGx frameworks that will leverage all relevant genomic information. This includes effect prediction of rare genetic variants and algorithms that allow for incorporation of multiple genes in one assay. In this study, we present a PGx variant scoring framework (PharmGScore) that overcomes multiple limitations of currently available variant-scoring approaches as it (1) assesses private, rare and common genetic variation altogether, (2) can be combined across multiple genes and (3) is not limited to known pharmacogenes. We then show the application of this framework to the 200k Whole Exome Sequences (WES) from the UK Biobank (UKB) to investigate ADRs to antidepressants.

Our approach is based on a pharmacogene-optimized variant prediction method. We introduced additional normalization steps within each gene to an exponential target function to optimize the score for effect prediction from sequencing data. We show that this normalized score effectively distinguishes no and decreased function star allele sequences from normal and increased function ones reported in PharmVar (AUC = 0.86). We then apply this framework to the WES data from the UKB and confirm its ability to effectively distinguish known normal, decreased and no function haplotypes for two pharmacogenes: CYP2C19 and CYP2D6. Finally, we focus on predicting severe ADRs to antidepressants ($n = 602$ participants) using the PharmGScore and an expanded list of 90 diverse genes that may play a role in governing antidepressant responses.

Overall our study proposes a novel paradigm to assess the compound genetic variant burden role in PGx studies that use sequencing data. It scores known pharmacogenetic variants correctly and aims to incorporate the effects of previously unreported mutations. The presented framework is an improvement of existing PGx tools and does not require star allele calling. Here we present an example use case - prediction of ADRs to antidepressants, but PharmGScore can be further developed and applied to a user-defined set of genes to investigate other pharmacological traits.

ISOFORM INSPECTOR: A JBROWSE 2 PLUGIN FOR VISUALIZATION AND ANALYSIS OF RNA-SPlicing PATTERNS.

Caroline Bridge¹, Scott Cain¹, Colin Diesh², Robert Buels², Garrett Stevens², Lincoln Stein¹, Ian Holmes²

¹Ontario Institute for Cancer Research, Computational Biology, Toronto, Canada, ²University of California, Berkeley, Computational Biology, Berkeley, CA

Genome browsers continue to be used in the field of human genetics for their usefulness in visualizing and analyzing biological information.

JBrowse 2 is one such genome browser that has uniquely positioned itself as an expandable and flexible application capable of adopting novel visualizations for nuanced cancer genomics.

Here, we present a plugin developed for JBrowse 2, the Isoform Inspector. The Isoform Inspector provides a novel visualization for viewing and analyzing alternative RNA-splicing patterns.

The Isoform Inspector features several components that serve this purpose, including a heatmap view of the read counts spanning across a given junction or mapped exon, an interactive annotations bar, an interactive gene sketch showing splice junctions and reads, and the ability to sort, cluster, and customize the view to best serve the user. The Isoform Inspector continues to be developed to further enhance the user experience.

The Isoform Inspector is compatible with both single cell sequencing and bulk sequencing derived from a cohort (e.g. a group of cancer patients), and aims to provide deeper insights related to transcript isoforms, such as the presence or absence of certain isoforms, and in what quantity amongst different cells or samples.

The utility of the Isoform Inspector will be presented through the lens of cancer data to communicate the power of the tool for interactive visualization and lightweight analysis of RNA-seq alignment data.

eQTL ANALYSIS OF CANINE TESTES SUGGESTS NOVEL GENE ASSOCIATIONS WITH MORPHOLOGICAL TRAIT LOCI

Reuben M Buckley¹, Alex C Harris¹, Susan E Lana^{2,3}, Elaine A Ostrander¹

¹National Human Genome Research Institute, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, ²Colorado State University, College of Veterinary Medicine and Biomedical Sciences, Department of Clinical Sciences, Fort Collins, CO, ³Colorado State University, Flint Animal Cancer Center, Fort Collins, CO

The role of breed-associated genetic variation in shaping canine morphological diversity is difficult to assess, as most of this variation is non-coding. To overcome this, we initiated the canine eQTL project. Bulk RNA-seq and low-pass WGS data was collected from over 140 canine testes samples and preliminary analysis on 39 samples was performed using the 750K Axiom Canine HD Array for genotyping. To address the effects of cell-type heterogeneity in bulk RNA-seq data, we used cell-type deconvolution with human single cell testes expression profiles to estimate cell-type composition of dog samples. We tested the efficacy of this approach by analyzing the well-characterized *IGF1* small body-size allele and its association with IGF1 levels in serum. We found that testes *IGF1* expression was only associated with the body-size allele when cell-type covariates were included in our models, indicating the importance of correcting for cell-type heterogeneity for detecting eQTLs. Moreover, by including cell-type covariates, we identified nearly twice as many eQTL-eGene pairs. To identify genes potentially regulating morphological traits in breeds, we performed colocalization analysis for 59 trait-associated loci. Two loci, one for height and one for weight showed suggestive evidence for eQTL colocalization. We detected a known association between a short stature allele and increased *LCORL* expression. However, this same allele was also associated with decreased expression of *SLC26A1*, a major determinant of sulfate homeostasis, which is linked to a form of intervertebral disc disorder and nephrolithiasis in humans. For the weight locus, an allele found primarily in very large breeds was associated with increased expression of *CNTROB*, a gene required for centriole duplication and cytokinesis. *CNTROB* expression drives cellular proliferation, suggesting this locus may contribute to large dog size through increased rates of cellular division. These eQTL associations provide novel hypotheses for how trait-associated variants mediate their impact. Analysis of our larger dataset will increase power to identify additional eQTLs and eGenes, and the use of low-pass WGS with imputation will improve genotype density from 750K markers to approximately 15M markers. This dataset will facilitate analysis of gene expression variability across species and throughout male canine reproductive development.

CHROMOSOME SUBSTITUTION FOR CHARACTERIZING
EPISTASIS IN THE BUDDING YEAST *SACCHAROMYCES*
CEREVISIAE

Cassandra Buzby, Federica Sartori, Mark Siegal

New York University, Center for Genomics and Systems Biology, New York, NY

Complex traits are the products of multiple genes and environmental factors, yet how these influences interact largely remains a mystery. The contribution of genetic interactions to natural trait variation is particularly challenging to estimate experimentally, and current approaches for detecting epistasis are often underpowered. Powerful mapping approaches such as bulk segregant analysis, wherein individuals with extreme phenotypes are pooled for genotyping, obscure epistasis by averaging over genotype combinations. To accurately characterize and quantify epistasis within natural trait variation, we have engineered *Saccharomyces cerevisiae* strains to enable crosses in which one parents' chromosome is fixed while the rest of the chromosomes segregate. Bulk segregant analysis then allows us to identify quantitative trait loci (QTL) whose effects depend on alleles on the fixed parental chromosome, indicating a genetic interaction with that chromosome. Using this method, we can thus identify interaction loci with high statistical power.

We demonstrate this approach in a cross of a yeast strain derived from a wine barrel (“Wine”) and a strain derived from an oak tree (“Oak”), where we obtained large pools of segregating progeny fixed for chromosome I from either Oak or Wine. We tested resistance traits by applying various selection agents to these pools and mapping the effects in each trait that do or do not depend on the parent of origin of Chromosome I. We identified multiple QTL for each trait, many of which correspond with previously identified resistance mechanisms. A subset of these QTL interact with Chromosome I, and these tend to correspond to the locations of additive QTL.

AN INTRON MOTIF-AWARE PIPELINE FOR THE ASSEMBLY OF SPLICED TRANSCRIPTS IN SPECIES OF THE NON-MODEL ORGANISM *TRICHOMONAS*.

Francisco Callejas-Hernández, Mari Shiratori, Krithika Shankar, Frances Blow, Jane M Carlton

Center for Genomics and Systems Biology, New York University, Biology, New York, NY

Conventional transcript assemblers cannot resolve splicing events in most of the non-model species, where the intron motifs and/or the splice junctions (SJs) are non-canonical. In the case of the non-model species *Trichomonas vaginalis* (the causative agent of trichomoniasis, the most prevalent sexually transmitted parasitic disease in humans), most of the aberrant SJs produced by RNAseq reads are caused by uncharacterized repeats and highly repetitive transposable elements (TEs). It is well described that the transcriptional activity of TEs can be affected by different cis-acting sequences (e.g., promoters, place of insertion, and maturation signals), and they can also be fragmented and insert themselves into or in the vicinity of genes. Therefore, transcribed TE units can be easily confounded with chimeric transcripts having putative multiple introns, an atypical intron length, and non-canonical intron motifs. Understanding these issues, we developed a pipeline to improve the assembly of *T. vaginalis* spliced genes using RNAseq data by filtering alignments containing conserved intron motifs at the SJ coordinates. The pipeline is divided into three main steps: (1) RNAseq mapping to the reference genome; (2) filtering of the BAM file by discarding the SJs missing the canonical intron motifs needed by the spliceosomal machinery (GTWBNNH(n)DBYHWNMHDYAG); and (3) assembling the transcripts using conventional bioinformatic tools. Our pipeline identified 32 new introns for a total of 64, all validated through wet-lab methods. Our customized pipeline increased the accuracy and sensitivity of the identification of spliced transcripts in the *T. vaginalis* reference strain G3 using RNAseq data by identifying the same number of introns as those identified manually. We tested our automated pipeline on the newly sequenced genomes of closely related bird parasite *Trichomonas stableri* (strains CA015840 and BTPI-3), identifying the existence of 80 introns per genome and a high conservation of the splicing motifs. These results confirm that our method of identification of spliced transcripts by using RNAseq data and custom intron motifs can be applied successfully. The pipeline and code are available at <https://github.com/biofcallejas/pysplicing>.

TOWARDS A COMPLETE CHARACTERIZATION OF HUMAN POLYMORPHIC INVERSIONS AND THEIR FUNCTIONAL EFFECTS.

Elena Campoy¹, Jon Lerga-Jaso¹, Marta Puig^{1,2}, Ruth Gómez Graciani¹, Illya Yakymenko¹, Teresa Soos¹, Alba Vilella-Figuerola¹, Ricardo Moreira¹, Alejandra Delprat¹, Marina Laplana¹, Mario Cáceres^{1,3}

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

Structural variation (SV) contributes substantially to genetic and phenotypic diversity, but the characterization of these variants is far from complete. Inversions are particularly interesting because of their effects on recombination. However, they are often missed due to their balanced nature, the repetitive sequences at their breakpoints and the fact that many are recurrent. The latest genomic techniques are finally allowing us to obtain a full catalogue of human inversions, although the number of inversions studied in detail is still limited. Here, by careful analysis of >350 predictions from different studies we have generated a reference benchmark of sequence-resolved and manually-annotated polymorphic human inversions (134) and inverted duplications (61). Moreover, each variant has been accurately genotyped in a large number of individuals from diverse populations, representing the most complete resource of this type of SVs to date. Among other things, our unique data set makes finally possible to analyse in depth the potential functional effects of inversions at multiple levels (including gene expression data from different tissues and cell lines and epigenetic signals, such as chromatin accessibility, histone modification and DNA methylation levels) and their association with disease or other phenotypic traits. As an example, a similar analysis of the well-known 8p23.1 and 17q21.31 inversions has found that the two inversions are associated to diverse phenotypes involving brain-related traits, red and white blood cells, lung function, anthropometric measures, male and female characteristics and disease risk. Also, the 17q21.31 inversion acts as lead eQTL of 29 genes located within or close to it. These findings highlight the important role that inversions can play in the human genome and that further investigation of their functional impact is needed.

TRANSCRIPTOME-WIDE CO-EXPRESSION OF SMALL NON-CODING RNAs AND GENES IN CANCER

Taylor B Cavazos¹, Aiden M Sababi¹, Jeffrey Wang¹, Alexander J Lazar², Patrick A Goodarzi¹, Hani Goodarzi³, Fereydoun Hormozdiari¹, Babak Alipanahi¹

¹Exai Bio, Computational Biology, Palo Alto, CA, ²MD Anderson Cancer Center, Department of Pathology, Houston, TX, ³UCSF, Department of Biophysics & Biochemistry, San Francisco, CA

Background: Small non-coding RNAs (smRNAs) are a diverse class of molecules with known or putative regulatory functions across human diseases. While some smRNAs are well characterized, the majority remain unannotated and have unknown biological functions. Understanding the co-expression of smRNAs with genes may provide insights into the regulatory functions of currently unannotated smRNAs. In this study, we systematically identify smRNAs co-expressed with proximal genes in tumors, investigate their tissue-specificity, and explore their possible regulatory links.

Methods: We identified smRNAs associated with changes in gene expression (smRNA-QTLs) for six major cancer types (breast, colorectal, kidney, lung, prostate, and thyroid) using smRNA-seq and RNA-seq data measured in tumors from The Cancer Genome Atlas (TCGA). We tested for associations between ~32k genes (lncRNAs and mRNAs), and ~530k smRNAs that were detected in at least 1% of samples. For each cancer, a quantitative trait loci (QTL) analysis was performed for all smRNAs, in *cis* (within a 1Mb region of each gene's transcription start site) and in *trans*, to identify the top smRNA-gene associations ($q < 0.05$).

Results: An average of 2,072 *cis*-smRNA-QTLs were identified per cancer, ranging from 7,639 in breast to 276 in prostate. Overall, there were 12,276 smRNA-QTLs for 13,325 genes with an average of 1.4 (median=1) co-expressed genes per smRNA. Among top smRNAs identified per gene, we saw high tissue specificity with 90.2% of smRNA-QTL, gene pairs uniquely identified in a single cancer. The smRNA-QTLs mapped primarily to exonic (27.8%), intronic (23.2%), and intergenic (20.1%) regions. While some represented known miRNAs (4.79%) or other smRNA biotypes (3.01%), the majority (93.2%) were previously unannotated smRNAs. Approximately 36% of smRNA-QTLs identified in *cis* were also significantly associated with a gene in *trans*. Genes with a significant co-expressed *cis*-smRNA were found to be overrepresented among known copy-number amplifications (CNA) in cancer ($\text{NES}=1.93$, $q < 1\text{e-}4$) using gene-set enrichment analysis and those smRNAs with additional *trans*-associations were 1.5 times more likely to be near a CNA.

Conclusions: We have demonstrated the ability to detect co-expression of smRNAs and genes transcriptome-wide with high-throughput sequencing data. Given the demonstrated prevalence of smRNAs in body fluids, our results highlight the potential of these molecules as harbingers of cancer-specific molecular signatures during tumor progression.

THE LANDSCAPE OF REGIONAL MISSENSE INTOLERANCE QUANTIFIED FROM 125,748 EXOMES

Katherine Chao, Lily Wang, Konrad Karczewski, Mark Daly, Kaitlin Samocha

Broad Institute, MPG, Cambridge, MA

Missense variants comprise the majority of rare coding variation in the genome and are known to contribute to human disease. However, they are notoriously difficult to interpret without functional follow-up, and most missense variants observed to date have unclear clinical significance. Building more precise tools that aid in missense variant interpretation will expand our understanding of this class of variation and improve the diagnostic rate for individuals with suspected genetic disease.

The MPC (Missense badness, PolyPhen-2, and Constraint) score was developed to incorporate information about both regional missense constraint (RMC) and variant-level metrics when predicting missense variant deleteriousness. However, the initial derivation of MPC and RMC had limited resolution due to the size of the population reference used (60,706 exomes). Here, we leverage the power of 125,478 exomes in gnomAD v2.1 to update MPC and RMC metrics for broader application in association studies and variant interpretation. Major method refinements improve the model of expected missense variation and introduce per-base resolution of constrained region breakpoints. The updated metrics will be available in the gnomAD browser, and the underlying code is open-source.

Using the updated metrics, we discover that 3,571 canonical gene transcripts harbor statistically significant regional differences in missense constraint, 52% of which had escaped previous detection. Genic regions predicted to be highly constrained (missense observed/expected ratio, or OE, ≤ 0.2) align closely with protein domains and have higher rates of ClinVar damaging variants in haploinsufficient genes as compared to unconstrained (missense OE ≥ 0.8) regions (rate ratio = 65.72; two-sided Poisson exact $p < 10^{-50}$). Initial analyses reveal a nearly 9-fold enrichment of de novo missense variants predicted to be highly deleterious ($MPC \geq 3$) in 31,058 individuals with neurodevelopmental disorders (NDD) and nearly 3-fold enrichment in 15,036 autistic individuals compared to unaffected individuals. MPC excels at differentiating between de novo missense variation in individuals with NDDs or autism vs. in unaffected individuals (odds ratio = 3.18; Fisher's exact $p = 6.93 \times 10^{-49}$), outperforming other in silico tools, including CADD and REVEL.

The improved metrics described here provide increased resolution on sub-genic missense depletion. We identify genes with RMC that were previously unappreciated and have corrected sub-genic regions erroneously identified as constrained based on the previous RMC derivation. We expect that our updated metrics will aid in variant interpretation and gene discovery efforts. Finally, larger population references will enable higher-resolution constraint calculations and enable us to extend regional constraint metrics across the coding and non-coding genome.

UNCOVERING SINGLE-CELL SPATIAL RELATIONSHIPS WITH HIGHLY MULTIPLEXED IMAGING

Erin Chung, Harald Voehringer, Anastasiia Horlova

European Molecular Biology Laboratory, Genome Biology Unit,
Heidelberg, Germany

The CODEX/PhenoCycler imaging system uses DNA-conjugated antibodies with fluorescently labelled probes which are added and removed in repeated stain-and-wash cycles to visualise up to 60 markers *in-situ*. Here, we present the analysis of multiplexed images to reveal mechanisms of B-cell non-Hodgkin's lymphoma (NHL) at spatial single-cell resolution.

We processed and analysed multiplexed images of formalin-fixed paraffin-embedded (FFPE) samples of lymph nodes from healthy patients and from patients with B-cell (NHL). Four different subtypes of B-cell NHL were present: follicular lymphoma, mantle cell lymphoma, marginal zone lymphoma, and diffuse large B-cell lymphoma (DLBCL). Preprocessing steps included nuclei segmentation, mask expansion, and single-cell fluorescent marker quantification for 56 proteins. From these data we identified 11 cell types.

Our analysis reveals disease-specific cellular interactions and tissue architecture changes in the tumour microenvironment. Most notably, we identified compositional and structural changes in follicle and germinal centre neighbourhoods, and altered cell-cell interactions between B-cell NHL subtypes. This is especially pertinent with regard to better understanding DLBCL, of which the mechanisms of pathogenesis and progression remain poorly understood. An aggressive subtype which represents a large percentage of NHL cases and often progresses quickly, identifying spatial biomarkers of DLBCL mechanisms is essential to advancing precise diagnosis and targeted treatments.

DIFFERENTIAL CELL-TYPE-SPECIFIC GENE EXPRESSION BY TYPE 2 DIABETES STATUS IN HUMAN SKELETAL MUSCLE

Dan L Ciotlos¹, Sarah C Hanks¹, Arushi Varshney², Michael R Erdos³, Nandini Manickam², Anne U Jackson¹, Heather M Stringham¹, Narisu Narisu³, Lori Bonnycastle³, Markku Laakso^{4,5}, Jaakko Tuomilehto⁶, Timo A Lakka⁴, Karen L Mohlke⁷, Michael Boehnke¹, Heikki A Koistinen⁶, Francis S Collins³, Stephen C J Parker^{2,8}, Laura J Scott¹

¹Univ of Michigan, Biostatistics, Ann Arbor, MI, ²Univ of Michigan, Bioinformatics, Ann Arbor, MI, ³NIH, NHGRI, Bethesda, MD, ⁴Inst Clinical Medicine, Internal Medicine, Kuopio, Finland, ⁵Kuopio Univ Hospital, Medicine, Kuopio, Finland, ⁶Finnish Inst for Health & Welfare, Public Health & Welfare, Helsinki, Finland, ⁷Univ of North Carolina, Genetics, Chapel Hill, NC, ⁸Univ of Michigan, Human Genetics, Ann Arbor, MI

Increased insulin resistance in skeletal muscle characterizes people with type 2 diabetes (T2D). In bulk skeletal muscle, we identified differences in gene expression between individuals with and without T2D; however, the effect of cell-type composition on our results is unknown. To understand T2D's effect on gene expression in cell types, we performed single nucleus RNA-sequencing of skeletal muscle biopsied from 282 Finnish adults (41.5% female, 25.9% T2D) from the FUSION Tissue Biopsy Study. Nuclei clustered into 13 cell types, including muscle fiber types (1, 2A, and 2X).

We tested for associations between cell-type composition and glucose and insulin-related traits using a negative binomial model. Individuals with higher fasting glucose or insulin had greater proportions of neuromuscular junctions (min p=0.0020). Individuals with higher 2-hour OGTT glucose had higher proportions of fast twitch type 2X muscle fibers (p=0.0010) and lower proportions of slow twitch type 1 muscle fibers (p=0.0012).

For each cell type, we tested for associations between gene expression and T2D using a negative binomial model (n genes tested=23,766). In each muscle fiber type, we found <15 genes differentially expressed by T2D status. However, using gene set enrichment analysis, we observed 6.4-10.1% of gene sets were enriched for different levels of gene expression by T2D status across muscle fiber types (n gene sets tested=5641). We observed consistent directions of enrichment of genes involved in catabolic processes across muscle fiber types, such as higher expression of cellular amino acid catabolic process in individuals without T2D than those with T2D (OR =0.51-0.65, min p=6.4e-10). In each muscle fiber type, genes associated with cellular respiration were enriched for higher expression in individuals without T2D than those with T2D (OR=0.51-0.65, min p=6.7e-6).

Overall, muscle fiber cell types show concordant enrichment of expression differences between individuals with and without T2D. This indicates T2D-associated differences in gene expression in catabolic and cellular respiration processes observed in bulk RNA-seq are detected in all muscle fiber types, rather than restricted to a specific one.

PREDICTING AND SPATIALLY LOCALISING BULK RNA-SEQ FROM HISTOLOGY ACROSS 39 HEALTHY HUMAN TISSUES

Francesco Cisternino¹, Soumick Chatterjee¹, Adam P Levine², Craig A Glastonbury¹

¹Human Technopole, Population and Medical genomics, Milan, Italy,

²University College London, Research Department of Pathology, London, United Kingdom

Accurate prediction of gene expression variation from healthy human Whole Slide Image (WSI) histology has yet to be described. Capturing the spatial resolution of gene expression is fundamental to better understanding gene function and involvement in biological and pathological processes. With the aim of spatially localising genes quantified in bulk RNA-seq across a wide variety of tissues, we developed a deep multiple instance learning model, called *RNAPath*. We trained tissue specific instances of *RNAPath* across a total of 13735 H&E WSI with paired RNA-seq from the Genotype Tissue Expression consortium (GTEx). Using *RNAPath*, we can predict RNA levels for thousands of genes and lncRNAs across diverse human tissues using WSI alone, providing interpretable, spatially resolved expression maps. We validate *RNAPath*'s ability to correctly spatially resolve the expression of individual genes, by demonstrating concordance for *CNN1* (smooth muscle), *PLIN1* (adipocytes) and *KRT5* (skin epidermis) predictions, with independent immunohistochemical (IHC) staining. Furthermore, we assess *RNAPath*'s predictions by comparison to ground truth RNA-levels for each gene (artery: 5183 genes, $r=0.56$; colon: 11119 genes, $r=0.43$; oesophagus mucosa: 10169 genes, $r=0.42$; skin: 10768 genes, $r=0.23$) and benchmark *RNAPath* against the state of the art method, *HE2RNA*, demonstrating a $2.9\times$ improvement ($r=0.56$ vs $r=0.19$, $n=5183$ genes), with 2043 genes predicted at $r>0.6$ (vs 232) from artery WSI. Many arterial samples in GTEx have significant peripheral adipose tissue attached to them. As a positive control, we demonstrate that *RNAPath* correctly predicts known adipocyte specific genes (e.g. *ADIPOQ*, *PLIN1*, *LIPE*) to be spatially restricted to adipose tissue present in artery samples (0.0139 FDR 5%). Having validated predictions from *RNAPath* using IHC and known tissue-specific genes, we sought to investigate *RNAPath*'s ability to derive spatial expression signatures for localised tissue pathologies in which molecular processes are poorly characterised. Arterial calcification, a characteristic of progressing atherosclerosis, is present in 255 GTEx samples (calcified tissue $> 1\%$). We found 37 genes (>2 -fold enrichment) that were spatially restricted to areas of calcification. These genes were significantly enriched for immune cells and inflammation, e.g. CD8+ T-cells ($p\text{-value} = 3.4 \times 10^{-5}$ FDR 5%). In summary, we introduce *RNAPath*, a state of the art multiple instance learning regression model able to predict the spatial expression of genes in healthy human histology.

PROCAPNET: DISSECTING THE CIS-REGULATORY SYNTAX OF TRANSCRIPTION INITIATION WITH DEEP LEARNING

Kelly Cochran¹, Melody Yin², Anshul Kundaje^{1,3}

¹Stanford University, Department of Computer Science, Stanford, CA, ²The Harker School, San Jose, CA, ³Stanford University, Department of Genetics, Stanford, CA

While many aspects of mammalian Pol II transcription initiation have been extensively characterized, our understanding of the DNA sequence determinants of initiation remains incomplete. Although overrepresented TF motifs in promoters have been identified, we still lack a base-resolution mapping of precisely how every sequence feature influences TSS positioning and promoter activity. A third of human promoters contain no known initiation motifs, and in promoters with known motifs, how those sequence features modulate transcription to be more bidirectional vs. unidirectional, or focused at one TSS vs. dispersed across many, is poorly characterized. We understand even less about transcription initiation at enhancers.

To address these knowledge gaps, we trained a deep learning model to learn the mapping between DNA sequence and transcription initiation, measured genome-wide at base resolution by PRO-cap experiments. The model accurately predicts both the exact locations of TSSs and the amount of initiation observed at them, and model performance is consistent across promoters and enhancers and within promoter classes (housekeeping and TCT promoters). We then applied a model interpretation framework to identify important sequence features for the model's predictions, obtaining a high-sensitivity collection of motifs relevant to transcription initiation that includes both core promoter sequence features (TATA box, Inr, BRE) and several other TF motifs (i.e. GABPA, NF-Y, CTCF, and ZBTB33).

To characterize the association between these identified sequence features and initiation, we performed *in silico* mutational experiments using our model. Our simulations suggest that the sequence features driving initiation are highly epistatic: many motifs play unique, specialized roles in combination with other motifs. Results recapitulated known spacing constraints between promoter features (e.g. the -30bp TATA box-Inr spacing) and revealed previously unreported constraints, including for periodic spacing of some TF motifs. We also identified motifs, such as NRF1 and YY1, that can function as initiation sites. Through systematic motif ablation, we quantified the contribution of all motifs to both TSS positioning and the rate of initiation, including "profile signatures" that suggest redistribution of initiation as a function of motif presence. Finally, we compared the sequence determinants of initiation across different classes of promoters, and between promoters and enhancers; our results support a unified, sequence-based model of transcription initiation genome-wide.

THE BATTLE OF THE SEXES IN HUMANS IS HIGHLY POLYGENIC

Jared M Cole¹, Peter R Golightly¹, Carly B Scott¹, Mackenzie M Johnson³,
Jedidiah Carlson^{1,2}, Matthew J Ming¹, Arbel Harpak^{1,2}, Mark Kirkpatrick¹

¹University of Texas at Austin, Department of Integrative Biology, Austin, TX, ²University of Texas at Austin, Department of Population Health, Austin, TX, ³Fred Hutchinson Cancer Center, Computational Biology Program, Public Health Sciences Division, Seattle, WA

Sex-specific selection, which occurs when the fitness effects of alleles differ between males and females, can have a profound impact on the maintenance of genetic variation, fecundity, and disease risk in natural populations.

Because the sexes mix their autosomal genomes each generation, quantifying the intensity of sex-specific selection has been difficult using conventional population genetic methods. Here, we introduce a novel method for estimating the strength of contemporary sex-specific selection that builds on subtle differentiation in haplotype structure between the sexes and apply it to haplotype data from 250K individuals in the UK Biobank. Though we find weak-to-undetectable sex-specific selection at any given individual locus, we uncover highly polygenic signals of sexually-antagonistic selection on both viability and reproductive success. Our method further allows us to decompose signals into selection on viability and fecundity. We find sexually-antagonistic viability selection favors in males, and disfavors in females, alleles that increase red blood cell count, BMI adjusted waist-to-hip ratio, and sex hormone binding globulin (SHBG). Fecundity selection favors alleles in males that increase forced vital capacity, waist circumference, waist-to-hip ratio, weight, BMI, and arm fat-free mass. Taken together, our findings marry the underwhelming evidence in human data to date with the long standing theoretical expectation of pervasive sex-specific selection.

DECODING THE INTERCELLULAR SIGNALS UNDERLYING HUMAN MICROGLIA AND ASTROCYTE PLASTICITY

Natacha Comandante-Lou¹, Masashi Fujita¹, Gilad S Green², David A Bennett³, Naomi Habib², Vilas Menon¹, Philip L De Jager¹

¹Center for Translational & Computational Immunology, Department of Neurology, Columbia University Medical Center, New York, NY, ²Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel, ³Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL

The cellular diversity of the human brain and the intricate crosstalk among cells of the neocortex contextualize the multicellular ecosystem that determines brain functions. Glia, most notably microglia and astrocytes, play major roles in shaping the microenvironmental context of the brain. As part of the innate immune system, microglia and astrocytes have been implicated in triggering cascades of neuroinflammatory events associated with neurodegeneration, and recent single-cell studies have identified a set of seemingly distinct microglia and astrocyte states linked to different events along the trajectory to Alzheimer's Disease (AD) and other forms of neurodegeneration. In particular, our analyses of single nucleus data from 424 brains in the Religious Order Study and Rush Memory and Aging project (ROSMAP) have prioritized Astrocyte subtype 10 (Ast.10) as mediating the effect of tau pathology on cognitive decline.

The origins of glial cells' remarkable plasticity, which enables them to occupy diverse cell states, remain a mystery. Moreover, the contexts underlying their cellular plasticity are likely intertwined, given by the ample evidence showing microglia and astrocytes secrete signals that alter each other's functions. Here, we performed a system-wide analysis to identify cell-state specific intercellular signals underpinning human microglia- and astrocyte-state plasticity.

Integrating cell-cell signaling inference and data-driven modeling, we analyzed over 86,000 microglia and 228,000 astrocyte transcriptomes from the 424 ROSMAP cohort. Using NicheNet, which computationally infers upstream ligands from target genes of interests based on prior knowledge, we identified 21 ligands and associated receptors which putatively regulate the Ast.10 gene signature in the cortex. Using partial least squares regression modeling, we further showed that the donor-to-donor variability in Ast.10 frequency can be sufficiently explained, based on the expressions of merely eight of the ligands or eight of the receptors prioritized by NicheNet. Ligand-receptor pairs that are among these top predictors in both the ligand and receptor models, such as SEMA4D-PLXNB1 and EFNA1-EPHA4, are well known astrocyte signaling axes associated with synapse assembly and transmission, and with AD neuropathologies. Independent validation of 240 ROSMAP donors is ongoing, as are a validation effort using a spatial transcriptomic platform MERSCOPE and in vitro validation. In all, our study provides a framework to systematically identify the signaling origins of human glial cell plasticity, revealing interdependences among glial subsets and prioritizing pharmacological targets to shift glia towards specific cellular states to restore homeostasis.

THE FAT-TAILED DUNNART GENOME REVEALS CIS-REGULATORY DRIVERS OF DEVELOPMENTAL HETEROCHRONY BETWEEN MARSUPIALS AND MAMMALS

Laura E Cook^{1,2,3}, Charles Y Feigin², Davide M Vespasiani^{2,3}, Andrew J Pask², Irene Gallego Romero^{2,3}

¹Lawrence Berkeley National Laboratory, Environmental Genomics & Systems Biology, Berkeley, CA, ²University of Melbourne, School of Biosciences, Melbourne, Australia, ³University of Melbourne, Melbourne Integrative Genomics, Melbourne, Australia

Marsupials and placental mammals exhibit significant differences in reproductive and life history strategies. Marsupials are born highly underdeveloped after an extremely short period of gestation, while placental mammals undergo substantial embryonic development during an extended pregnancy. The short *in utero* window in marsupials requires prioritization of the development of structures critical for post-birth survival in the pouch. Of particular interest is the accelerated development of craniofacial structures compared to placentals (an example of heterochrony), to allow post-birth suckling. The fat-tailed dunnart (*Sminthopsis crassicaudata*) is a small Australian marsupial mouse that has one of the shortest gestations of any mammals, being just 13.5 days long, and gives birth to one of the smallest and most embryonic young. For these reasons, the dunnart presents an extreme example of craniofacial heterochrony in marsupials. Cis-regulatory elements (CREs) play a central role in morphological divergence. By combining genome comparisons of the mouse and dunnart with functional data for the enhancer-associated chromatin modifications, H3K4me3 and H3K27ac, we investigated divergence of craniofacial cis-regulatory landscapes in these species. We found that the majority of craniofacial CREs in each species were not functionally conserved in the other species, and that active CREs were associated with distinct sets of genes. In particular, CREs in the dunnart were significantly enriched around genes that exhibit increased expression in the mouse embryonic face from E10.5-E15.5, supporting that the CREs regulate genes involved in embryonic craniofacial ossification. This included critical craniofacial developmental genes such as the master bone growth regulator, *Runx2*, as well as *Bmp6*, *Mef2c* which have previously been suggested to contribute to vertebrate craniofacial diversity. Using mouse-dunnart comparisons, we also identified dunnart-specific CREs active near genes enriched for development of mechanosensory structures in the facial epidermis. Accelerated development of the dunnart sensory system likely relates to the sensory cues received by the nasal-oral region during the distinctive postnatal journey to the pouch and uncovers an intriguing marsupial-specific example of heterochronic adaptation. Heterochrony is a fascinating evolutionary process found in many parts of the animal kingdom, and our study highlights the power of marsupial-placental comparative genomics for understanding the role of CREs in driving temporal shifts in evolution and development.

A PHENOTYPIC PATIENT MATCHING ALGORITHM TO IMPROVE DIAGNOSES IN RARE DISEASE COHORTS

Isabelle B Cooperstein¹, Alistair Ward^{1,2}, Gabor Marth¹

¹University of Utah School of Medicine, Human Genetics, Salt Lake City, UT,

²Frameshift Genomics, Inc., Boston, MA

Background The genome sequence of patients with rare monogenic diseases often lacks a clearly identifiable disease-causing genetic variant or mutation. In many cases, such patients remain undiagnosed despite the application of numerous computational variant prioritization tools. Identifying phenotypically and genotypically similar patients can be essential for diagnosing these cases and help significantly advance knowledge of a rare disease. Available prioritization tools do not focus on patient-to-patient similarities within existing datasets, necessitating the development of novel tools to increase diagnostic rates. Here, we have developed an approach that matches undiagnosed patients to phenotypically similar diagnosed patients and then utilizes the known diagnostic genes of these patient matches to prioritize variants in the undiagnosed case.

Methods The hierarchically organized Human Phenotype Ontology (HPO) terms used to describe disease phenotypes enable the phenotypic similarity of pairs of patients to be calculated. These patient-to-patient similarity scores are highest for the most clinically similar patients, without requiring that they share exact phenotypes. We then use these scores to identify the most similar diagnosed patients to an undiagnosed patient, build a list of diagnostic genes, and then search for plausible diagnostic variants within these genes for the undiagnosed case.

Results We have applied our methods to the difficult-to-diagnose patient cohort within NIH's Undiagnosed Diseases Network (UDN). This patient cohort is characterized by diverse and complex phenotypes across a large variety of genetic disorders. To validate our methods, we first show that patients with matching clinical diagnoses exhibit significantly higher phenotypic similarity scores compared to patients with different clinical diagnoses. Further, we have shown that patients with matching genetic diagnoses score significantly higher phenotypic similarity scores compared to patients with different genetic diagnoses. We have applied these methods to search for candidate diagnostic variants in currently undiagnosed patients, and have identified multiple cases for whom this procedure identifies compelling candidates. We are currently working to validate these candidates in collaboration with the diagnostic team at the University of Utah UDN clinical site.

Conclusion Our methods enable the prioritization of candidate variant lists of undiagnosed patients based on their phenotypic similarity to previously diagnosed cases. Initial successes suggest that this will be an exciting and valuable tool for diagnostic variant prioritization for new patients under study as well as patients left undiagnosed in projects such as the UDN. This algorithm will be available as an easy-to-use web tool accessible to any member of the clinical team.

DIRECT CONVERSION OF PEDIATRIC ALL TO AML AFTER CAR-T CELL AND BLINATUMOMAB THERAPY

Tim Coorens¹, Grace Collord², Taryn Treger^{3,4,5}, Stuart Adams², Emily Mitchell^{3,4}, Barbara Newman⁴, Gad Getz^{1,6,7}, Anna Godfrey⁴, Jack Bartram², Sam Behjati^{3,4,5}

¹Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA,

²Great Ormond Street Hospital for Children, London, United Kingdom,

³Wellcome Sanger Institute, Hinxton, United Kingdom, ⁴Cambridge

University Hospitals NHS Foundation Trust, Cambridge, United Kingdom,

⁵University of Cambridge, Pediatrics, Cambridge, United Kingdom, ⁶MGH, Pathology, Boston, MA, ⁷Harvard Medical School, Pathology, Boston, MA

Children with acute lymphoblastic leukemia (ALL) undergoing anti-CD19 directed therapy can develop phenotypically distinct acute myeloid leukemia (AML). The origin of such class switch leukemias remains unresolved. The AML can be an independent cancer induced by treatment, a descendant of a precursor clone shared with the ALL, or the result of transdifferentiation of the ALL. These scenarios can be distinguished using somatic mutations.

Here, we reconstructed the phylogeny of multiple leukemias in a child who, following an initial diagnosis of B-ALL, relapsed twice over nearly five years. For the second ALL relapse, the patient received anti-CD19 cellular and antibody treatment and subsequently developed AML. Whole-genome sequencing to 100X coverage each of eight samples at six disease time points revealed that the AML was a direct descendant of the initial ALL. The AML exhibited the same V(D)J recombinations as the ALL and its relapses, confirming its previous state as a B cell and the direct conversion to a myeloid cell type.

Strikingly, none of the leukemia relapses directly descended from one another, suggesting a reservoir of persistent clones as the source of disease progression. Disease stages were characterized by distinct driver mutations, detectable prior to clinical emergence. All relapses derived from a subclone with double loss of CDKN2A and PAX5, caused by off-target effects of the V(D)J recombination machinery. Loss of PAX5 is known to enable B-cell transdifferentiation *in vitro* and likely underpins the lineage switch.

Mutational signature analysis shows the AML lineage diverged from the ALL lineage during the chemotherapy for the initial ALL, nearly 5 years prior to CAR-T and blinatumomab treatment. Therefore, the anti-CD19 treatment imposed a strong selection pressure in favor of a pre-existing AML lineage, with its drivers detectable long before its emergence.

This study showcases the highly dynamic nature of cancer genome evolution under pressure of various treatments. More importantly, our findings highlight that complete genomic monitoring of primary childhood leukemias and relapses is essential to predict therapy resistance, especially in the context of anti-CD19 treatment.

DEEP WHOLE-GENOME SEQUENCING OF GTEx TISSUES REVEALS DEVELOPMENTAL PATTERNS AND SOMATIC EVOLUTION

Tim Coorens¹, Danielle Firer¹, Oliver Priebe¹, Julian Hess¹, Gad Getz^{1,2,3}, Francois Aguet⁴, Kristin Ardlie¹

¹Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA,

²Massachusetts General Hospital, Pathology, Boston, MA, ³Harvard Medical School, Pathology, Boston, MA, ⁴Illumina, Inc., Artificial Intelligence Laboratory, San Diego, CA

From fertilization onwards, the cells of the human body continuously experience DNA damage and accumulate somatic mutations. The somatic genome of a cell is a record of its life history: mutations shared with other cells indicate a shared ancestry and can be used to retrace early development. Early somatic mutations have also been linked to various disease phenotypes, such as childhood cancer and developmental disorders.

Recent studies on a few donors have used somatic mutations to show that embryonic cells contribute asymmetrically to the adult body, such that one daughter cell of the zygote has twice as many descendant cells as the other, likely due to cellular bottlenecks in embryogenesis. It is unclear how variable this pattern is across the human population.

Here, we sequenced 308 whole genomes derived from six tissue types of 55 donors within the GTEx cohort to a mean depth of 195x. These tissue types span the embryonic germ layers: ectoderm (brain and skin), mesoderm (heart) and endoderm (esophagus, thyroid and lung). Facilitated by the breadth in sampling and depth of sequencing, we can efficiently detect embryonic mutations in all donors and estimate the contribution of embryonic progenitor cells to the donors in our cohort. We observe a highly variable asymmetry in zygotic daughter cell contribution, hinting at a stochastic bottleneck during early embryogenesis. We model that the major source of this asymmetry is the early split between trophectoderm and the inner cell mass, but further bottlenecks during gastrulation may modulate contributions to specific germ layers.

Our analysis also detects large clonal expansions in thyroid, esophagus and skin, each harboring distinct imprints of known mutagenic processes. The severity of these expansions is highly variable, with one notable clone in skin carrying over 20,000 somatic single-base substitutions. Most of these expansions can be explained by somatic mutations in genes known to be under selection, such as NOTCH1.

Taken together, this study reveals patterns of embryonic development and later somatic evolution from deep whole-genome sequencing data of normal tissues and will serve to provide context to diseases rooted in abnormal development. As costs for whole-genome sequencing continually decrease and technologies advance, the approach outlined here will substantially increase our understanding of human cellular evolution in health and disease.

GENETIC ARCHITECTURE OF IMMUNE CELL DNA METHYLATION IN FREE-RANGING RHESUS MACAQUES

Christina E Costa¹, Marina M Watowich², Rachel M Peterson², Elisabeth A Goldman³, Kirstin Sternier³, Michael J Montague⁴, Michael Platt⁴, Josue E Negron-Del Valle⁵, Daniel Phillips⁵, Lauren J Brent⁶, James P Higham¹, Noah Snyder-Mackler⁵, Amanda J Lea^{2,7}

¹NYU, Anthropology, NY, NY, ²Vanderbilt University, Biological Sciences, Nashville, TN, ³University of Oregon, Anthropology, Eugene, OR, ⁴University of Pennsylvania, Neuroscience, Philadelphia, PA, ⁵ASU, Life Sciences, Tempe, AZ, ⁶University of Exeter, Psychology, Exeter, United Kingdom, ⁷CIFAR, Azrieli Global Scholars, Toronto, Canada

A major goal in evolutionary biology is understanding genotype-phenotype relationships and the heritable basis of complex traits. Gene regulatory changes, like DNA methylation, may link genotypes to phenotypes, yet these links are rarely studied in natural primate populations relevant for understanding human evolution. We used reduced representation bisulfite sequencing to measure DNA methylation at 555,856 CpGs in peripheral whole blood of 573 free-ranging rhesus macaques (*Macaca mulatta*) on the island of Cayo Santiago, Puerto Rico. We used allele-specific methods to map cis methylation quantitative trait loci (meQTL) testing for effects of 243,389 single nucleotide polymorphisms (SNPs) on local DNA methylation. Of 776,092 tested SNP-CpG pairs, we identified 516,213 meQTL (FDR 5%). meQTL explain an average 21.2% of methylation variance, significantly more than demographic factors. CpGs with sex or age effects are also more likely to be associated with an meQTL, and therefore genetically influenced. meQTL are most often found in functional regions like promoters, enhancers, and open chromatin, and explain more methylation variation than meQTL in inactive regions. meQTL SNPs are also enriched in transcription factor (TF) binding sites of the methylation-sensitive ETS family, suggesting genotype-dependent TF-binding might link QTLs to nearby methylation. Overall, our results indicate widespread proximal genetic effects on inter-individual DNA methylation profiles, primarily in active regions where methylation is likely to impact gene expression and downstream phenotypes. In support, meQTL were enriched for associations with complex traits in humans, such as immune, cardiovascular, and neurological risk phenotypes. Furthermore, with a dataset of 172 mRNA-seq samples 332 were confirmed as cis expression QTL (eQTL) in the same population. Of 966 SNP-CpG-Gene trios, 4% showed methylation-expression correlations. Macaque meQTL-eQTL genes are enriched for immune response functions, like antigen presentation and inflammation. These may be candidate loci for studies using macaques as models for human health, and more generally points to the biological mechanisms driving regulatory variation in our close relatives.

DETECTION OF GENETIC AND EPIGENETIC ALTERATIONS DRIVEN BY LOSS OF TET PROTEINS AT SINGLE BASE RESOLUTION

Hugo Sepulveda¹, Robert Crawford², Fabio Puddu², Yang Liu², Ankita Singhal², Gary Yalloway², Helen Sansom², Jens Fullgrabe², Nikolay Pchelintsev², Lidia Prieto-Lafuente², Audrey Vandomme², Philippa Burns², David M Morley², Rosie Spencer², Páidí Creed², Joanna D Holbrook², Anjana Rao¹

¹La Jolla Institute for Immunology, Division of Signaling and Gene Expression, San Diego, CA, ²Cambridge Epigenetix, Cambridge, United Kingdom

The three TET (Ten-Eleven Translocation) dioxygenases promote cytosine demethylation by successive rounds of oxidation of the methyl group of 5-methylcytosine (5mC). The balance between DNA methylation and demethylation influences many biological processes including epigenetic regulation, development, and oncogenesis. Indeed, TETs act as known or putative tumour suppressor genes, and TET2 loss-of-function mutations are frequently observed in hematopoietic malignant and pre-malignant syndromes.

We previously generated TET inducible triple knockouts (Tet iTKO) mouse embryonic stem cells (m(ESC)) and showed that acute TET deletion leads to genome-wide loss of 5-hydroxy-methyl cytosine (5hmC) and an altered gene expression profile that results in increased genome instability and aneuploidy. However, this work did not compare methylation and hydroxymethylation levels genome-wide at single-base resolution. Existing single-base resolution methods for measuring methylation either conflate mC and hmC or only measure one of the two, limiting their applicability in analyses where changes in both markers are of interest. Six-letter seq is a single base-resolution sequencing methodology that sequences complete genetics and cytosine modifications in a single workflow, separately reading 5mC and 5hmC (and unmodified cytosine) at each CpG in the genome.

Here we use six-letter seq to determine the changes in 5mC and 5hmC induced by simultaneous deletion of TET genes in mESC, genome-wide, and at single-base resolution. We show how deletion of TET genes alters the distribution of 5hmC and 5mC in different genomic compartments. The ability to simultaneously detect genetic and epigenetic information allows us to explore the interaction of differential cytosine methylation and genomic sequence in a single dataset.

SIMULTANEOUS MEASUREMENT OF GENETICS AND EPIGENETICS ENABLES NEW BIOLOGICAL INSIGHT

Nicholas J Harding¹, Páidí Creed¹, David Currie¹, Casper K Lumby¹, David M Morley¹, Fabio Puddu¹, Jean Teyssandier¹, Michael Wilson¹, Jens Fullgrabe¹, Audrey Vandomme¹, Aurel Negrea¹, Alexandra Palmer¹, Phillipa Burns¹, Shirong Yu¹, Diljeet Gill^{2,3}, Aled Parry^{2,3}, Wolf Reik^{2,3}, Joanna D Holbrook¹

¹Cambridge Epigenetix Ltd, Cambridge, United Kingdom, ²Altos Labs, Cambridge, United Kingdom, ³Babraham Institute, Cambridge, United Kingdom

DNA comprises molecular information stored in genetic and epigenetic bases, both of which are vital to our understanding of biology. The interaction of genetics with the DNA epigenome plays a causal role in cell fate, ageing, response to environment and disease development. Methods widely used to detect epigenetic DNA bases do not distinguish unmodified cytosines and thymine, therefore fail to capture common C-to-T mutations and thus capture incomplete genetic information. Five-letter seq is a single base-resolution sequencing methodology that sequences complete genetics and cytosine modification in a single workflow.

Five-letter seq generates high quality genetic and epigenetic information, even from low DNA input, enabling the identification of genetic variants and quantification of modified cytosine levels in a single experiment. The phased nature of the technology, whereby genetic and epigenetic information is available jointly at read-level, enables the study of genetic and epigenetic co-variation. For example, allele-specific methylation (ASM), whereby differential methylation patterns are observed between heterozygous variants. We identify ASM across the genomes of all 7 Genome-in-a-bottle samples. We go on to show that a substantial increase in the degree of ASM is associated with successful “maturation phase transient reprogramming” (MPTR) whereby the transcriptome and epigenome of fibroblasts from middle-aged donors are rejuvenated about 30 years. No such increase in ASM is associated with fibroblasts that were treated by MPTR but failed to rejuvenate. This work demonstrates not only that ASM can be directly identified using five-letter sequencing, but that ASM is associated with cellular ageing and function.

DETECTING SOMATIC LINE-1 RETROTRANSPOSON INSERTIONS IN SINGLE NEURONS

Michael S Cuoco^{1,2,3}, Meiyang Wang¹, Rohini Gadde², Iryna Gallina¹,
Reicardo Jacobini⁴, Daniel R Weinberger⁴, Jennifer A Erwin⁴, Eran A
Mukamel², Apua Paquola⁴, Fred H Gage¹

¹Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, CA,

²University of California, San Diego, Computational Neural DNA

Dynamics Lab, Department of Cognitive Science, La Jolla, CA, ³University
of California, San Diego, Bioinformatics and Systems Biology Graduate
Program, La Jolla, CA, ⁴Johns Hopkins University, Lieber Institute for
Brain Development, Baltimore, MD

Long interspersed nuclear elements (LINE)-1 (L1) retrotransposons represent a family of mobile genetic elements distinguished by their ability to autonomously mobilize in the human genome via a copy-and-paste mechanism. While undetectable in most healthy tissues, L1 insertional activity has been found in the brain, resulting in somatic mosaic insertions. Many have speculated that somatic insertions may alter cellular phenotypes, disrupt neuronal circuits, and contribute to neurological disease. However, reported L1 insertion rates are largely inconsistent due to limited sensitivity and specificity of insertion profiling methods. Here we employ Somatic L1-associated Variant sequencing (SLAV-seq) to survey L1 insertions in over 4000 single neurons from the prefrontal cortex and hippocampus of 14 healthy donors and 14 donors with schizophrenia. We use a machine learning approach to accurately classify and remove amplification artifacts from the sequencing data, enabling us to characterize the somatic L1 insertion burden of each cell. With somatic L1 insertion calls, we plan to characterize the rate, insertional preferences, and added schizophrenia risk of somatic L1 insertions in human neurons.

RETROCOPIES: GENE COPIES IDENTIFIED IN VERTEBRATES AND INVERTEBRATES' GENOMES AND THEIR ORTHOLOGY.

Helena B da Conceição^{1,2}, Rafael L Mercuri^{1,2}, Matheus P Castro¹, Daniel T Ohara¹, Gabriela Guardia¹, Pedro F Galante¹

¹Hospital Sírio Libanês, Molecular Oncology Center, São Paulo, Brazil,

²University of São Paulo, Institute of Mathematics and Statistics, São Paulo, Brazil

Retrocopies are copies of mRNAs that are reverse-transcribed into the genome as a result of LINE1 activity. They are characterized by the conservation of only their parental exons, the frequent presence of poly(A) tail, and lack of their parental promoter regions. These characteristics have been used to identify retrocopies since the 1980s when many human duplicated genes were first reported. However many advancements in biology and bioinformatics have made it possible to systematically search for them. In 2003, the first wide list of processed pseudogenes was generated (HOPPSIGEN), and nowadays, we have a handful of databases, such as RCPedia, Pseudogene.org, and RetrogenesDB. In this current work, we developed a novel pipeline to identify retrocopies that can be confidently applied to different species. The pipeline aligns the full set of coding mRNA sequences against the reference genome, selects alignments based on size and distance from the gene that originated the mRNA, searches for exon-exon junctions from the top 3' most exons, and eliminates candidates composed mainly of repetitive elements. The pipeline has been applied successfully to 44 organisms ranging from mammals to invertebrates. Importantly, we have built a pipeline to determine the orthology of these retrocopies between themselves. Briefly, we retrieve the genomic sequence around the retrocopy (3000bp up- and downstream) and perform pairwise alignment using Lastz. We subsequently use filters of coverage and identity to determine the best match for each retrocopy. We find that, on average, 30% of the retrocopies identified are conserved within most vertebrates (mammals and birds). Specifically, humans have 96,5% of their retrocopies conserved with other species, mainly other primates. In general, 82% of primate retrocopies are conserved - mostly with other primates. Surprisingly, most retrocopies of rodents are mainly species-specific, with only 10% of retrocopies shared with other species. For other mammals, we find that on average 30% of retrocopies have orthologs, but prominent cases are observed, such as the sloth that has more than 96% of its retrocopies being species-specific. In summary, we have developed a novel and robust approach to identify retrocopies and have, in addition, made a major contribution to the study of the evolutionary history of these species. This is a substantial improvement, for humans and other mammals, in the knowledge of these understudied gene copies.

Support: FAPESP 2018/13613-4

EXPLORING THE REGULATORY LANDSCAPE OF CANCER ADAPTION IN ACIDIC EXTRACELLULAR MATRIX

Yifan Dai^{*1,2}, Arnaud Stigliani^{*1,2}, Jiayi Yao^{*1,2}, Renata Ialchina³, Dominika Czaplinska³, Stine F Pedersen³, Albin G Sandelin^{1,2}

¹Section for Computational and RNA biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark, ²Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark, ³Section for Cell Biology and Physiology Department of Biology, University of Copenhagen, Copenhagen, Denmark

A dense and acidic extracellular matrix (ECM) is a hallmark of pancreatic ductal adenocarcinoma (PDAC). Recent research has demonstrated a causal relationship between acidic microenvironment adaptation and cancer progression, as PDAC cells that have adapted to acidic ECM exhibit aggressive phenotypes in proliferation, metastasis, and chemotherapy resistance. However, it is still largely unclear how ECM and acidity interplay during the adaptation. To decompose those effects, we conducted in-vitro adaptation experiments using murine pancreatic cancer (Panc02) spheroids in different 3D cultures. In each culture, treatment groups were adapted to gradually decreasing pH to 6.7 in 5 weeks, whereas the corresponding control groups were passaged in constant physiological pH of 7.4. We applied Cap Analysis of Gene Expression (CAGE) to measure the transcription intensity of promoters and enhancers as a function of pH and ECM conditions. Our results demonstrate that acid-adapted spheroids exhibit depleted activity in EGFR and INF pathways, but enhanced cell cycle and metabolism, indicating that adaptation may occur through the inhibition of inflammatory machinery and the activation of alternative growth pathways. These changes in transcriptional regulation largely persist even after acid-adapted cells return to a physiological pH. We also find that the regulatory mechanism of adaptation to extracellular acidity is partially dependent on the existence of ECM. Through TF activity regression, we identified FOX and KLF TF families as potential regulators of acid adaptation in ECM-free and ECM-rich microenvironments, respectively. Overall, our study provides important insights into the mechanisms underlying the adaptation of pancreatic cancer cells to an acidic, dense extracellular microenvironment. The decomposition of microenvironmental factors allows for a better understanding of causal relationships, which may ultimately benefit future clinical research.

DIVERSITY AND REPRESENTATION OF SOUTH ASIAN GENOMES

Arun Das, Michael C Schatz

Johns Hopkins University, Computer Science, Baltimore, MD

The rapid growth in genomics has not been uniform across the full range of human diversity, with the vast majority of widely used genomic data generated from just a handful of populations. This leaves the majority of the world's populations poorly represented, and has resulted in systemic biases that can have serious impacts on downstream analysis. In this work, we aim to utilize recent advancements in sequencing and assembly to better catalog the variation present in South Asian populations.

We conducted a pilot study focusing on ten individuals from the 1000 Genomes Project (1KGP), with two from each of the five 1KGP populations of South Asian origin. Using high quality short read data from 1KGP, we investigated the variation between these individuals relative to both GRCh38 and T2T-CHM13. To do this, we follow a similar pipeline used in the creation of the African Pan-Genome (Sherman et al. 2019) and align reads from these individuals to the chosen reference, and then assemble unaligned and/or poorly aligned reads into contigs. We then attempt to place the larger contigs in the reference genome, allowing us to identify variants and novel sequences.

Using CHM13 versus GRCh38, we observe slightly higher average alignment rates (98.0% vs 97.5%), and approximately 50% fewer large contigs assembled from the unaligned reads. Nevertheless, CHM13's assembled contigs still contain 6-9 Mbp of sequence per individual, highlighting widespread population-specific sequence missing. When placing these contigs against the references using pair-end alignments, we achieve more successful placements against CHM13 than against GRCh38. The placed contigs are evenly distributed across the chromosomes, and overlap a range of biologically significant regions.

We similarly evaluated the recently released draft human pan-genome references from the Human Pangenome Reference Consortium (HPRC), built from 47 individual genomes plus GRCh38 or CHM13. We observe higher alignment rates (+0.3-0.6%) in pangenomes than in their corresponding linear reference, and find that the contigs assembled from unaligned reads are a subset of those assembled from the corresponding linear references. Interestingly, we still find 5-8 Mbp of sequence per individual was assembled from unaligned reads, including contigs that are over 60Kbp in size, many of which still overlap biologically significant regions.

We are currently working to expand this pilot study for all individuals of South Asian descent in the 1KGP, with initial results showing similar trends for alignment rates and the sizes of assembled contigs. We also plan to extend this to include other relevant genomic datasets as well as RNA-seq data from South Asian individuals to investigate the overlap between placed contigs and clinically relevant variants. Overall, we hope that these analyses will improve clinical sequencing and diagnostics in South Asian populations.

MULTI-MODAL ASSESSMENT OF FUNCTIONAL IMPACT OF MUTATIONS ON THE GENOME

Maitreya Das^{1,2}, Deepro Banerjee^{1,3}, Jiawan Sun^{1,2}, Saie Mogre³, Ayaan Hossain², Adam Glick^{1,4}, Santhosh Girirajan^{1,3}

¹Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA, ²Pennsylvania State University, Molecular, Cellular and Integrative Biosciences program, University Park, PA,

³Pennsylvania State University, Bioinformatics and Genomics program, University Park, PA, ⁴Pennsylvania State University, Department of Veterinary and Biomedical Sciences, University Park, PA

Fundamental principles governing how genetic mutations cause ripple effects through changes in gene expression and regulatory activity in the genome are not completely understood. We tested enhancer activity of 253,632 fragments covering 46,142 ChIP-seq sites of enhancer marks and 211 transcription factors. We also deleted six master regulators and developmental TFs (*ATF2*, *CTCF*, *FOXA1*, *LEF1*, *TCF7L2* and *SCRT1*) using CRISPR/Cas9 and quantified the downstream effects on enhancer activity and gene expression using UDI-UMI-STARR-seq and RNA-seq. First, we found 2,653 fragments that showed enhancer peaks in wild type, indicating high non-functional binding of TFs. Next, we compared enhancer activity in the deletion lines to wild-type and using logistic regression, classified all fragments as induced or repressed in response to the deletion (logistic regression; average precision score >0.7). For example, we observed that 12863/20987 fragments in *ATF2* and 6803/10145 in *LEF1* lines lost activity, 29948/35245 in *FOXA1* and 9036/9365 *TCF7L2* lines gained enhancer activity. In fact, our results identified 2024 inactive fragments that were induced in response to the deletion across all lines. Analysis of binding motifs found that fragments repressed in *ATF2* and *LEF1* lines and those induced in *FOXA1* were enriched for TP53 family motifs. We further incorporated STARR-seq data to the Activity By Contact model to connect responsive enhancers to target genes, and correlated differential enhancer activity to target gene expression. Finally, we applied our approach to assess the impact of the neurodevelopment disease-associated 16p12.1 deletion, and discovered 3278 repressed and 2052 induced fragments that trigger a diverse set of downstream functional effects associated with the disease. Overall, our results show a paradigm for quantifying the downstream effect of a mutation and provide a conceptual framework for understanding regulatory networks.

DISCOVERY OF TYPE 2 DIABETES GENES USING AN ACCESSIBLE TISSUE

David Davtian¹, Theo Dupuis¹, Dina Mansour-Aly², Naeimeh Atabaki-Pasdar³, Mark Walker⁴, Paul W Franks³, Femke Rutters⁵, Hae Kyung Im⁶, Ewan R Pearson¹, Martijn Van de Bunt⁷, Ana Viñuela⁸, Andrew A Brown¹

¹University of Dundee, Population Health and Genomics, Dundee, United Kingdom, ²Lund University, Genomics, Diabetes and Endocrinology, Lund, Sweden, ³Lund University, Genetic and Molecular Epidemiology, Lund, Sweden, ⁴Newcastle University, Faculty of medical Sciences, Newcastle, United Kingdom, ⁵Amsterdam University Medical Center, Epidemiology and Biostatistics, Amsterdam, Netherlands, ⁶University of Chicago, Medicine-Genetic Medicine, Chicago, IL, ⁷University of Oxford, Wellcome Centre for Human Genetics, Oxford, United Kingdom, ⁸Newcastle University, Bioscience Institute, Newcastle, United Kingdom

In the context of studying complex diseases, identification of the genes mediating the activity of GWAS variants with methods such as Transcriptome-wide Association Studies (TWAS) has been limited by the relative scarcity of data on the genetic effects on molecular traits, particularly from disease relevant tissues often difficult to sample. Using multiple reference datasets (DIRECT consortium [N = 3029], GTEx consortium [N = 70-706] and InsPIRE [N = 420]), GWAS summary statistics for T2D (DIAGRAM consortium) and a TWAS method (MetaXcan), we identified 1,818 unique genes associated with T2D. Comparing the performance of different reference datasets, we found that sample size, and not the relevance of the tissue to the disease, was the critical factor in identifying disease genes (404 genes from whole blood compared to 3 to 299 from multiple tissues, including pancreatic islets). Genes from well powered reference datasets were more likely to have multiple lines of genetic evidence (31 genes validated by MR in blood, compared to 1 to 9 from GTEx tissues). Moreover, we show that accounting for BMI reduced the number of associated genes by ~ 30% across all tissues, suggesting that many GWAS links to T2D may be mediated by BMI. Finally, using GWAS studies of more specific characterised subtypes of T2D, we uncovered genes directly relevant to that subtype, such as *LST1*, an immune response gene for Severe Autoimmune Diabetes and *TRMT2A*, involved in beta cell apoptosis, for Severe Insulin Deficient Diabetes, but only using the well powered, whole blood reference. Our work demonstrates that well powered gene expression reference panels can identify disease relevant genes, even when the tissue is not directly relevant for the disease. Moreover, it shows the added value of well-defined disease subtypes when studying complex diseases involving multiple tissues, and how a well powered molecular reference can compensate for an underpowered GWAS.

BIOCHEMICAL ACTIVITY IS THE DEFAULT DNA STATE IN EUKARYOTES

Ishika Luthra, Xinyi E Chen, Cassandra Jensen, Abdul Muntakim Rafi,
Asfar Lathif Salaudeen, Carl G de Boer

University of British Columbia, School of Biomedical Engineering,
Vancouver, Canada

Genomes encode for genes and the regulatory signals that enable those genes to be transcribed, and are continually shaped by evolution. Genomes, including those of human and yeast, encode for numerous regulatory elements and transcripts that have limited evidence of conservation or function. Here, we sought to create a genomic null hypothesis by quantifying the gene regulatory activity of evolutionarily naïve DNA, using RNA-seq of evolutionarily distant DNA expressed in yeast and computational predictions of random DNA activity in human cells and tissues. In yeast, we found that >99% of bases in naïve DNA expressed as part of one or more transcripts. Naïve transcripts are sometimes spliced, and are similar to evolved transcripts in length and expression distribution, indicating that stable expression and/or splicing are insufficient to indicate adaptation. However, naïve transcripts do not achieve the extreme high expression levels as achieved by evolved genes, and frequently overlap with antisense transcription, suggesting that selection has shaped the yeast transcriptome to achieve high expression and coherent gene structures. In humans, we found that, while random DNA is predicted to have minimal activity, dinucleotide content-matched randomized DNA is predicted to have much of the regulatory activity of evolved sequences, including active chromatin marks at between half (DNase I and H3K4me3) and 1/16th (H3K27ac and H3K4me1) the rate of evolved DNA, and the repression-associated H3K27me3 at about twice the rate of evolved DNA. Naïve human DNA is predicted to be more cell type-specific than evolved DNA and is predicted to generate co-occurring chromatin marks, indicating that these are not reliable indicators of selection. However, extreme high activity is rarely achieved by naïve DNA, consistent with these arising via selection. Our results indicate that evolving regulatory activity from naïve DNA is comparatively easy in both yeast and humans, and we expect to see many biochemically active and cell type-specific DNA sequences in the absence of selection. Such naïve biochemically active sequences have the potential to evolve a function or, if sufficiently detrimental, selection may act to repress them.

UNIDIRECTIONAL EXPRESSION OF ENHancers WITH CELL TYPE-DEPENDENT DIRECTION OF TRANSCRIPTION

Emi Kanamaru, Yoriko Saito, Fumihiko Ishikawa, Michiel J de Hoon

RIKEN, Center for Integrative Medical Science, Yokohama, Japan

Enhancers are genomic regulatory elements that can affect expression of genes over megabases of genomic distances. Enhancers are not transcriptionally silent, but produce capped RNAs known as enhancer RNAs (eRNAs) in a cell type-specific manner, with enhancer transcription associated with the regulatory activity of the enhancer.

As transcription at enhancers has been observed to occur in a bidirectional fashion, bidirectional transcription has been used as a signature feature for genome-wide enhancer identification. However, as enhancer RNAs are rapidly degraded by the exosome, their abundance is typically low, making it difficult to assess whether enhancer transcription is genuinely bidirectional.

Here, we analyze enhancer expression directionality in very deep CAGE (Cap Analysis Gene Expression) data from the 6th edition of the Functional Annotation of the Mammalian Genome consortium project (FANTOM6). With more than 1 billion mapped CAGE tags in a single cell type, this data set allows the expression directionality of each enhancer to be calculated accurately. Our analysis of the FANTOM6 and other deep CAGE data sets revealed that enhancers have a clear and highly significant preference for unidirectional transcription.

Next, we constructed a beta-binomial model to reliably analyze enhancer directionality in 808 CAGE libraries from FANTOM5, with on average 5 million mapped CAGE tags per library. We found predominantly unidirectional enhancer expression in 547 libraries, predominantly bidirectional enhancer expression in 110 libraries, and no significant preference for unidirectional or bidirectional enhancer expression in 151 libraries.

In contrast, expression of enhancers was highly significantly bidirectional when aggregating CAGE tag counts across libraries, suggesting that the preferred direction of transcription of an enhancer switches between libraries. Applying our model to the FANTOM5 set of enhancers revealed 20,956 switching enhancers, which are unidirectionally expressed but with a cell type-dependent direction of transcription, 14,275 enhancers with unidirectional expression consistently in the same direction, and 7,240 enhancers that were consistently bidirectionally expressed. To compare the biological relevance of the three categories of enhancers, we calculated the GWAS SNP enrichment levels for each category, and found them to be similar to each other.

UNDERSTANDING THE CONTEXT-DEPENDENT ROLE OF PBAF-SPECIFIC SUBUNIT PBRM1 IN CHROMATIN REGULATION AND CANCER

Alisha Dhiman, Emily Dykhuizen

Purdue University, Medicinal Chemistry and Molecular Pharmacology,
West Lafayette, IN

Mammalian SWI/SNF or BAF are multi-subunit complexes that regulate cell-type specific gene expression by regulating chromatin accessibility, in coordination with transcription factors. PBRM1 is the complex defining subunit of PBAF class of SWI/SNF complexes. It is mutated in ~40% of clear cell renal cell carcinoma cases but also associated with cancer progression and therapy resistance, indicating context-dependent function in cancer. It uniquely has six tandem bromodomains, unlike any other mammalian bromodomain containing protein, which suggests it could act as the targeting subunit of PBAF complex. Bromodomains are ~110 amino acid protein modules which bind to acetyl-lysine, a histone modification associated with open chromatin. Individual bromodomains of PBRM1 have been shown to bind to H3K14Ac, a histone mark associated with the promoters of inducible genes in mammals. We have previously reported that loss of PBRM1 in normal epithelial cells results in a partial Epithelial-to-Mesenchymal Transition (EMT). Further probing on this observation, we observed a requirement of PBRM1 function for metastasis *in vivo*. In this study, we have utilized epigenomic and transcriptomic approaches to understand the mechanistic role of PBRM1 in EMT-regulated gene expression, and its relationship to genome-wide H3K14Ac histone modification. Using RNA-seq and ATAC-seq, we identified PBRM1-dependent genes regulating cell polarity, cell-cell junctions, and cell migration in EMT-inducing conditions. ChIP-seq studies using a PBAF-specific subunit revealed a predominantly promoter localization for PBAF under both basal and EMT-inducing conditions. Using PBAF genome occupancy at PBRM1-dependent genes, in combination with HOMER motif analysis, we further identified ATF3 as the transcription factor important for regulation of these genes. H3K14Ac was observed at both promoters and distal intergenic regions in unstimulated conditions. However, under EMT-inducing conditions, there was a significant increase in the number of promoters marked by H3K14Ac. There was a high correlation between PBAF genomic localization and H3K14Ac along with H3K4Me3 histone mark under these conditions. A subset of these genes was dependent on PBRM1 for expression, spanning processes like cell differentiation and regulation of transcription. This suggests that PBRM1 can target PBAF to genomic loci either by recognizing acetylation on histones or possibly on transcription factors. Deciphering the role of PBRM1 and the genomic context of its function can present targeting opportunities by small molecule bromodomain inhibitors.

RHINOVIRUS INFECTED EPITHELIAL CELLS DRIVE GENETIC SUSCEPTIBILITY TO CHILDHOOD-ONSET ASTHMA

Sarah Djeddi^{1,2,3}, Daniela Fernandez-Salinas^{1,2,3,4}, George Huang^{5,6}, Chitrasen Mohanty⁷, Christina Kendzierski⁷, Joshua Boyce^{5,6}, James Gern⁸, Nora Barrett^{5,6}, Maria Gutierrez-Arcelus^{1,2,3}

¹Boston Children's Hospital, Boston, MA, ²Boston Children's Hospital, Boston, MA, ³Broad Institute of MIT and Harvard, Boston, MA,

⁴Undergraduate Program in Genomic Sciences, Cuernavaca, Mexico,

⁵Harvard Medical School, Boston, MA, ⁶Brigham and Women's Hospital, Boston, MA, ⁷University of Wisconsin-Madison, Madison, WI, ⁸University of Wisconsin School of Medicine and Public Health, Madison, WI

Asthma is a complex disease caused by genetic and environmental factors. Genome-wide association studies (GWAS) have identified hundreds of genetic variants contributing to asthma susceptibility. Studies that identify the cell-types mediating genetic risk for complex diseases are lacking in epithelial cells. Furthermore, epidemiological studies for asthma indicate that wheezing illnesses with rhinovirus (RV), which is the most frequent cause of common cold, increases the risk of developing asthma in children. Rhinovirus infects airway epithelial cells (AECs), and these cells are known to have roles in type 2 inflammation.

Here we hypothesized that particular cell states of AECs may play a role in mediating genetic susceptibility to asthma. We compiled bulk and single-cell transcriptomic and epigenomic datasets of AECs that were exposed to different stimuli such as pro-inflammatory cytokines and viruses. We applied methods that use GWAS summary statistics from 4 different cohorts that assessed multiple asthma endotypes and associated diseases to characterize AEC states that mediate genetic risk. We used a single-cell disease-relevance score (scDRS) that identifies single cells over-expressing GWAS genes in a weighted manner. We also used Linkage Disequilibrium Score-regression in Specifically Expressed Genes (LDSC-seg) to identify cell-state specific annotations with enrichment of heritability.

First, using immune cell datasets we validated that T cells mediate significant genetic risk to asthma. Then, we found that RV-infected AECs significantly mediate genetic susceptibility to childhood-onset asthma.

Single-cell data indicates non-ciliated epithelial cell subsets are the main mediators, specifically at 24 and 42 hours post infection. Furthermore, our data suggest that influenza virus-infected AECs may mediate genetic risk to asthma, while Sars-Cov2-infected AECs do not. Finally, we found that RV-infected AECs from asthmatic patients showed a stronger enrichment for asthma risk compared to healthy individuals.

Overall, our results suggest that part of the “missing regulatory effects” for childhood-onset asthma are hidden in RV-infected non-ciliated epithelial cells and in epithelial cell states of patients with asthma.

THE GENETIC ARCHITECTURE OF ADAPTIVE PIGMENTATION TRAITS IN SWORDTAIL (*XIPHOPHORUS*) FISHES

Tristram O Dodge¹, Daniel L Powell¹, John J Baczenas¹, Theresa R Gunn¹, Shreya M Banerjee^{1,2}, Manfred Schartl^{3,4}, Molly Schumer¹

¹Stanford University, Biology, Stanford, CA, ²U.C. Davis, Evolution and Ecology, Davis, CA, ³University of Wuerzburg, Developmental Biochemistry, Wuerzburg, Germany, ⁴Texas State University, Xiphophorus Genetic Stock Center, San Marcos, TX

Background: Variation in pigmentation patterns have long interested evolutionary biologists as striking examples of phenotypic diversification, and is a tractable system to understand molecular genetics and development. Teleost fish provide an exciting complement to studies of pigmentation in mammals, due to their increased repertoire of 5 pigment cell types and the diversification of key pigmentation genes stemming from their whole-genome duplication 350 million years ago. Swordtail fish in the genus *Xiphophorus* are polymorphic for a dizzying array of pigmentation traits that have been under long-term balancing selection, providing unique opportunities to characterize the genetic architecture and evolution of pigmentation.

Results: We leverage long-read sequencing, genome-wide association studies, and ATAC-seq to uncover the genetic basis of multiple melanic pigmentation traits across 2 *Xiphophorus* clades.

- In one *Xiphophorus* clade, we find the false gravid spot—a male-limited sexual mimicry polymorphism—is controlled by variation at a 30 kb complex structural rearrangement upstream of *kit-ligand a*. By quantifying allele-specific expression in hybrids, we discover tissue-specific cis-acting regulatory changes are responsible for increased *kit-ligand a* expression in false gravid spot tissue.
- We map two other pigmentation ornaments in this clade—the peduncle edge and the caudal blotch tailspot—to *kit-ligand a*. The fine resolution of our GWAS peaks, in combination with ATACseq data, indicate the caudal blotch tailspot maps to a narrow region upstream of the structural variant associated with the false gravid spot, possibly representing an additional tissue-specific enhancer.
- In another clade, we find that five distinct tailspot patterns map to two linked loci: *kit-ligand b* and *wnt7ba*. Through association mapping analyses, we discover *wnt7ba* controls tailspot presence versus absence and distinct *kit-ligand b* alleles control tailspot pattern.

Our results suggest that ancient *kit-ligand* paralogs underpin the convergent evolution of pigmentation patterns in two *Xiphophorus* clades. In one clade, distinct mutations in the *kit-ligand a* regulatory region underpin multiple phenotypes in different tissues, whereas two linked pigmentation genes, *wnt7ba* and *kitlgb* interact to generate 5 distinct tailspot patterns in a second clade.

Conclusions: These results represent a remarkable case of convergent evolution, where ancient *kit-ligand* paralogs are recurrently modified to produce an array of distinct phenotypes. More generally, our results show that even ancient whole-genome duplications can allow for recent phenotypic diversification. Long-read sequencing has allowed us to uncover complex structural variants in regulatory regions, which may play an underappreciated role in regulatory evolution.

LEVERAGING POLYGENIC ENRICHMENTS FOR RISK GENE PRIORITISATION FROM GWAS SUMMARY STATISTICS

Theo Dupuis^{1,2}, Will Macnair¹, Andrew A Brown², Martin Ebeling³, Julien Bryois¹

¹Roche Pharma Research and Early Development, Neuroscience and Rare Diseases Research, Basel, Switzerland, ²University of Dundee, School of Medicine, Population Health and Genomics, Dundee, United Kingdom,

³Roche Pharma Research and Early Development, Pharmaceutical Sciences, Basel, Switzerland

Despite the success of genome-wide association studies (GWAS) in linking genetic regions with disease risk, finding the genes that mediate that risk remains challenging. Other properties of genes can be used to prioritise causal candidates, such as expression in relevant cells or tissues, presence in pathways and protein-protein interaction with other disease-related genes. We developed PERiGene (Polygenic Enrichments for Risk Gene Prioritization), a method which combines these properties with gene-level genetic association statistics, such as MAGMA z-scores, to calculate gene prioritisation scores. Training PERiGene independently on two GWAS of Alzheimer's disease (AD), we found that its predictions replicated better than scores based only on the underlying genetic signal ($\rho_{\text{PERiGene}} \sim 0.58$, $p=0$; $\rho_{\text{MAGMA}} \sim 0.33$, $p=0$), showing the benefits of exploiting multiple sources of information. The top genes identified by PERiGene were also enriched in known AD pathways, such as the microglia phagocytosis pathway (OR=45; adjusted $p=5.8e-21$). Applied to AD, Parkinson's disease, schizophrenia and height, we compared PERiGene's predictions against genes identified in familial forms of the diseases, in OMIM, or with independent methods, such as whole exome sequencing. We observed significant GSEA enrichments ranging from 1.7 to 4.7 depending on the disease (10k permutations); these values are consistently as large as or greater than the enrichments obtained with MAGMA on the same validation sets (GSEA NES: 1.4-2.7). Where the causal gene is known and near a GWAS locus, PERiGene identified it in 39% of cases with the highest score, compared to 27% when using only genetic information. Our results demonstrate the value of PERiGene for post-GWAS analyses, providing a more comprehensive picture of disease risk and leading to more accurate identification of causal genes in complex traits.

STRATOMOD: PREDICTING SEQUENCING AND VARIANT CALLING ERRORS WITH INTERPRETABLE MACHINE LEARNING

Nathan Dwarshuis¹, Peter Tonner², Nathan Olson¹, Fritz Sedlazeck³, Justin Wagner¹, Justin Zook¹

¹National Institute of Standards and Technology, Materials Measurement Laboratory, Gaithersburg, MD, ²National Institute of Standards and

Technology, Information Technology Laboratory, Gaithersburg, MD,

³Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

The Genome in a Bottle (GIAB) consortium generates variant benchmarks for a set of human genomes to enable evaluation and comparison of sequencing technologies and variant detection methods. While these technologies have advanced significantly, correctly calling variants in complex or repetitive regions remains a challenge. We generally understand that sequencing biases and short read lengths can lead to incorrectly-called variants; however, we lack a data-driven model that uses GIAB benchmarking metrics to link variant caller performance to specific, quantifiable features of the context surrounding a given variant.

We aim to make such a model using explainable boosting machines (EBMs). EBMs create models that are a linear combination of arbitrary univariate and bivariate functions (a generalized additive model with interaction terms). Despite being flexible, the relative simplicity of EBMs allows a human to easily understand the functional relationship and relative contribution of each feature. We exploit this transparency here by fitting EBMs to variant call errors as a function of genomic context, which enables understanding when and to what extent a given feature will impact performance.

We demonstrated this strategy with two main use cases. First, we compared false positive rates for Illumina PCR-free and PCR-plus sequencing technologies. Within homopolymers, INDEL false positives were much more prevalent in PCR-plus, and that the error rate for both methods began to increase after 8 bp for both A/T and G/C homopolymers. SNP errors in contrast were not different between the two technologies; however, the error rate increased beyond 10 bp for A/T homopolymers and any length in the case of G/C homopolymers. For our second use case, we compared the likelihood of callsets from Illumina PCR-free or PacBio HiFi data to miss clinically relevant variants. Most false negatives occurred in hard-to-map regions, but some errors occurred in homopolymer regions.

Ultimately, this will provide a data-driven mechanism for understanding sources of error within different variant caller methods and sequencing technologies as they relate to calling variants in difficult regions of the genome. We also plan to use this model to improve genome stratifications for creating and using GIAB benchmarks. The code for this model is at <https://github.com/ndwarshuis/stratomod>.

PROMOTER SEQUENCE AND ARCHITECTURE DETERMINE EXPRESSION VARIABILITY AND CONFER ROBUSTNESS TO GENETIC VARIANTS

Hjörleifur Einarsson, Marco Salvatore, Christian Vaagenso, Nicolas Alcaraz, Sarah Rennie, Jette Bornholdt, Robin Andersson

University of Copenhagen, Department of Biology, Copenhagen, Denmark

Genetic and environmental exposures cause variability in gene expression. Although most genes are affected in a population, their effect sizes vary greatly, indicating the existence of regulatory mechanisms that could amplify or attenuate expression variability. Here, we investigate the relationship between the sequence and transcription start site architectures of promoters and their expression variability across human individuals. We find that expression variability can be largely explained by a promoter's DNA sequence and its binding sites for specific transcription factors. We show that promoter expression variability reflects the biological process of a gene, demonstrating a selective trade-off between stability for metabolic genes and plasticity for responsive genes and those involved in signaling. Promoters with a rigid transcription start site architecture are more prone to have variable expression and to be associated with genetic variants with large effect sizes, while a flexible usage of transcription start sites within a promoter attenuates expression variability and limits genotypic effects. Our work provides insights into the variable nature of responsive genes and reveals a novel mechanism for supplying transcriptional and mutational robustness to essential genes through multiple transcription start site regions within a promoter.

DISCOVERING MACROEVOLUTIONARY TRENDS FOR HUMAN CELL TYPES

Christiana Fauci^{1,2}, Craig B Lowe^{1,2}

¹Duke University Medical Center, Molecular Genetics and Microbiology, Durham, NC, ²Duke University Medical Center, University Program in Genetics and Genomics, Durham, NC

As we sequence additional species and bioinformatic tools advance, we are able to answer increasingly complex questions about macroevolutionary trends throughout vertebrate evolution. The vertebrate lineage includes animals displaying a multitude of phenotypes despite the similarities of their gene sets, which is consistent with non-coding evolution playing a major role in phenotypic diversity. To study the macroevolutionary trends of gene regulation in vertebrates, we have inferred the branch of origin for every region of open chromatin across hundreds of cell types in the human body, which we use as an estimate of the time and magnitude of regulatory changes related to that cell type. The cell types with the most distantly conserved regulatory regions over vertebrate evolution include neurons and skeletal muscle. Regions of open chromatin in several neuronal and skeletal muscle cell types show enrichment for being ancient, over 400 million years old. This suggests that the regulatory landscapes of these cell types have remained relatively unchanged throughout the vertebrate lineage, unlike many other regulatory landscapes. The cell types with the youngest gene regulatory regions in the human genome include placental-, gastrointestinal-, and immune-related regions, with few regulatory regions having orthologs older than placental mammals. These cell types represent what we think are relatively new cell types (e.g. placenta) and cell types that are rapidly changing the regulation of their genes (e.g. immune-related). Going forward we are using this method to better understand the life history of our genomes and the trends of macroevolution that shaped them.

CHROMOSOME-SCALE AND HAPLOTYPE-RESOLVED SEQUENCE ASSEMBLY OF *HORDEUM BULBOSUM* GENOMES

Jia-Wu Feng¹, Maria Cuacos¹, Hélène Pidon¹, Thomas Lux², Heidrun Gundlach², Yi-Tzu Kuo¹, Jörg Fuchs¹, Axel Himmelbach¹, Manuel Spannagl², Jochen Kumlehn¹, Stefan Heckmann¹, Andreas Houben¹, Frank Blattner¹, Nils Stein^{1,3}, Martin Mascher^{1,4}

¹Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, IPK, Seeland, Germany, ²Helmholtz Center Munich, Plant Genome and Systems Biology (PGSB), Neuherberg, Germany, ³Georg-August-University, Center for Integrated Breeding Research (CiBreed), Göttingen, Germany, ⁴German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, iDiv, Leipzig, Germany

Hordeum bulbosum is the closest wild relative of barley (*Hordeum vulgare*). Wide crosses between both species have been used to obtain doubled haploid progeny and develop disease-resistant barley introgression lines. The *H. bulbosum* genome is highly heterozygous and, in some genotypes, autotetraploid. Advances in genome assembly methodology, notably PacBio accurate-long-read (HiFi) and chromosome conformation capture sequencing (Hi-C), have greatly expanded our ability to assemble and phase heterozygous and autopolyploid genomes. Here, we report on developing diploid and tetraploid *H. bulbosum* genome sequence assemblies using the TRITEX computational pipeline. We validated the assembly of one diploid clone by genetic mapping and fluorescence in situ hybridization of haplotype-specific chromosome painting probes. The diversity between *H. bulbosum* haplotypes was higher than those of *H. vulgare*. *H. bulbosum* has more resistance genes of the NLR class than *H. vulgare*. Introgression lines carrying *H. bulbosum* chromatin in a *H. vulgare* background frequently harbored genomic segments rich in NLR genes of the alien donor. These genomic resources will further our understanding of the variation between diploid and tetraploid cytotypes of *H. bulbosum* and the characterization of barley crop-wild introgression lines.

WIDESPREAD TRANSPOSABLE ELEMENT DYSREGULATION IN HUMAN AGING BRAINS

Yayan Feng¹, Feixiong Cheng^{1,2,3}

¹Cleveland Clinic, Genomic Medicine Institute, Lerner Research Institute, Cleveland, OH, ²Case Western Reserve University, Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Cleveland, OH, ³Case Western Reserve University School of Medicine, Case Comprehensive Cancer Center, Cleveland, OH

Background: Transposable elements (TEs), known as “viral elements”, constitute approximately 45% of the human genome. Recent evidences have shown that TEs expression are upregulated with aging, and contribute to aging-associated disorders, including Alzheimer’s disease (AD); yet, the mechanistic basis of AD-risk TE expression upregulation has been relatively understudied.

Method: We investigated TE expression upregulation across diverse pathobiology of Alzheimer’s disease (AD), including Tau and amyloid beta, APOE genotypes, and sex differences using bulk RNA-seq data among three human brain biobanks, including 1) 284 AD and 150 cognitive healthy controls from Mount Sinai Brain Bank (MSBB) brain biobank, 2) 163 AD and 196 cognitive healthy subjects from Mayo clinic brain biobank (Mayo), and 3) 109 AD and 44 cognitive healthy controls from the Rush Memory and Aging Project (ROS/MAP). We next leveraged whole-genome sequencing (WGS) and matched RNA-seq data to identify genome-wide significant TE expression QTLs (teQTLs) in human AD brains. We then used colocalization analysis to integrate six AD GWAS summary statistical datasets with xQTLs, including teQTLs, gene expression QTLs (eQTLs), DNA methylation QTLs (meQTLs), and H3K27 histone acetylation QTLs (haQTLs), for identifying likely causal TEs involved in AD. Finally, we examined the regulatory roles of AD-related TEs using brain cell-type specific chromatin looping data.

Results: We identified 110, 5614, and 181 locus-based TEs showed elevated expression level across MSBB, Mayo, and ROS/MAP brain biobanks, respectively. We showed that TE dysregulation in AD were associated with tau pathology and amyloid neuropathology, and APOE4 genotypes, and acted in sex-specific and cell type-specific manners. Joint-analysis of WGS and matched RNA-seq data identified 38,398 genome-wide significant TE expression QTLs (teQTLs) in human AD brains. Colocalization analysis of teQTLs with AD GWAS loci identified key AD likely causal genes regulated by brain teQTLs. This identified new AD risk genes, including complement C1q tumor necrosis factor-related protein 4 (*C1QTNF4*) and farnesyl-diphosphate farnesyltransferase 1 (*FDFT1*). Regulatory roles of TEs were further confirmed using brain cell-type specific interactomes.

Conclusion: These findings show that teQTLs offer a powerful QTL analytic approach to identify TE-specific risk genes in AD and other neurodegenerative disease if broadly applied.

SEGMENTAL DUPLICATION-MEDIATED VARIATION ACROSS DIVERSE MOUSE GENOMES

Eden Francoeur^{1,2}, Ardian Ferraj^{1,2}, Peter A Audano¹, Parithi Balachandran¹, Christine R Beck^{1,2}

¹The Jackson Laboratory for Genomic Medicine, Farmington, CT,

²University of Connecticut Health Center, Department of Genetics and Genome Sciences, Farmington, CT

Segmental duplications (SDs) are amongst the most rapidly evolving regions of mammalian genomes and are a major source of variation within a species. SDs are large (>1kb), highly homologous DNA sequences (>90%) that constitute over 5% of the mouse genome. The homology between SDs and the orientation of the repeats—direct or indirect—can lead to ectopic rearrangements resulting in large structural variants (SVs), specifically deletions, duplications, and inversions. SDs and SD-rich regions of the genome often contain genes including KRAB zinc-finger proteins, which are known to bind and repress transposable elements. Previous studies have identified SD-mediated rearrangements responsible for reproductive, immune, and behavior related differences between mouse strains. While these SVs have been identified at targeted loci, SD-mediated rearrangements have not been studied as extensively in mice as they have been in human genomes. Consequently, the extent of SD recombination resulting in SVs between diverse subspecies of mice is unknown. Using parameters from the literature on human rearrangements, we have identified SD paralogues on the same mouse chromosome that have >95% sequence identity, are longer than 1kb, and have <10Mb of intervening sequence; with this work, we have determined that ~22% of direct and ~24% of indirect SDs present in the mouse genome are potential candidates for recombination. To test our hypothesized regions prone to SD-mediated variation in mice, we used our orthogonal datasets of short-read sequencing, long-read sequencing, and optical mapping from 10 diverse strains to identify 9 inversions, 155 duplications, and 321 deletions with support from at least two of the three orthogonal methods. These 485 rearrangements contribute to over 650kbp of differences per strain when compared to the mouse reference genome. Many of these SVs lead to copy number variation of genes, including olfactory receptor genes. We further examined the breakpoints of SDs, and found a significant enrichment for LINE elements flanking SDs in mice. This is an important finding that gives insights into SD origin in the mouse genome and is analogous to the enrichment of Alu transposable elements at the flanks of human SDs. With our SD-mediated variation predictions, callset, and breakpoint enrichment analysis, we are investigating how transposons and SDs act as drivers of genomic and transcriptomic evolution within a species, and how the requirements for SD-mediated recombination vary within and between species.

PROTEOME-SCALE PROBABILISTIC MODELING OF HUMAN GENETIC VARIATION

Jonathan Frazer

The Centre for Genomic Regulation, Computational Biology and Health Genomics, Barcelona, Spain

Whole exome sequencing is having a transformative impact on diagnosis of rare disease patients and yet our ability to interpret this data is still in its infancy. Currently, less than 35% of rare disease patients thought to have a monogenic disease receive a genetic diagnosis following sequencing. A major limitation is our inability to rank the pathogenicity of missense variants, especially in genes which have so far not been associated with disease. To tackle this problem, we develop the first generative model of genetic variation across the full human proteome, by leveraging both variation across diverse organisms and variation within the human population. We find this model performs well at identifying pathogenic variants in patient data, distinguishing patients with severe disorders from unaffected individuals with high precision. We then use this model to reanalyze 31k trios with developmental disorders. By incorporating our model in burden testing, we find evidence for 13 novel genetic disorders. Finally we argue that our model can play a valuable role in identifying the many genetic disorders that are so rare they may never be discoverable with burden tests, finding evidence for a further 81 novel genetic disorders in this cohort.

HIGH-THROUGHPUT PROFILING OF ONCOFUSION-INDUCED GENOMIC DYSREGULATION WITH SINGLE-CELL SEQUENCING

Max Frenkel^{1,2}, Zachary Morris¹, Vatsan Raman¹

¹Department of Biochemistry, University of Wisconsin, Madison, WI,

²MSTP, University of Wisconsin School of Medicine and Public Health, Madison, WI, ³Department of Human Oncology, University of Wisconsin School of Medicine and Public Health, Madison, WI

Advances in DNA sequencing have allowed cancer biologists to catalog genetic variants, identify cancer drivers, and define common themes of oncogenesis. One theme is the recurrence of chromosomal translocations that often fuse DNA binding domains of one protein to the effector domain of another. As DNA sequencing has progressed, the list of fusions has grown dramatically. While a handful of well-characterized oncofusions exist, there are thousands of cancer-associated fusions for which we have no mechanistic insight or targeted therapies. Within the growing list of cancer-causing oncofusions are dozens of DNA binding, activating, repressing, and chromatin remodeling domains that can cause complex rewiring of genomic architecture adding to the difficulty of understanding these heterogeneous cancers.

To address both the scale and complexity of this problem, we created a generalizable pipeline for mapping hundreds of protein perturbations to their respective genome-wide effects. Pooled protein libraries are expressed *in vitro* and single-cell sequencing is used to resolve their genome-wide effects. As a proof of concept, we expressed a pilot library of 8 diverse oncofusions in HEK293T cells. Using modified 10X Genomics' single-cell protocols, we link oncofusion identity in individual cells to the resulting chromatin state (scATAC-seq) and gene-expression profile (scRNA-seq). In just one experiment we show that cells expressing EWSR1-FLI1, PAX3-FOXO1, and EWSR1-ATF1 all have increased chromatin accessibility at loci enriched for their known DNA binding motifs and at loci bound in cancer-specific models spanning rhabdomyosarcoma, Ewing sarcoma, and clear cell sarcoma. As well, ETV6-NTRK3 increased accessibility at FOS/JUN motifs likely representing the fusion's known role in activating the AP-1 complex. Despite our assay's non-native context, it recapitulates diverse genomic mechanisms of fusion proteins representing equally diverse cancer types. Unlike existing techniques, this method did not require bespoke customization for each oncofusion and is both high-throughput and high-resolution: a critical advantage allowing us to now rapidly profile hundreds of fusions representing dozens of cancer types. Among seemingly disparate cancer drivers, we hope that the resulting genotype-phenotype maps will reveal common mechanisms of cell-state disruption, shared target genes and networks, and fusion-specific mechanisms that can be exploited for molecular cancer classification and precision therapies.

THE SEX-SPECIFIC GENETIC ARCHITECTURE OF CHILDHOOD ASTHMA

Amelie Fritz^{1,2}, Anders Ulrik Eliasen¹, Kasper Rasmussen¹, Casper Emil Tingskov Pedersen¹, Klaus Bønnelykke¹, Anders Gorm Pedersen²

¹Technical University of Denmark, Health Tech Department, Kgs. Lyngby, Denmark, ²Copenhagen University Hospital, Herlev-Gentofte, Copenhagen Prospective Studies on Asthma in Childhood, Gentofte, Denmark

Childhood asthma is the most common reason for hospitalization in early childhood. From epidemiological studies, it is evident that the prevalence is higher in boys than girls. After puberty, it is more prominent in women than men. The heritability of childhood asthma is estimated to be between 60 and 90%. This suggests that the genetic components driving the development of childhood asthma have a sex-specific effect. Yet, most association studies do not consider gender in their analysis.

In this project, a Bayesian logistic regression model with a variant-sex interaction term was developed in RStan to identify SNPs that have a sex-specific effect on childhood asthma. Discovery studies were conducted in a dataset of 1189 children with severe asthma (2-6 hospitalizations) from Copenhagen Prospective Studies on Asthma in Childhood (COPSAC) and 5094 non-asthmatic controls. 77 variants have a posterior probability of interaction higher than 95%.

A subset of individuals with severe asthma (6+ hospitalizations, 372 individuals) suggests 26 variants with a posterior probability higher than 95% of having a sex interaction.

Sex-stratified analysis confirms the sex-specific effect in both data sets. Two interacting variants are found to be part of the genes IL1R1 and CLEC16A, known for being associated with asthma previously, and 4 of the top 9 interacting variants are expressed in lung tissue. A variant nearby CEP68 is additionally expressed in testis and ovary which strongly suggests an interaction with gender.

Three of the suggested variants replicate in UK Biobank (5581 cases, 88094 controls). Further replication is planned in the iPSCYH data set as well as (sex-specific) eQTL analysis.

SPECIES-AWARE DNA LANGUAGE MODEL

Dennis Gankin, Alexander Karollus, Julien Gagneur

Technical University of Munich, Informatics, Munich, Germany

Predicting gene expression from DNA is an open field of research. As in many areas, labeled data is dwarfed by unlabeled data, i.e. species with a sequenced genome but no gene expression assay data. Pretraining on unlabeled data using masked language modeling has proven highly successful in overcoming data constraints in natural language and proteomics. However, in genomics, this approach has so far been applied only to single genomes, neither leveraging conservation of regulatory sequences across species nor the vast amount of available genomes.

Here we train a masked language model on more than 800 species spanning over 500 million years of evolution. We show that explicitly modeling species is instrumental in capturing conserved yet evolving regulatory elements and in controlling for oligomer biases. Notably, the model captures the evolution of regulatory elements across distant species, which is impossible through alignments due to the evolutionary distance.

We extract embeddings for 3' untranslated regions of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and use them to achieve prediction of mRNA half-life that is better or on-par with the state-of-the-art. Further demonstrating the utility of the approach for regulatory genomics, we also verify that the embeddings are predictive of reporter expression. Moreover, we show that the per-base reconstruction probability of our model directly and significantly predicts sites bound by RNA-binding proteins.

Altogether, our work establishes a self-supervised framework to leverage large genome collections of evolutionarily distant species for regulatory genomics and contributes to alignment-free comparative genomics. It highlights the potential of multi-species DNA language models for studying regulatory evolution and DNA variation.

The preprint is available at: <https://doi.org/10.1101/2023.01.26.525670>

INVESTIGATING RNA-BASED GENE DUPLICATION, A MAJOR FORCE GENERATING GENETIC NOVELTIES IN HUMAN AND OTHER GENOMES.

Pedro A Galante, Helena B Conceição, Gabriela D Guardia, Rafael L Mercuri

Hospital Sirio-Libanes, Centro de Oncologia Molecular, São Paulo, Brazil

The emergence of new genes is a critical source of evolutionary innovation, contributing to lineage or species-specific phenotypic novelties. In mammals, RNA-based gene duplication has emerged as the most common but least studied source of new genes. This process is carried out by two LINE1 (long interspersed nuclear element 1) proteins: one protein that exhibits reverse transcriptase and endonuclease activities, and an RNA-binding protein. Together, the two proteins hijack mRNAs (of protein-coding genes), concomitantly synthesize a molecule copy (retrocopies), and integrate it into the nuclear genome, generating a RNA-based gene duplicate (mRNA retrocopy). Since mRNA retrocopies lack introns and the upstream regulatory regions of their parental genes, they have been thought to be "dead on arrival gene copies" or just (processed) pseudogenes. However, these processed pseudogenes (retrocopies) are common in mammal genomes, and there is growing evidence that some of them are functional. In this study, we investigate the RNA-based gene duplication, or retrocopies, and their role as a major force in generating novel genes and gene functionalities in the human and other mammal genomes. To achieve our goals, we developed and used bioinformatics tools to analyze large-scale genomics and transcriptomics data and investigate several characteristics of retrocopies, such as their genomic features, their expression, coding potential, and polymorphism in the human population. First, our investigation of 32 human tissues revealed approximately 4,000 expressed retrocopies, with 30% of them showing restricted expression (mostly in testis) and 35 expressed in more than 20 human tissues. We also found that roughly 40% of expressed retrocopies are located in intronic regions of protein-coding genes, mostly in the first intron. In addition, we found that most intergenic retrocopies that are expressed are located near already transcribed regions (<8kb). Next, investigated retrocopies as a source of exaptation and identified 17 human microRNAs that originated from retrocopies. Finally, we investigated non-fixed retrocopies in the human genomes, and we identified approximately 600 polymorphic retrocopies in 2,500 individuals from 26 human populations, with the oldest human populations (Africans) showing more population-specific polymorphic retrocopies. In summary, our results shed light on retrocopy functionalities and their contribution to shaping and creating variabilities and novelties in human and other genomes. Support: FAPESP

MACHINE LEARNING BASED PATIENT STRATIFICATION TO ENHANCE PHEWAS ANALYSES USING UK BIOBANK DATA

Manik Garg*¹, Marcin Karpinski*¹, Ryan S Dhindsa^{2,3}, Dorota Matelska¹, Amanda O'Neill¹, Quanli Wang⁴, Andrew Harper¹, Slavé Petrovski¹, Dimitrios Vitsios¹

¹Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, United Kingdom, ²Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX,

³Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, ⁴Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, MA

*These authors contributed equally.

Identifying confident genetic signals from genome-wide association studies (PheWAS) requires sufficient sample sizes and refined patient stratification. Inaccuracies in self-reported diseases, missingness in electronic health records or undiagnosed conditions at the time of assessment among other factors pose challenges in defining patient cohorts. Here we used a wealth of phenotypic information collected by UK Biobank (UKB) to find participants that share similar quantitative traits for a given phenotype. The trained machine learning models also considered the time-lag between sample collection for quantitative trait measurement and diagnosis. Those participants who shared similar quantitative trait profile for a given phenotype formed a part of an extended cohort on which PheWAS rare-variant collapsing analysis was performed. These results were compared with baseline PheWAS collapsing analysis performed on 450K WES samples and presented on AstraZeneca PheWAS portal (Wang et al. 2021, <https://azphewas.com/>). Across 2,770 ICD10 codes (number of samples ≥ 122 , MILTON AUC ≥ 0.6), binary PheWAS analysis on MILTON extended cohorts was able to recover 1,268 out of 3,279 gene signals ($p < 5 \times 10^{-8}$) that were previously missed by binary PheWAS analysis but identified by quantitative PheWAS analysis. Furthermore, across 113 ICD10 codes with MILTON AUC > 0.9 , we found 322 potential novel associations ($p < 5 \times 10^{-8}$) not previously identified by either binary or quantitative PheWAS analysis and require further investigation.

RECOMBINATION BETWEEN HETEROLOGOUS HUMAN ACROCENTRIC CHROMOSOMES

Andrea Guerracino^{1,5}, Silvia Buonaiuto², Leonardo Gomes de Lima³, Tamara Potapova³, Arang Rhie⁴, Sergey Koren⁴, Boris Rubinsteiⁿ, Christian Fischer⁵, Jennifer L Gerton³, Adam M Phillip^y, Vincenza Colonna^{2,5}, Erik Garrison⁵

¹Human Technopole, Genomics Research Centre, Milan, Italy, ²National Research Council, Institute of Genetics and Biophysics, Naples, Italy,

³Stowers Institute for Medical Research, St. Louis, MO, ⁴National Human Genome Research Institute, National Institutes of Health, Genome Informatics Section, Computational and Statistical Genomics Branch, Bethesda, MD, ⁵University of Tennessee Health Science Center, Genetics, Genomics and Informatics, Memphis, TN

The short arms of the human acrocentric chromosomes 13, 14, 15, 21, and 22 share large homologous regions, including the ribosomal DNA repeats and extended segmental duplications. While the complete assembly of these regions in the Telomere-to-Telomere consortium's CHM13 provided a model of their homology, it remained unclear if these patterns were ancestral or maintained by ongoing recombination exchange. Here, we show that acrocentric chromosomes contain pseudo-homologous regions (PHRs) indicative of recombination between non-homologs. Considering an all-to-all comparison of the high-quality human pangenome from the Human Pangenome Reference Consortium (HPRC), we find that contigs from all of the acrocentric short arms form a community similar to those formed by single chromosomes or the sex chromosome pair. A variation graph constructed from centromere-spanning acrocentric contigs indicates the presence of regions where most contigs appear nearly identical between heterologous CHM13 acrocentrics. Except on chromosome 15, we observe faster decay of linkage disequilibrium in the PHRs than in the corresponding short and long arms, indicating higher rates of recombination. The PHRs include sequences previously shown to lie at the breakpoint of Robertsonian translocations, and we show that their arrangement is compatible with crossover in inverted duplications on chromosomes 13, 14, and 21. The ubiquity of signals of recombination between heterologous chromosomes seen in the HPRC draft pangenome's acrocentric assemblies suggests that these shared sequences form the basis for recurrent Robertsonian translocations, providing sequence and population-based confirmation of hypotheses first developed cytogenetically fifty years ago.

DETECTING DECOHERENT GENE CO-EXPRESSION PATTERNS ASSOCIATED WITH A RURAL-TO-URBAN LIFESTYLE TRANSITION IN TURKANA

Kristina M Garske^{1,2,3}, Diogo Melo², Marina M Watowich⁴, Varada Abhyankar¹, Echwa John³, Michael Gurven⁵, John Kahumbu³, Joseph Kamau^{6,7}, Patriciah Kinyua³, Dino J Martins^{2,3,8}, Charles Miano³, Benjamin Muhoya^{2,3}, Julie Peng^{1,2}, Amanda J Lea^{4,9}, Julien F Ayroles^{1,2}

¹Princeton University, Lewis Sigler Institute for Integrative Genomics, Princeton, NJ, ²Princeton University, Department of Ecology and Evolutionary Biology, Princeton, NJ, ³Mpala Research Centre, Turkana Health and Genomics Project, Nanyuki, Kenya, ⁴Vanderbilt University, Department of Biological Sciences, Nashville, TN, ⁵University of California Santa Barbara, Department of Anthropology, Santa Barbara, CA, ⁶National Museums of Kenya, Institute of Primate Research, Nairobi, Kenya, ⁷University of Nairobi, Department of Biochemistry, School of Medicine, Nairobi, Kenya, ⁸Stony Brook University, Turkana Basin Institute, Stony Brook, NY, ⁹Canadian Institute for Advanced Research, Child and Brain Development ProgramToronto, Canada

Transitions from subsistence-level lifestyles toward urban centers are routinely accompanied by an increased risk for non-communicable diseases (NCDs). Identifying the molecular mechanisms underlying drastic lifestyle and health changes can inform policy and interventions for uniquely at-risk populations that are currently undergoing these urban transitions. The Turkana of northern Kenya represent one such group, and we have partnered with them to understand how their evolutionary past and current lifestyles shape traits and NCDs across environments. We have previously shown that genes can exhibit a loss of their pairwise correlations in conditions of stress or disease. This ‘decoherence’ can be used as a biomarker for the development of NCDs, as well as shed light on relevant pathways in chronic gene regulatory dysregulation. To identify genes that are likely to exhibit decoherence in response to their lifestyle transition, we performed RNA-sequencing of peripheral blood mononuclear cells (PBMCs) collected from individuals from the Turkana community, stratified over rural (n=110) and urban (n=229) locations. Differentially expressed (DE) genes are enriched for numerous GO terms and KEGG pathways involving inflammatory mechanisms and response to infection (e.g., response to virus and regulation of inflammatory response), in line with the immune function of these cells. Notably, we also observe an enrichment for pathways related to neurological and metabolic NCDs (e.g., Alzheimer’s disease and non-alcoholic fatty liver disease). Decoherence analysis of the DE gene -enriched pathways highlights *NFE2L2*, the gene encoding the NRF2 transcription factor. This gene exhibits reduced expression correlation with many pathway genes in the urban when compared to rural environments in the Turkana. These pathways include atherosclerosis and response to tumor necrosis factor (TNF). We incorporate genotyping data to map expression quantitative trait loci in the PBMCs and identify genetic polymorphisms regulating the genes that exhibit expression decoherence. In summary, we explore the molecular and genetic basis of physiological responses to lifestyle transitions in the Turkana population, which will ultimately discover environment-driven mechanisms underlying NCD etiology and pathology in a population-aware manner.

USING LONG-READ SEQUENCING TO IDENTIFY METHYLATION DIFFERENCES RELATED TO ALZHEIMER'S DISEASE

Rylee M Genner^{1,2}, Melissa M Meredith³, Kimberley J Billingsley¹, Pilar Alvarez Jerez¹, Laksh Malik¹, Winston Timp², Miten Jain⁴, Cornelis Blauwendraat¹

¹Center for Alzheimer's and Related Dementias, National Institutes of Health, Bethesda, MD, ²Johns Hopkins University, Biomedical Engineering, Baltimore, MD, ³UC Santa Cruz University, Genomics Institute, Santa Cruz, CA, ⁴Northeastern University, Bioengineering, Boston, MA

DNA methylation is an epigenetic mechanism that involves adding a methyl group to cytosine residues in DNA and is generally associated with transcriptional repression. Methylation differences affecting gene transcription levels have been shown to play a role in the development and progression of neurodegenerative diseases including Alzheimer's disease (AD). Methylation levels have traditionally been measured using bisulfite conversion followed by methylation arrays and short read sequencing techniques, however these methods detect only short-range methylation patterns, are not optimized for identifying methylation in traditionally challenging genomic regions, and do not allow for phased haplotype specific methylation detection.

New “long-read” sequencing approaches have been developed by PacBio and Nanopore that can directly sequence reads ranging from 10kb to 1Mb+ in length. This results in increased sequencing accuracy and resolution, which is particularly helpful for analyzing complex regions of the genome. Long-read sequencing can also directly detect base modifications such as 5MC methylation, allowing for additional genetic information to be extracted such as haplotype phasing for the detection of allele-specific methylation differences.

NIH's Center for Alzheimer's Disease and Related Dementias (CARD) has developed protocols designed to streamline and automate the tissue processing and long-read sequencing of thousands brain samples from individuals with and without AD. Generation of this sequencing data provides a unique opportunity to analyze genome-wide, population scale methylation patterns and assess the methylation levels of poorly resolved genomic regions in human brain tissue.

We are currently using computational methods and algorithms to identify allele-specific, cell type-specific and genome-wide methylation differences in long read sequencing data from the first ~250 brain samples. This information will provide new insights into the epigenetic mechanisms of the brain and the underlying etiology of AD that could eventually lead to the development of novel and improved treatment strategies.

THE TOPOGRAPHY OF NULLOMER-RESURFACING MUTATIONS AND THEIR RELEVANCE TO HUMAN DISEASE

Candace Chan^{1,2}, Ioannis Mouratidis³, Georgios G Tsatsianis³, Sarah Fong^{1,2}, Martin Hemberg⁴, Nadav Ahituv^{1,2}, Ilias Georgakopoulos-Soares³

¹Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA., Department of Bioengineering and Therapeutic Sciences, San Francisco, CA, ²Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA, Institute for Human Genetics, San Francisco, CA,

³Department of Biochemistry and Molecular Biology, The Pennsylvania State University, Department of Biochemistry and Molecular Biology,

Hershey, PA, ⁴Evergrande Center for Immunologic Diseases, Harvard

Medical School and Brigham and Women's Hospital, Boston, MA, USA, Evergrande Center for Immunologic Diseases, Boston, CA

Nullomers are DNA sequences that are absent from a genome; however they can resurface due to mutations. Here we characterize all possible nullomer-resurfacing single base pair mutations, population variants and disease-causing mutations. We report that the primary factor of nullomer formation in the human genome is CpG dinucleotides and methylated cytosines. Among repeat elements, Alu repeats exhibit pronounced enrichment for nullomer-resurfacing mutations at specific positions. We also find that nullomer-resurfacing mutations are enriched at genomic elements, including transcription start and end sites, splice sites and transcription factor binding sites and observe that regions with nullomer-resurfacing mutations are show preferential location relative to nucleosomes. We also report that pathogenic mutations are significantly more likely to cause resurfacing of nullomers than their benign counterparts. Therefore, the increased mutation rate at these sequences reflects the vast majority of nullomers observed while we also provide evidence for selection pressures at specific loci. We conclude that methylated cytosines and CpG dinucleotides are the primary determinant of nullomer-resurfacing in the human genome.

EFFECTS OF PARENTAL AGE AND POLYMER COMPOSITION ON SHORT TANDEM REPEAT *DE NOVO* MUTATION RATES

Michael E Goldberg^{1,2}, Michelle D Noyes², Evan E Eichler^{2,3}, Aaron R Quinlan¹, Kelley Harris^{2,4}

¹University of Utah, Departments of Human Genetics and Biomedical Informatics, Salt Lake City, UT, ²University of Washington, Department of Genome Sciences, Seattle, WA, ³University of Washington, Howard Hughes Medical Institute, Seattle, WA, ⁴Fred Hutchinson Cancer Research Center, Computational Biology Division, Seattle, WA

Short tandem repeats (STRs) are hotspots of genomic variability in the human germline because of their high mutation rates, which have long been attributed largely to polymerase slippage during DNA replication. This model suggests that STR mutation rates should scale linearly with the number of cell divisions in the male germline, where progenitor cells continually divide after puberty. In contrast, STR mutation rates should not scale with the age of the mother at her child's conception, since oocytes spend a mother's reproductive years arrested in meiosis II and undergo a fixed number of cell divisions prior that are independent of the age at ovulation.

We tested this prediction using *de novo* mutation calls from the Simons Simplex Collection, a cohort of nearly 1600 human quad families, each consisting of two children plus parents whose ages at the birth of the children are known. Contrary to expectation, STR mutation rates covary with maternal age as well as paternal age, implying that some STR mutations are caused by DNA damage in quiescent cells rather than the classical mechanism of polymerase slippage. Our results echo both the recent finding that DNA damage in quiescent oocytes is a significant source of *de novo* SNVs in addition to evidence of STR expansion in postmitotic cells. However, we find that the maternal age effect is not confined to previously discovered hotspots of oocyte mutagenesis, nor are post-zygotic mutations likely to contribute significantly. STR nucleotide composition demonstrates divergent effects on DNM rates between sexes. The mutation rate at GC-containing STRs is significantly more associated with paternal age than the rate at STRs composed solely of AT nucleotides, whereas the opposite association exists with maternal age. These observations suggest both the mechanism and developmental timeline of certain STR mutations, and are especially surprising in light of the prior belief in replication slippage as the dominant mechanism of STR mutagenesis.

THE ROLE OF RECOMBINATION IN THE ORIGIN AND EVOLUTION OF HUMAN INVERSIONS.

Ruth Gómez-Graciani¹, Antonio Barbadilla^{1,2}, Mario Cáceres^{1,3}

¹Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain, ²Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain, ³ICREA, -, Barcelona, Spain

Chromosomal inversions are structural variants that alter recombination in heterozygosity and are involved in multiple evolutionary processes. This recombination inhibition could be caused by a physical limitation to synapse during meiosis, or by purifying selection against early gametes carrying aberrant chromosomes resulting from a crossover between opposite orientations. The relevance of each of both mechanisms driving recombination reduction in humans is unknown. To counter purifying selection, new inversions either should arise in regions of low recombination and/or confer positive evolutionary effects that overcome the potential fitness loss caused by their effect on fertility. Here, we investigate the relationship between inversions and recombination in humans by combining different types of recombination maps and a highly accurate set of 133 polymorphic inversions ranging from 0.2 kb to 5.7 Mb. Using pedigree and likelihood-based population recombination maps, we found that inversion location patterns are mainly related to the generation mechanism of inversions: NAHR-mediated inversions are more frequent in repetitive regions, while NH inversions are generated at random in the genome. Large inversions at frequencies above 0.2 overlap less than expected with the highest 10% recombination rates in the genome, and are enriched within the 0.01% lowest recombination rate bins. Using a published set of 813,122 crossovers and 787 aneuploid chromosomes from 20 sperm donors, we observed a decrease in crossover rates within the inverted region in heterozygotes compared to homozygotes, which is especially noticeable in inversions >25 kb. Moreover, we found a positive correlation between the genetic length affected by inversions in heterozygosity and the number of aberrations detected on each chromosome. These results suggest that the reduction of recombination between orientations is due to the unviability of recombinant chromosomes rather to asynapsis. Therefore, inversions are more likely to survive when the crossover probability within the inversion region is low.

IDENTIFICATION OF HAPLOTYPE EPISTASIS IN GENE EXPRESSION REGULATION USING DEEP LEARNING MODELS

Saideep Gona¹, Temidayo Adeluwa¹, Andy Dahl¹, Boxiang Liu², Ravi Madduri³, Hae Kyung Im¹

¹University of Chicago, Genetic Medicine, Chicago, IL, ²National University of Singapore, Biomedical Informatics, Singapore, Singapore,
³Argonne National Lab, Mathematics and Computer Science, Chicago, IL

Gene regulation underlies how our single genome gives rise to higher level phenotypes. Understanding it provides insight into how homeostasis is maintained, and how misregulation can lead to disease. A form of gene regulation is epistatic effects, where the potency of gene regulatory regions is conditional on other genomic regions. We are interested more specifically in whether total gene expression in diploid organisms is just the sum of haploid expression or if epistatic interactions between the haplotypes help explain the total expression. An extreme example of this is imprinting, where one parental haplotype is silenced conditional on the simple presence of another. In less extreme cases it can be hard to disentangle whether differences in each parental haplotype's contribution to total expression are due to their own cis regulation or, like with imprinting, are somehow conditional on characteristics of the other parental haplotype. If we knew the expression level of each haplotype in isolation from the other (haploid expression), we could test whether their sum is indeed equivalent to the observed total diploid expression, with deviations from this likely due to epistasis. Interestingly, deep learning (DL) models trained to predict gene expression from DNA sequence are usually designed to only predict using a single haplotype input. As a result, their predictions are analogous to isolated haploid expression, thus providing an avenue to test the epistasis hypotheses. In this work, we use Enformer, the latest in the line of convolutional sequence-to-DL models to make haploid expression predictions. Due to its use of transformers, Enformer can accept an input sequence of 400kb, providing extensive context for cis regulatory variation affecting the isolated haploid predictions. Enformer expression predictions were made for 463 individuals with GEUVADIS lymphoblastoid (LCL) expression data using genotype data from the 1000 Genomes Project. Predictions were restricted to the ten thousand most variable expression GEUVADIS genes. Of these, 635 showed significant additive association between the Enformer predictions and GEUVADIS ground truth (BH adj.p < 0.05), suggesting their expression to be genetically controlled. A further 13 genes showed significant interaction, suggesting probable epistasis (BH adj.p < 0.05). Among this enriched set, we observed a significant enrichment of ribosomal genes (Fisher's test p=0.0073) against the test set background, a finding which may indicate increased utilization of epistasis in the regulation of ribosomal activity. Our study highlights the potential of using deep learning models to improve our understanding of gene regulation and its impact on higher level phenotypes.

TERMINATOR: RISE OF THE PLANT MACHINE. HOW LARGE SCALE CHARACTERIZATION OF PLANT TERMINATORS REVEALS SPECIES-SPECIFIC EXPRESSION LEVELS

Sayeh Gorjifard¹, Tobias Jores¹, Jackson Tonnies¹, Nicholas A Meuth¹,
Kerry Bubb¹, Travis Wrightsman², Joshua T Cuperus¹, Edward S Buckler²,
Stanley Fields¹, Christine Queitsch¹

¹University of Washington, Genome Sciences, Seattle, WA, ²Cornell University, Integrative Plant Science Plant Breeding and Genetics, Ithaca, NY

The 3' end of a gene, often called terminator, plays critical roles in plant gene regulation by modulating translation efficiency, polyadenylation, mRNA localization, and mRNA stability. In plants, terminator activity has not yet been systematically interrogated. Here, we adapted Plant STARR-seq, a massively parallel reporter assay, to determine the effects of sequence composition on terminator activity. To do so, we assayed 54,621 terminators of nearly all annotated genes of the crucifer model *Arabidopsis thaliana* and the crop *Zea mays*, in addition to 4 commonly used non-plant terminators and 5378 control sequences designed to test hypotheses on terminator elements, in tobacco leaves and maize protoplasts. Our results successfully recapitulate known terminator biology and enable us to determine the relative contributions of terminator motifs to overall terminator strength. We find thousands of plant terminators that outperform bacterial terminators commonly used to engineer transgenic crops. The strongest terminators tend to reside in metabolic genes of both *Arabidopsis* and maize. We observe species-specific differences between *Arabidopsis* and maize terminators and determine how these differences affect terminator activity in a monocot and a dicot assay system. Further, we discover alternative polyadenylation sites across tens of thousands of terminators; however, the strongest terminators in our assay tended to have a single dominant cleavage site. Enabled by the large size of our data set, we built a computational model (CNN) to predict terminator strength in both assay systems ($R^2=0.76$ for tobacco; $R^2=.67$ for maize), and conducted *in silico* evolution to optimize terminators. Our results establish principles of terminator function in plants and identify natural and synthetic terminator sequences that will improve performance of transgenes in crop plants.

INTEGRATION OF 170,000 SAMPLES REVEALS GLOBAL PATTERNS OF GUT MICROBIOME DIVERSITY

Samantha P Graham^{*1}, Richard J Abdill^{*2}, Frank W Albert¹, Ran Blekhman²

¹University of Minnesota, Department of Genetics, Cell Biology, and Development, Minneapolis, MN, ²University of Chicago, Section of Genetic Medicine, Department of Medicine, Chicago, IL

* - These authors contributed equally to this work

Important relationships exist between human health and disease and the microbes that comprise our microbiome. Beneficial microbes can provide energy sources like short-chain fatty acids to the host, whereas others can have detrimental effects, and have been linked to diseases such as inflammatory bowel disease and certain cancers. However, microbiome studies are inherently limited due to the sparse, high-dimensional, and compositional nature of the data. While increasing sample size can help mitigate the issues associated with such sparse and high-dimensional data, most human microbiome studies are small, with a median size of only 79 samples. Here, we have leveraged publicly available 16S rRNA microbiome data to create the Human Microbiome Compendium, the largest collection of uniformly processed human gut microbiome data. The compendium currently comprises over 170,000 samples from 109 countries and 6 continents.

Analysis of the compendium reveals Firmicutes to be the most prevalent phylum, consistent with previous reports. Present in 99.8% of samples, the relative abundance of Firmicutes ranges almost uniformly from 0 to nearly 100% across samples. We found strong associations between geographic region and microbiome composition. We note that the alpha diversity, a measure of the number of species per sample, is highest in Latin America and the Caribbean. Prevotella, which has previously been associated with a non-westernized diet and lifestyle, was highest in abundance in Sub-Saharan Africa. Principal coordinates analysis revealed that Bacteroidia and Actinobacteria abundances are important in differentiating samples. We show that the large sample size helps reveal signals in the microbiome that were previously difficult to detect. Unique to Sub-Saharan Africa is a strong, positive correlation between the abundance of Bifidobacteriaceae and Streptococcaceae, as well as a negative correlation between Ruminococcaceae and Staphylococcaceae. We discovered significant correlations between the abundance of Lachnospiraceae and Ruminococcaceae in every geographic region.

In summary, our integrated analysis of large-scale publicly-available microbiome data found previously-unknown broad-scale regional patterns in microbiome variation. We anticipate that the Human Microbiome Compendium will be useful as a resource for others, particularly researchers working with machine learning tasks that require an otherwise impractical amount of training data to assess large-scale patterns.

REGULATORY ENHANCER-GENE INTERACTIONS IN THE HUMAN GENOME

Andreas R Gschwind^{*1}, Kristy Mualim^{*1}, Alireza Karbalayghareh^{*2}, Maya U Sheth^{*1}, Kushal K Dey^{*3}, Ramil N Nurtdinov^{*4}, Evelyn Jagoda^{*5}, Wang Xi^{*6}, Alkes L Price³, Michael Beer⁶, Roderic Guigo⁴, Lars M Steinmetz¹, Christina Leslie², John Stamatoyannopoulos⁷, Erez Aiden⁸, William J Greenleaf¹, Anshul Kundaje¹, Jesse M Engreitz¹

¹Stanford University, Department of Genetics, Stanford, CA, ²Memorial Sloan Kettering Cancer Center, New York, NY, ³Harvard T.H. Chan School of Public Health, Department of Epidemiology, Boston, MA, ⁴Universitat Pompeu Fabra (UPF), Institute for Science and Technology (BIST), Centre for Genomic Regulation (CRG), Barcelona, Spain, ⁵Broad Institute of MIT and Harvard, Cambridge, MA, ⁶Johns Hopkins University School of Medicine, Department of Biomedical Engineering, Baltimore, MD, ⁷University of Washington, Department of Genome Sciences, Seattle, WA, ⁸Baylor College of Medicine, The Center for Genome Architecture, Houston, TX

The ENCODE4 Consortium

Identifying *cis*-regulatory elements, such as enhancers, and their target genes is essential for understanding gene regulation and the impact of human genetic variation. Here we create and evaluate a resource of >20 million regulatory enhancer-gene interactions across 352 cell types and tissues, by integrating predictive models, measurements of chromatin state and 3D contacts, and large-scale genetic perturbation datasets generated by the ENCODE4 Consortium. We first create a systematic benchmarking pipeline to compare predictive models, assembling a dataset of 10,411 element-gene pairs measured in CRISPR interference perturbation experiments, >30,000 fine-mapped eQTL variants, and 750 fine-mapped GWAS variants linked to a likely causal gene. Using this framework, we develop a new predictive model, ENCODE-E2G, that achieves state-of-the-art performance across multiple prediction tasks, including generalizing to make predictions in new cell types based on cell-type specific measurements of chromatin state and 3D physical interactions. By interpreting the model, we find evidence that ubiquitously expressed genes are insensitive to distal regulatory input by enhancers, and that enhancers can act synergistically to affect chromatin state at nearby enhancers and expression of nearby genes. We examine the global properties of the predicted enhancer networks, and identify differences in the functions of genes that have more or less complex regulatory landscapes. Finally, we provide guidelines for using ENCODE-E2G predictions to link noncoding variants to target genes and cell types for common, complex diseases. Altogether, these genome-wide maps of regulatory enhancer-gene interactions, benchmarking software, predictive models, and insights about enhancer function will provide a valuable resource for future studies of gene regulation and human genetics.

THE RAPID EVOLUTION OF TBC1D3-A GENE IMPLICATED IN HUMAN CORTICAL EXPANSION.

Xavi Guitart¹, Philip Dishuck¹, David Porubsky¹, PingHsun Hsieh³, Evan Eichler^{1,2}

¹University of Washington, Genome Sciences, Seattle, WA, ²Howard Hughes Medical Institute, Chevy Chase, MD, ³University of Minnesota, Institute of Health Informatics, Minneapolis, MN

TBC1D3 is a primate-specific gene family implicated in human cortical expansion. Transgenic experiments suggest that *TBC1D3* increases basal radial glial cell divisions and as a result the number of neurons in the human cortex. Unfortunately genetic variation of this locus has been impossible to study using traditional methods because the gene is embedded in high-identity segmental duplications. Using long-read sequencing data (HiFi & ONT), we sequenced, assembled and validated the two 1 Mbp-size *TBC1D3* clusters on human chromosome 17 in 47 human genomes and 9 non-human primate genomes. Only 46% (44/94) of haplotypes were fully resolved in the HPRC consortium, however, reassembly using the hybrid assembler recovered an additional 16 haplotypes, providing 60 fully resolved human loci for comparison to NHP. Using maximum likelihood methods, pan-genome graph as well as structural variation analysis tools, we reconstructed the evolution and population genetic diversity of *TBC1D3*. There were three important conclusions. First, while humans have increased copy number compared to chimpanzee, *TBC1D3* appears to have duplicated independently in human, gorilla, orangutan, and gelada lineages with copy numbers ranging from 9 to 23. In humans we estimate that the *TBC1D3* expansion occurred 700 kya- 3 mya consistent with fossil record evidence for the tripling of ancestral human cranial volume. Second, human haplotypes can vary by more than a Mbp in length due to palindromic expansions of *TBC1D3*, which varies between 3 and 28 copies. We estimate structural heterozygosity of *TBC1D3* exceeds 60% making it among the most structurally variable loci in the human genome. Interestingly, individuals of African descent tend to have higher copy-number suggesting higher *TBC1D3* copy number in the ancestral human lineage that may be reducing in Out-of-Africa populations. Third, our investigation into the gene structure using full-length transcript sequencing from IsoSeq suggest that the *TBC1D3* gene model has been altered between humans and NHP. Remarkably, unique sequence flanking *TBC1D3* genes mapping to the second cluster show a significant dearth of common variants (Tajima's D=-1.79) suggesting ongoing positive selection. We hypothesize that extraordinary turnover of these duplicated genes during primate evolution provided the milieu for neofunctionalization in humans, but the extraordinary genetic diversity is difficult to reconcile with its purported function in the expansion of the human prefrontal cortex.

SINGLE CELL MULTIOME ANALYSIS REVEALS CANCER CELL PLASTICITY

Yasuhiko Haga, Masahide Seki, Ayako Suzuki, Yutaka Suzuki

The University of Tokyo, Graduate School of Frontier Science, Kashiwa, Japan

It is known that after treatment with anticancer drugs, some cancer cells survive by acquiring drug-resistance. While the resistance is acquired by genomic mutations in some cases, the resistance is acquired by changes in the transcriptome or epigenome in other cases. It is important to elucidate the changes in gene expression and epigenomic status for the acquisition of resistance in the so-called “persistent cells”.

In this study, we used single-cell multiome (scGEX-ATAC) technology, which can simultaneously obtain single nuclei RNA-seq and single cell ATAC-seq data derived from the same cell, and NCC oncopanel, which can get captured deep-sequencing genome data, to decipher the persistent state in cancer cells from both the transcriptomic and epigenomic levels until acquisition of genomic mutation.

For this purpose, a lung cancer cell line PC9, have EGFR exon19 deletion as a driver mutation, were cultured with treatment of gefitinib, a first-generation EGFR-TKI, for 3 days. After that, the cells were cultured in the condition without gefitinib to recover for 3 days. These treatments were repeated 20 times. The cells were sampled at multiple timepoints, and scGEX-ATAC and deep-sequencing were performed. As a result, ~8,030 cells on average scGEX-ATAC data from 9 samples, and about 5,200 ~ 7,800 depth deep sequencing data from 16 samples, were obtained. We analyze these data focusing on the difference of changes between transcriptome and epigenome. We found that the cells were divided two populations in early stages after treatment, and then, cells in one population began to occur dedifferentiation, while cells in another population retained epithelial cell status. Then, as the treatment continued, the latter population was vanished. While, famous drug-resistant mutations, such as EGFR T790M mutation, were not detected in deep-sequencing data. But, in some cells, the non-synonymous genomic mutation on JAK3 was occurred. Now, we analyze which this mutation is functional or not.

TRANSCRIPTOME ANALYSIS OF FAMILIAL DYSAUTONOMIA REVEALS TISSUE-SPECIFIC GENE EXPRESSION DISRUPTION IN THE PERIPHERAL NERVOUS SYSTEM

Ricardo S Harripau^{1,2,3}, Elisabetta Morini^{1,2}, Monica Salani¹, Emily Logan¹, Emily G Kirchner¹, Jessica Bolduc¹, Anil Chekuri^{1,2}, Benjamin Currall^{1,2,3}, Rachita Yadav^{1,2,3}, Serkan Erdin^{1,2,3}, Michael E Talkowski^{1,2,3}, Dadi Gao^{1,2}, Susan Slaugenhouette^{1,2}

¹Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, ²Department of Neurology, Massachusetts General Hospital Research Institute and Harvard Medical School, Boston, MA, ³Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA

Familial Dysautonomia (FD) is a rare recessive neurodevelopmental disease caused by a splice site mutation in the Elongator acetyltransferase complex subunit 1 (ELP1) leading to tissue-specific skipping of exon 20 and reduction of the ELP1 protein, distinctly in the central nervous system (CNS) and peripheral nervous system (PNS). Here we performed a transcriptome-wide study to dissect the molecular mechanisms underlying FD in specific neuronal tissues from the FD phenotypic mouse which expresses human ELP1, including the dorsal root ganglion (DRG), trigeminal ganglion (TG), medulla (MED), cortex, and spinal cord (SC). We focused our analyses on differentially expressed genes (DEGs) representing the most dominant transcriptomic alterations; and on genes in co-expression modules that are highly correlated with full-length ELP1 expression (ELP1 dose-responsive genes). We identified a higher number of DEGs (342) in the PNS (DRG, TG) as compared to the CNS (MED, SC, Cortex) (143). ELP1 dose-responsive genes are only found in DRG, TG, and MED, not in Cortex or SC, tissues. Gene Ontology analyses of both DEGs and ELP1-dose-responsive genes highlight the regulation of neurotransmitters. The transcriptome-wide signals were highly convergent between PNS tissues (DRG and TG) but not among CNS tissues. Those convergent genes were enriched for known protein-protein interactions and cell type-specific markers defining myelinated neurons and peptidergic nociceptors. Our findings support the involvement of specific neuronal subtypes underlying the PNS phenotypes in FD. Our study comprehensively investigates transcriptome-wide alterations in FD neuronal tissues and identifies the functional dysregulations in the peripheral nervous system contributing to disease.

CENTURIES OF GENOME INSTABILITY AND EVOLUTION IN SOFT-SHELL CLAM TRANSMISSIBLE CANCER

Samuel F Hart^{1,2}, Marisa A Yonemitsu¹, Rachael M Giersch¹, Brian F Beal³, Gloria Arriagada⁴, Brian W Davis⁵, Elaine A Ostrander⁶, Stephen P Goff⁷, Michael J Metzger^{1,2}

¹Pacific Northwest Research Institute, Seattle, WA, ²University of Washington, Molecular and Cellular Biology Program, Seattle, WA,

³University of Maine at Machias, Environmental and Biological Sciences, Machias, ME, ⁴Universidad Andres Bello, Instituto de Ciencias Biomedicas, Facultad de Medicina y Facultad de Ciencias de la Vida, Santiago, Chile,

⁵Texas A&M University School of Veterinary Medicine, Veterinary Integrative Biosciences, College Station, TX, ⁶National Human Genome Research Institute, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, ⁷Columbia University, Microbiology and Immunology, New York, NY

Transmissible cancers are infectious parasitic clones of malignant cells that metastasize to new hosts, living past the death of the founder animal in which the cancer initiated. We investigated the evolutionary history of a cancer lineage that has spread through the soft-shell clam (*Mya arenaria*) population by assembling a chromosome-scale 1.2 Gb soft-shell clam reference genome and characterizing somatic mutations from cancer sequences. We observe high mutation density, widespread copy number gain, structural rearrangement, loss of heterozygosity, variable telomere lengths, mitochondrial genome expansion, and transposable element activity, all indicative of an unstable cancer genome. We also discover a previously unreported mutational signature that predicts the cancer lineage to be at least two hundred years old. Our study reveals the ability for an invertebrate cancer lineage to survive for centuries while its genome continues to structurally mutate, likely contributing to the ability of this lineage to adapt as a parasitic cancer.

WHOLE GENOME BURDEN TESTING IN 333,100 INDIVIDUALS OF RARE NON-CODING GENETIC VARIATION ON COMPLEX PHENOTYPES

Gareth Hawkes¹, Robin N Beaumont¹, Zilin Li³, Ravi Mandla², Xihao Li³, Alisa K Manning², Xihong Lin³, Caroline F Wright¹, Andrew R Wood¹, Timothy M Frayling¹, Michael N Weedon¹

¹University of Exeter, Clinical and Biomedical Sciences, Exeter, United Kingdom, ²Harvard, Broad Institute, Boston, MA, ³Harvard, T.H. Chan School of Public Health, Boston, MA

Most sequence-based association studies for common human phenotypes have focussed on rare variants that reside in the coding regions of the genome. However, the recent release of whole-genome-sequence (WGS) data in 100,000s of individuals from several studies, such as the UK Biobank, provides an unprecedented opportunity to examine rare, non-coding variants and their contribution towards the genetic architecture of common traits.

Using height as an exemplar trait, we analysed 333,100 individuals from three studies: UK Biobank (N=200,003), TOPMed (N=87,652) and AllOfUs (N=45,445). TOPMed and AllOfUs in particular are diverse cohorts with >50% individuals of non-European ancestry. To facilitate our discovery efforts in the UK Biobank, we developed a novel analytical pipeline with the aim of finding novel rare (<0.1% minor allele frequency) non-coding genetic associations. We tested 75,311,546 variants which had at least 20 carriers in the UK Biobank, and performed 52,749,161 genomic aggregate tests based on a hierarchical classification of genetic variants into either gene-centric (coding, proximal, intronic) and non-gene-centric (regulatory, intergenic) groupings. We additionally grouped variants by measures of conservation, constraint and deleteriousness. Finally, we performed a hypothesis-free 2kb sliding window analysis, with 1kb overlap.

We observed 30 independent novel rare variants associated with height using an empirical significance threshold of $p<6.3\text{E-}10$, after adjusting for 12,111 common variants previously reported by the Genetic Investigation of ANthropometric Traits (GIANT) consortium. We additionally conditioned on 606 variants reported from an exome array-based analysis and 12,796 variants that were genome-wide significant in an exome-wide analysis of height by Regeneron. We observed effect sizes range from -7cm to +2cm, and replicated 3 rare single variant associations. We also observed evidence for 9 non-coding genomic regions associated with adult height, including regions proximal to *HMGA1* and *GHI*, and a novel association downstream of *C17orf49* overlapping pseudo-exons of miRNA, all three of which showed evidence of replication. Finally, we will discuss the challenges of whole-genome association testing and demonstrate our analytical pipeline.

Our approach provides a template for the analysis of non-coding rare variants for common human phenotypes.

HOW DOES SOMATIC MOSAICISM VARY ALONG THE LENGTH OF THE HUMAN COLON?

Laurel Hiatt¹, Jason Kunisaki¹, Suchita Lulla¹, Xichen Nie², James Hotaling³, Aaron Quinlan¹

¹University of Utah, Human Genetics, Salt Lake City, UT, ²University of Utah, Oncological Sciences, Salt Lake City, UT, ³University of Utah, Surgery, Salt Lake City, UT

Somatic mosaicism results from accumulated mutations in non-germline cells throughout an organism's lifetime. Somatic mutations play a central role in pathogenesis, from developmental syndromes to cancer, and there is growing consensus that somatic mosaicism in healthy tissue influences fitness and disease predisposition. Recent experimental advances have made it possible to investigate the somatic mosaicism of healthy organs and somatic evolution in disease pathogenesis which can, in turn, facilitate predicting disease risk and optimizing disease management.

The colon is ideal for studying mosaicism it has significant associations with colorectal disease: somatic mutations are a known causative factor in colorectal cancer and are proposed to contribute to inflammatory bowel disease. These two umbrellas of colorectal diseases cause extensive mortality and morbidity: colorectal cancer is the 2nd leading cause of US cancer death, and inflammatory bowel disease affects one in 200 individuals of European ancestry. These diseases have well-established and intriguing regional presentations, with distinct pathologies along the length of the colon, such as “right” and “left” colorectal cancer and differently clustered subtypes of inflammatory bowel disease. However, the genetic etiologies underlying these region-specific pathologies are unknown, given previous limitations in mosaic research. Evaluating mosaicism across the colon’s length will provide insight into whether region-specific mutageneses, such as developmental events or microbiome exposures, may explain variation in regional pathogenesis.

We are leveraging access to cadaveric tissue and colonoscopy samples acquired from the University of Utah Gastrointestinal Biobank to investigate the mutational landscape of distinct colorectal regions. We have two pilot studies aimed at resolving two clinical states: we are examining healthy regional tissue via detailed cadaveric biopsy (10 standardized sites, five donors) and analyzing colorectal carcinogenesis by comparing colonoscopy biopsies from patients in different stages of disease progression (5 standardized sites, six patients across three clinical cohorts). By extracting mutational signatures indicative of mutational etiology, we will determine whether there are endogenous or exogenous processes that may be enriched in specific parts of the organ. We will also evaluate mutation rate gradients as they may vary across the colon. Investigating these processes will provide insight into regional pathologies and their genetic etiologies.

THE IMPACT OF GENOMIC VARIATION ON FUNCTION CATALOG

Benjamin C Hitz¹, Pedro R de Assis¹, Idan Gabdank¹, Shencheng Dong¹, Meenakshi S Kagda¹, Mingjie Li¹, Otto Jolanki¹, Jennifer Jou¹, Kalina Andreeva¹, Khine Lin¹, Ian Whaling¹, Wenjin Zhang², Xiaowen Ma², Daofeng Li², Heather Lawson², Feng Yue³, Ting Wang², J Michael Cherry¹

¹Stanford School of Medicine, Genetics, Stanford, CA, ²Washington University, Genetics, Saint Louis, MO, ³Northwestern University, Biochemistry and Molecular Genetics, Chicago, IL

The IGVF (Impact of Genomic Variation on Function) Consortium aims to understand how genomic variation affects genome function, which in turn impacts phenotype. The NHGRI has funded this collaborative international program to bring together teams of investigators who will use state-of-the-art experimental and computational approaches to model, predict, characterize and map genome function, learn how genome function shapes phenotype, and how these processes are affected by genomic variation. The ultimate goal of these joint efforts is to generate a resource centered on a Catalog of the experimental results, models, and predictions of the IGVF and to put them in context of public genomic data.

Our approach to this problem is to create a massive genomics knowledge graph, powered by ArangoDB and hosted in the cloud. The graph will have human genetic variants, genes, transcripts, proteins, regulatory elements (putative and experimentally validated), and binding sites represented as nodes in the graph, while their relationships -- defined by experimental and theoretical methods will be represented as edges. For example, QTLs (expression, chromatin accessibility, splicing etc) will be represented as links between genomic variants and genes (or accessible elements, or transcripts). Methods such as ABC model predictions will be used to make links between "enhancers" (regulatory elements) and genes. GWAS studies can be used to link genomic variants to phenotypes, and Linkage Disequilibrium links variants to each other within a particular cohort or ancestry group. In this way we plan to represent the vast variety of products from the IGVF in a graph. ArangoDB is an efficient, hybrid database system that can be used to represent and query graphs along side other types of data storage as well. Specifically it has indices that allow efficient genomic interval queries (i.e "Does this variant's position intersect the coordinates of which set of regulatory intervals or elements; where the number of intervals is in the 100s of millions).

Finally we have wrapped this database in a tRPC base API to allow efficient access to remote queries which will be used for visualizations and analysis of the IGVF data.

The IGVF-catalog frontend UI is proposed to be a user-friendly, responsive web application. It aims at providing intuitive navigation, search functions and visualizations of variation data. We hope the catalog would enable researchers to visualize and interpret variation data from multiple sources with ease.

TRANSCRIPTIONAL PROFILING OF 3D LOWER AIRWAY TISSUE CULTURE MODEL EXPOSED TO PHYLOGENETICALLY DIVERSE MICROBES

Mian Horvath^{1,2}, Elizabeth Fleming¹, Ruoyu Yang^{1,2}, Diana Cadena Castaneda¹, Megan Callendar¹, Jose Fachi³, Marco Colonna³, Karolina Palucka¹, Julia Oh¹

¹The Jackson Laboratory, Genomic Medicine, Farmington, CT, ²UConn Health, Molecular Biology and Biophysics, Farmington, CT, ³Washington University, Pathology and Immunology, St Louis, MO

According to the CDC, each year viral respiratory infections cause over 60,000 deaths in the USA and hundreds of thousands of deaths worldwide. New interventions may be found in the respiratory microbiome, because the microbiome is vital to immune regulation. However, the effects of microbial colonization are species specific. Due to extensive biodiversity in the respiratory tract, host-microbe interactions are poorly understood. My goal is to characterize the effects of microbial colonization of the respiratory epithelium. I cultivated 56 genetically diverse microbes obtained from the respiratory tract of lung cancer patients. Each microbial isolate was applied to tracheal/bronchial air-liquid interface (ALI) tissue cultures (n=3). At 18 and 48 hours, RNA sequencing of the ALI cultures was conducted. The RNA sequencing data was analyzed using differential gene expression analysis and gene set enrichment analysis. My analysis focused on innate immunity and epithelial barrier integrity functional classes. Microbial phylogenetic similarity did not guarantee similar gene expression changes. For example, one *Streptococcus* isolate caused gene expression changes that were overrepresented in a gene set involved in mucosal innate immunity. However, another isolate of the same species caused changes that were underrepresented. Heterogeneity was also seen for this gene set across different *Streptococcus* species. Secretion of cytokines was quantified. The most notable change in cytokine levels was seen in IL-8, which summons neutrophils to a site of inflammation. IL-8 levels were increased in almost all samples. Transepithelial electrical resistance (TEER), which is a measure of epithelial permeability, was measured for a subset of microbes. Colonization with *Lactobacillus rhamnosus* increased TEER, implying improved epithelial barrier integrity, while colonization with a more aggressive lab strain of *Staphylococcus epidermidis* decreased TEER. In conclusion, the effects of microbial colonization on the respiratory epithelium are isolate dependent, with variation even between isolates of the same species. Colonization with some microbes in the respiratory tract appears to improve immune defense while others decrease it. Future studies will include co-colonization of ALI with microbial isolates and influenza to determine the effect of different microbes on influenza infection and the mechanisms that cause these changes.

HUMAN DISEASE GENETICS AND REAL-WORLD PATIENT DATA FUEL TARGET AND DRUG DISCOVERY IN EUROPEAN AND AFRICAN AMERICAN ANCESTRIES

Yuan Hou¹, Pengyue Zhang², Jeffrey Cummings³, Andrew A Pieper⁴, James B Leverenz⁵, Feixiong Cheng¹

¹Cleveland Clinic, Lerner Research Institute, Cleveland, OH, ²Indiana University, Department of Biostatistics and Health Data Science, Indianapolis, IN, ³University of Nevada Las Vegas, Department of Brain Health, Las Vegas, NV, ⁴Case Western Reserve University, Department of Neuroscience, Cleveland, OH, ⁵Cleveland Clinic, Neurological Institute, Cleveland, OH

High-throughput DNA/RNA sequencing technologies have generated massive genetic and genomic data in human disease, and selecting genetically supported targets can double the success rate in clinical drug development. However, translation of these findings into new patient treatment has not materialized, in particular for race-conscious target identification from diverse population genetics data. To address this problem, we have used Mendelian randomization (MR) and large patient's genetic and functional genomic data to evaluate druggable targets using Alzheimer's disease (AD) as a prototypical example. We utilized the genetic instruments from 9 expression quantitative trait loci (eQTL) and 3 protein QTL (pQTL) datasets across five human brain regions from three human brain biobanks: a) Mount Sinai Brain Bank, b) Mayo clinic, and c) Religious Orders Study (ROS) or the Rush Memory and Aging Project (MAP). We tested the outcome of MR independently across 7 genome-wide association studies (GWAS) datasets of European-American (EA) and African-American (AA) ancestries, with 275,540 Alzheimer's disease (AD) cases and 1.55 million controls. We identified 25 drug targets in EAs and 6 new drug targets in AAs among 1,176 investigated druggable targets. Among 6 AA-specific targets, we identified that TRPV3 is a potent drug target for AD individuals with AAs and replicated our finding in AA-specific eQTL data from the metabrain database. We further identified that elevated TRPV3 expression in the human cortex was significantly associated with both Braak stage and clinical cognitive diagnosis dcfdx scores, revealing important clinical roles of TRPV3 in AD. We also identified 23 candidate drugs associated with reduced risk of AD in mild cognitive impairment (MCI) patients after analysis of ~80 million electronic health records. We identified that usage of either apixaban (hazard ratio [HR] = 0.74, 95% confidence interval [CI] 0.69 – 0.80) and amlodipine (HR = 0.91, 95% CI 0.88 – 0.94) were both significantly associated with reduced progression to AD in people with MCI, mechanistically supported by MR analysis. In summary, combining genetics and real-world patient data identified ancestry-specific therapeutic targets and medicines for AD and other complex diseases if broadly applied.

COMPREHENSIVE MAP OF INTRORESSED STRUCTURAL VARIATION IN THE HUMAN GENOME

PingHsun Hsieh¹, William T Harvey², Katherine M Munson², Kendra Hoekzema², Francois-Xavier Ricaut³, Nicolas Brucato³, Irene G Romero⁴, Murray Cox⁵, Evan E Eichler²

¹University of Minnesota, Genetics, Cell Biology, and Development, Twin Cities, MN, ²University of Washington, Genome Sciences, Seattle, WA,

³University of Toulouse III Paul Sabatier, Department Evolution & Biological Diversity, Toulouse, France, ⁴University of Melbourne, Melbourne Integrative Genomics, Parkville, Australia, ⁵Massey University, School of Natural Sciences, Palmerston North, New Zealand

Genetic Introgression is a ubiquitous phenomenon across the tree of life. In humans, there is ample evidence of admixture events occurring between modern and archaic humans, such as the Neanderthal and Denisovan. Single nucleotide variation (SNV) data indicate that on average 2-4% of the present-day non-African human genomes derive from Neanderthal and/or Denisovan DNA. However, the actual genetic contribution of archaic hominins to modern humans remains elusive because structural variants, which alter at least 5X more bases than SNVs in the genome, have yet to be systematically studied due to the limitation of short-read sequencing data. Here we leverage high-quality haplotype-phased long-read assemblies from the 47 publicly available genomes of the Human Pan-genome Reference Consortium and two newly generated Melanesian genomes to build a comprehensive map of introgressed archaic variants that include both SNVs and SVs. 724,021 SVs (≥ 50 base pairs [bps]) and 18,771,089 SNVs from these 49 genomes are included in this study; of which 13,126 SVs are found explicitly in the two new Melanesian genomes. To resurrect introgressed segments from these modern-day humans, we used genotypes from high-coverage Neanderthal (n=3) and Denisovan (n=1) genomes and complementary haplotype-based and allele frequency-based methods to increase specificity. While the amount of archaic hominin DNA varies from 0.52% to 4.31% in our modern-day genomes, the Melanesian carries the highest amount (n=2, mean%: 3.97%), followed by East Asian (n=5, mean%: 2.14%), South Asian (n=1, 2.12%), European (n=1, 1.48%), and Admixed Americans (n=16, mean%: 1.12%). Melanesians also carry the highest amount of introgressed SVs (mean: 1,157 SVs; 486 deletions, 671 insertions), which together affect 1,264,492 bps in the genome. In contrast, Admixed Americans have the least amount of introgressed SVs (mean: 299 SVs; 125 deletions, 174 insertions) affecting only 186,686 bps. Of note, ~50% of introgressed SVs overlap with genes, including a previously reported 383 kbp positively selected Denisovan-introgressed insertion in the Melanesian. We will genotype and discuss the frequencies of these introgressed SVs in a large, short-read genome cohort of >1,000 samples, including 250 Melanesians, and highlight their possible adaptive roles and functional properties. Our study represents one of the most comprehensive genomic surveys to date for evidence of archaic SV introgression in anatomically modern humans outside Africa.

CELL DIVISION HISTORY ENCODES DIRECTIONAL INFORMATION OF FATE TRANSITIONS

Kun Wang^{1,4}, Liangzhen Hou¹, Xionglei He², Christina Curtis³, Da Zhou⁴,
Zheng Hu¹

¹Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Institute of Synthetic Biology, Shenzhen, China, ²Sun Yat-Sen University, School of Life Sciences, Guangzhou, China, ³Stanford University School of Medicine, Department of Medicine, Stanford, CA,
⁴Xiamen University, School of Mathematical Sciences, Xiamen, China

Single-cell RNA-sequencing (scRNA-seq) enables systematic mapping of cellular differentiation trajectories. However, accurately inferring the cell-fate transitions in the context of diseases or perturbations remains challenging due to high cellular plasticity. Here, we introduce a computational framework (PhyloVelo) that leverages monotonically expressed genes (MEGs) during cell divisions to estimate the past extrapolated state of single cells. Using simultaneous scRNA-seq and lineage information, PhyloVelo can identify MEGs and reconstruct a novel transcriptomic velocity field. PhyloVelo accurately recovered linear, bifurcated and convergent differentiations in simulation and *C. elegans* data. Application to eight lineage tracing datasets including CRISPR/Cas9 editing, lentiviral barcoding and lymphocyte receptor repertoires showed that PhyloVelo can robustly infer complex lineage trajectories with superior performance relative to RNA velocity. We also found MEGs across tissues and organisms had similar functions in translation and ribosome biogenesis, indicating an internal cellular clock. Together, our study presents a powerful method for cell-fate analysis in diverse biological contexts. Documentation and detailed examples for using PhyloVelo are available at <https://phylovelo.readthedocs.io/>.

LINKING PROCESS-SPECIFIC POLYGENIC RISK SCORES OF LIPODYSTROPHY TO DIFFERENTIATION-DEPENDENT ADIPOCYTE SUBPOPULATIONS

Yi Huang¹, Joaquín Pérez-Schindler^{*1}, Thiago M Batista^{*1}, Hesam Dashti^{1,2}, Miriam S Udler^{1,2,3}, Melina Claussnitzer^{1,2,3,4}

¹Broad Institute of MIT and Harvard, Metabolism Program, Cambridge, MA,

²Harvard Medical School, Department of Medicine, Boston, MA,

³Massachusetts General Hospital, Center for Genomic Medicine, Boston, MA,

⁴Hebrew SeniorLife and Harvard Medical School, Institute for Aging Research, Boston, MA

*These authors contributed equally

The individual polygenic risk of type 2 diabetes and insulin resistance, particularly lipodystrophy, is modified by changes to the mass, distribution and function of adipose tissue. The effects of metabolically-relevant genetic variation on adipocyte molecular and cellular programs are highly cell stage- and subpopulation-dependent, and require to study the polygenic context-dependent gene regulation at single cell resolution. Here, we use a combination of bulk and single nucleus multiomics joint profiling to map adipocyte subpopulations across differentiation and link these to gene regulatory networks of polygenic metabolic disease risk. We first profiled donor-paired subcutaneous and visceral adipose-derived mesenchymal stem cells derived from 26 individuals across four differentiation stages (Day 0, 3, 8 and 14) using bulk RNA-seq. We used process-specific polygenic risk scores (PRSSs) of lipodystrophy to ascertain the effect of polygenic risk for lipodystrophy on gene regulatory networks and cell states. We found a cell stage- and depot-specific PRS effect on gene expression in differentiated subcutaneous adipocytes. We observed that the score associates with genes enriched in fatty acid beta oxidation and mitochondrial activity. To link the PRS to specific adipocyte subpopulations, we established paired single nucleus gene expression and chromatin accessibility profiles across adipocyte differentiation using 10X Multiome in subcutaneous adipocytes. We identified in total nine adipocyte subpopulations, including two pre-adipocyte populations, five intermediately differentiated adipocyte populations, and two fully differentiated adipocyte populations, and validated their differentiation status using Cytotrace. Using this single nucleus multiomics profile as an anchor, we performed BayesPrism deconvolution to jointly predict cellular composition from the donor-paired adipocyte bulk RNA-seq data described above. The data confirmed a differentiation-dependent shift in the proportion of individual adipocyte subpopulations in both depots. Ascertaining the effect of lipodystrophy PRS on different adipocytes subpopulations, we found that the scores were specifically associated with an increased proportion of a small fully differentiated subcutaneous adipocyte subpopulation at day 14 of differentiation. This subpopulation was characterized by lipolysis and mitochondrial function-related gene signatures. We validated this PRS effect on mitochondrial cellular processes using our novel high-dimensional image-based profiling tool called LipocyteProfiler. In conclusion, using joint bulk and single nucleus multiomics together with high content image-based profiling we found a novel association between clinically informed lipodystrophy PRS and a subcutaneous adipocyte population that act on mitochondrial activity at differentiation day 14.

INVESTIGATING THE EFFECTS OF *INDY* REDUCTION ON A FLY MODEL OF ALZHEIMER'S DISEASE

Billy J Huggins, Jacob Macro, Kali Meadows, Blanka Rogina

UConn Health, Genetics and Developmental Biology, Farmington, CT

Alzheimer's Disease (AD) is a debilitating, age-related disease, characterized by cognitive and functional decline. The preclinical and clinical stages are marked by metabolic dysfunction, oxidative stress, accumulation of amyloid plaques, and comorbid pathologies. One striking presentation of AD is temporoparietal hypometabolism, indicative of reduced glucose metabolism and neuronal damage. We propose to study the potential of *Indy* reduction to ameliorate multiple aspects of AD pathology as a novel AD intervention. *Indy* codes for a plasma membrane citrate transporter and genetic reduction has been demonstrated to extend longevity in flies and worms. *Indy* is a fly homologue of the SLC13A5 mammalian citrate transporter, and its function is highly conserved across species. Reduced *Indy* levels in flies, worms, and *Indy* knock-out in mice, alter metabolism in a manner similar to calorie restriction. This results in reduced glucose levels, increased insulin sensitivity, reduced oxidative damage, increased mitochondrial biogenesis, preserved memory, and more. Taken together, evidence from flies and other model systems suggests that *Indy* reduction may rescue many of the detrimental effects associated with AD pathology including metabolic and mitochondrial dysfunction. We are using a fly model of human AD, which contains APPswe mutation, in order to study *Indy*'s potential as a novel intervention for AD. Preliminary data in the lab demonstrates that we can drive APPswe mutation in neuronal and glial cells of flies, which leads to reduced longevity. Strikingly, when *Indy* activity is reduced in APPswe flies, lifespan is largely rescued. Preliminary data has also revealed that *Indy* reduction can ameliorate some of the metabolic changes associated with APPswe overexpression. To further investigate the effects of *Indy* reduction on AD, future experiments include mRNA sequencing to obtain an unbiased view of genes and pathways modulated in APPswe/*Indy* flies. We expect to observe transcriptional differences in genes associated with oxidative phosphorylation, insulin-like signaling, among others, in APPswe/*Indy* flies, indicating improvements in metabolism and resembling longevity associated transcriptional profiles found in *Indy* flies.

SINGLE-NUCLEOID ARCHITECTURE REVEALS HETEROGENEOUS PACKAGING OF HUMAN MITOCHONDRIAL DNA

R. Stefan Isaac¹, Thomas W Tullius^{*1}, Katja G Hansen^{*1}, Danilo Dubocanin², Mary Couvillion¹, Andrew B Stergachis^{2,3}, L. Stirling Churchman¹

¹Blavatnik Institute, Harvard Medical School, Department of Genetics, Boston, MA, ²Division of Medical Genetics, University of Washington, Department of Medicine, Seattle, WA, ³University of Washington, Department of Genome Sciences, Seattle, WA

Cellular metabolism relies on the regulation and maintenance of mitochondrial DNA (mtDNA), an extrachromosomal genome contained within the mitochondrial network. Hundreds to thousands of copies of mtDNA exist in each cell, yet because mitochondria lack histones or other machinery important for nuclear genome compaction, it remains unresolved how mtDNA is packaged into individual nucleoids. In this study, we used long-read single-molecule accessibility mapping to measure the compaction of individual full-length (16.5 kb) mtDNA molecules at nucleotide resolution. We found that, unlike the nuclear genome, human mtDNA largely undergoes all-or-none global compaction, with the majority of nucleoids existing in an inaccessible, inactive state in multiple cell types. The fraction of highly accessible mitochondrial nucleoids is co-occupied by transcription and replication machinery and selectively forms a triple-stranded D-loop structure, which is formed by paused replication initiation. In addition, we showed that the primary nucleoid-associated protein TFAM directly modulates the fraction of inaccessible nucleoids both *in vivo* and *in vitro* and acts via a nucleation-and-spreading mechanism, preferentially binding higher affinity sites throughout the genome and spreading to coat and compact mitochondrial nucleoids. Together, these findings reveal the primary architecture of mtDNA packaging and regulation in human cells.

PRODUCING ARTIFICIAL ANTIBODIES FROM FFPE LUNG TISSUE

Sadahiro Iwabuchi, Shinichi Hashimoto

Wakayama Medical University, Department of Molecular Pathophysiology,
Wakayama, Japan

Recombinant monoclonal antibodies generated by the B cell receptor (BCR) repertoire of patients have been useful for treating the respiratory syncytial virus infection. Rearrangement of the BCR encoding genes induces recombination of variable (V), diversity (D), and joining (J) segments of the third complementarity determining region (CDR3), leading to considerable diversification. Understanding the diversity of BCRs, their response to virus infection, and detection of specific BCRs may help in the development of therapeutic antibodies for patients. Candidate antibodies are usually obtained by analyzing BCRs in the peripheral blood mononuclear cells (PBMCs) obtained from patients with virus infectious disease compared with those from healthy donors (HDs). Here, BCR repertoire analysis was performed using a unique method than the conventional approach by analyzing of BCRs in FFPE lung lobes of COVID-19 patient. We developed several artificial antibodies using pairs of IgG heavy and light chains, which were frequently detected in the inflamed lung lobe. The B-cells and plasma cells existed in the inflamed lung lobes; therefore, we hypothesized the presence of antibodies with neutralizing activity against SARS-CoV-2 more in the lung lobes. To realize the significance of the detected BCR repertoires, single-cell BCR (scBCR) repertoires obtained from the PBMCs of HDs who had recovered from COVID-19 or had received the mRNA vaccine twice or had not received the vaccine were also analyzed. The pair of highly expressed IgH and IgK chains in FFPE samples (*IGHV1-69/IGKV2-28*, *IGHV1-69/IGKV3-20*) were not specifically expressed in the PBMCs of HDs, indicating the BCRs may be novel antibody candidates that cannot be highly detected by conventional methods or stimulated by vaccination. Moreover, we evaluated SARS-CoV-2 neutralizing activity along with the artificial antibodies. The results of this study shed light on the development of vaccines and neutralizing antibodies using FFPE samples against future unknown emerging infectious disease.

ASSESSING TISSUE-SPECIFIC EFFECTS OF RARE AND STRUCTURAL VARIANTS TOWARDS GENE REGULATION WITH THE EN-TEX PERSONAL GENOME RESOURCE

Matthew Jensen^{1,2}, Tai Michaels¹, Anna Su², Timur Galeev^{1,2}, Sushant Kumar^{1,2}, Kun Xiong^{1,2}, Beatrice Borsari^{1,2}, Joel Rozowsky^{1,2}, Mark Gerstein^{1,2}

¹Yale University, Molecular Biophysics and Biochemistry, New Haven, CT,

²Yale University, Computational Biology and Bioinformatics, New Haven, CT

Comprehensive tissue-specific analyses of how variants alter molecular phenotypes have greatly improved our understanding of genomic mechanisms for complex traits. For example, the EN-TEx resource, consisting of long read-based personal genomes and a full battery of functional assays across 25 tissues in four donors, allowed us to systematically assess the allelic activity of common SNVs. The long-read personal genomes in the EN-TEx resource are also ideal for studying the functional effects of a full spectrum of genomic variants, in particular rare SNVs and all types of structural variants (SVs), which are often under-represented in functional genomics studies. In this study, an extension of work from the EN-TEx Consortium, we aligned 381 tissue-specific and single-cell ATAC-Seq, DNase-Seq, and ChIP-Seq datasets within EN-TEx onto personal genomes, and prioritized functional signals within variant regions adjacent to genes with altered expression. Alternate-aware mapping of functional sequencing data to individual haplotypes allowed us to better identify peaks in heterozygous regions, especially in novel insertions distinct from the reference genome. In these datasets, we found ~293 SV-eQTLs per individual and linked key functional signals within these variants to genes with altered expression; for instance, deletion of an upstream H3K27Ac peak led to reduced expression of ZFAND2A. We also specifically identified ~152 functional elements within novel genomic insertions per individual, and found that 16% of these variants could contribute to perturbed expression patterns. For example, the tumor suppressor ASMTL-AS1 showed 4.2-fold increased expression when coupled with an upstream duplication spanning a novel ATAC-Seq peak. Beyond SVs, we identified ~620 rare SNVs per individual that disrupted candidate cis-regulatory elements (cCREs) near protein-coding genes, and prioritized variants in cCREs located near genes with tissue-specific altered expression. Overall, our study emphasizes the broad effects of both rare variants and SVs towards tissue-specific gene regulatory patterns, and highlights the utility of using personal genomes for accurate and comprehensive functional genomics studies.

NOVEL EPISTATIC INTERACTIONS BETWEEN *KIT* AND *MITF* CAUSE BREAKTHROUGH PIGMENTATION IN REGIONS OF WHITE ON THE COAT IN CATTLE

Swati Jivanji¹, Anna Yeates¹, Chad Harland¹, Charlotte Gray¹, Christine Couldrey¹, Gemma Worth¹, Isabelle Gamache², John A A Tabares², Lorna McNaughton¹, Marie-Pier Cloutier¹, Jade Desjardins², Mitra Crowan², Tony Fransen¹, Tracey Monehan¹, Richard Spelman¹, Richard Mort³, Yojiro Yamanaka^{2,4}, Mathew D Littlejohn^{1,5}

¹Livestock Improvement Corporation, Research & Development, Hamilton, New Zealand, ²McGill University, McGill Integrated Core for Animal Modeling, Montreal, Canada, ³Lancaster University, Division of Biomedical and Life Sciences, Lancaster, United Kingdom, ⁴Goodman Cancer Institute, Department of Human Genetics, Montreal, Canada, ⁵Massey University, School of Agriculture and Environment, Palmerston North, New Zealand

The white-face trait characteristic of Hereford cattle is assumed to be dominantly inherited, where the inheritance of just one copy of the candidate causal serial duplication upstream of the *KIT* gene is sufficient to cause the white-face trait. This mutation has been proposed to modulate *KIT* expression and prevent pigment expression in the face. However, a proportion of Hereford crossbred calves appear to have ‘rescued’ pigmentation in the face. A genome-wide association analysis (GWAS) for face colour in 128 Hereford-cross calves using genotype data from a medium density SNP chip found that the adulteration of the white-face trait is caused by an epistatic interaction between mutations at the *KIT* and *MITF* loci. We found that a single copy of a mutation within a highly conserved SOX10 transcription factor binding site in the *MITF* promoter region was sufficient to rescue pigmentation, suggesting that the white-face trait is not technically dominant. The same *MITF* mutation was also found to be the most significantly associated variant in a GWAS investigating speckling within white spots in 256 white spotted Jersey and Holstein x Jersey crossbred bulls. We propose that the *MITF* mutation interacts with another mutation at the *KIT* locus responsible for white spotting on the coat in Jersey and Holstein x Jersey crossbred cattle and partially rescues pigmentation in these regions. Targeted editing of this variant in mice supported the causality of this mutation. Somewhat surprisingly, different mutations at the transcription factor binding site created large phenotypic diversity in the edited mice, speaking to the critical role this regulatory element plays in coat colour and patterning determination. Our results demonstrate how a single regulatory mutation, inherited on different genetic backgrounds, can interact with mutations at the *KIT* locus to generate phenotypic diversity in cattle and enhance our understanding of the molecular mechanisms that influence these traits more broadly.

APPLICATION OF NETWORK-BASED HETEROGENEITY
CLUSTERING FOR INVESTIGATION OF GENOTYPE-PHENOTYPE
CORRELATIONS IN BIOME BIOBANK

Meltem Ece Kars¹, Yiming Wu¹, Cigdem Sevim Bayrak², Bruce D Gelb^{2,3,4},
Yuval Itan^{1,2}

¹Icahn School of Medicine at Mount Sinai, The Charles Bronfman Institute for Personalized Medicine, New York, NY, ²Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences, New York, NY, ³Icahn School of Medicine at Mount Sinai, Department of Pediatrics, New York, NY, ⁴Icahn School of Medicine at Mount Sinai, Mindich Child Health and Development Institute, New York, NY

The genetic basis of human diseases often exhibits heterogeneity, resulting in different genes in the same or related biological pathways being responsible for similar or close phenotypes. Conventional frequency-based gene burden methods, which are used to identify genetic signals in case-control studies, lack statistical power for significant associations in small and genetically heterogeneous cohorts. To address this issue, we have previously developed the network-based heterogeneity clustering (NHC) method, that can detect genetic heterogeneity in physiologically homogenous cohorts with small sample sizes, and demonstrated the method's capacity to effectively converge genes that are biologically related on a protein-protein interaction network and accurately identify gene clusters with potentially deleterious variants. Here, we utilized NHC to effectively analyze three disease cohorts from the Mount Sinai BioMe BioBank, where previous gene burden approaches were ineffective in identifying candidate genes. The cohorts comprised 50 individuals with lactation disorders (agalactia or hypogalactia) and 500 controls, 71 patients with severe COVID-19 and 390 controls with mild-moderate disease, and 431 children with food allergy and 2,155 controls. NHC analysis in the lactation disorders cohort identified gene clusters involved in intra-golgi vesicle transport, mammary stem cell differentiation and proliferation, and trans-differentiation of white adipocytes in the mammary gland during lactation. NHC also uncovered two gene clusters potentially related to increased risk for severe COVID-19, including genes that preserve endothelial barrier function and genes involved in snRNA processing, which have previously been implicated in susceptibility to HSV-1 infection. Furthermore, NHC facilitated the identification of candidate gene clusters in the food allergy cohort with population-specific signatures, including genes proposed to be associated with allergic asthma and food allergies, such as DYNC1H1, toll-like receptor encoding genes, genes involved in IL-17 signaling and cell adhesion, and genes encoding mitochondrial ribosomal proteins. These findings indicate that NHC is a powerful approach for uncovering candidate genes in disease groups with genetic heterogeneity, especially in smaller population-specific subgroups, outperforming traditional case-control studies for such cohorts.

SINGLE CELL ANALYSIS OF IMMUNE CELL LANDSCAPE OF HEALTHY INDIVIDUALS

Yukie Kashima, Yutaka Suzuki

The University of Tokyo, Department of Computational Biology and Medical Sciences, Kashiwa, Japan

It is well known that the immune cells are collectively responsible for the diverse immune responses of an individual by shaping the immune landscape. Immune responses differ between individuals and personal health history and unique environmental conditions should collectively determine the current state of immune cells. However, the molecular systems underlying such heterogeneity remain elusive. To address these questions, we performed a systematic single-cell analysis, scRNA-seq, scVDJ-seq and CyTOF, in healthy individuals.

First, as a small pilot study, we collected datasets from seven healthy individuals. We found substantial diversity in immune cell profiles between different individuals. These patterns showed daily fluctuations even within the same individual. Similar diversities were also observed for the T cell receptor and B cell receptor repertoires.

Second, we collected PBMCs during influenza and SARS-CoV-2 vaccination. We hypothesized that immune cell profiles from healthy individuals provide an essential information to understand complex immune landscapes, when the individual is exposed to different environmental conditions. PBMCs from vaccinated individual showed common key responses and individual-dependent responses.

Third, based on these results, we expanded our study to 50 individuals who lived in the same area, Japan. The donors included in the current cohort is in their 50s- 80s. We collected their medical history and daily habits together with their PBMCs. These datasets highlighted age-associated diversity and individual-dependent diversity.

As a conclusion, our single cell immune cell profile data provide a fundamental data resource to understand variable immune responses, which are unique to each individual.

ROBUST DIFFERENTIAL EXPRESSION TESTING FOR SINGLE-CELL CRISPR SCREENS

Timothy Barry¹, Kathryn Roeder^{1,2}, Eugene Katsevich³

¹Carnegie Mellon University, Department of Statistics and Data Science, Pittsburgh, PA, ²Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA, ³University of Pennsylvania, Department of Statistics and Data Science, Philadelphia, PA

Single-cell CRISPR screens have emerged as a standard tool for mapping genetic perturbations to phenotypic changes in single cells. A fundamental task in single-cell CRISPR screen data analysis is to test for differential expression (DE), i.e. association between a CRISPR perturbation and the expression of a gene or protein. There does not currently exist a standard methodology for carrying out this task, with several distinct approaches being applied in practice. Furthermore, there does not exist an established framework for evaluating and comparing the calibration of single-cell CRISPR screen DE methods. To address these gaps, we first propose a simple framework for evaluating the calibration of such a method. This framework uses negative control perturbations, while ensuring unbiased evaluation of methods that are themselves based on negative controls. Second, focusing on low multiplicity-of-infection (MOI) screens (which account for the vast majority of existing single-cell CRISPR screens), we use this framework to conduct the first-ever comprehensive benchmarking study of single-cell CRISPR screen DE methods, applying five leading methods to analyze six diverse datasets. We found that existing methods exhibit significant miscalibration, leaving them prone to an excess of false discoveries. Third, we conducted an extensive investigation into the causes of this miscalibration, identifying three core analysis challenges: data sparsity, confounding, and expression model misspecification. No existing method adequately deals with all three challenges. Fourth, we developed a principled resampling-based statistical methodology (SCEPTRE, an extension to the low MOI setting of a methodology with the same name we recently developed for high MOI) that does address all three of the aforementioned challenges, and verified its excellent calibration on all six datasets under consideration. On average across datasets, SCEPTRE reduces the number of false positive findings compared to the best existing method by a factor of 10 (from 7.1 to 0.7). Using positive control data, we also found that SCEPTRE is not only well-calibrated but also has better sensitivity than existing methods to detect true perturbation-expression associations. Therefore, SCEPTRE makes fewer false discoveries and more true discoveries compared to existing methods. Fifth, we developed an efficient, user-friendly, open-source software implementation of SCEPTRE handling both low and high MOI single-cell CRISPR screen data.

CHROMATIN MODIFIERS AS DRIVERS IN ENDOMETRIAL CANCER

Katarzyna Z Kedzierska¹, Yannick Comoglio¹, Matthew W Brown¹, Endometrial Cancer GeCIP Domain², Dan J Woodcock³, David N Church¹

¹University of Oxford, Wellcome Centre for Human Genetics, Oxford, United Kingdom, ²Genomics England, 100,000 Genome Programme, London, United Kingdom, ³University of Oxford, Nuffield Department of Surgical Science, Oxford, United Kingdom

Chromatin organization, a central regulator of gene expression and cell phenotype, is recurrently dysregulated in cancer. Accumulating evidence suggests that chromatin dysregulation through genomic alteration of its modifiers is particularly common in endometrial cancer (EC); the most common gynecological malignancy and a cause of substantial morbidity and mortality. *ARID1A* - a key component of the SWI/SNF remodeling complex - is one of the most frequently mutated EC drivers, while remodelers, such as *CHD4*, have also been identified as drivers in EC. However, systematic analysis of alteration in chromatin modifiers and their functional consequences in EC has not been performed.

We analyzed whole genome sequence (WGS) data from EC in the Genomics England (GEL) 100,000 Genomes Project (n=665) along with WES and RNA-seq from The Cancer Genome Atlas (TCGA) pan-cancer (n=10,295) and EC (UCEC and UCS, n=586) cohorts. Driver analysis of GEL EC by IntOGen identified 15 chromatin modifiers as EC drivers, including known EC drivers such as *ARID1A*, *ARID1B*, *ARID5B*, *CTCF*, and additional pan-cancer drivers *CREBBP*, *SETD1B*, *SETDB1*, and *SMARCA4* new to EC. *ARID1A* mutations are highly over-represented in EC (TCGA EC vs. pan-cancer cohort, Fisher's Exact Test, Odds ratio=10.53, p-value<2.2e-16), with 40-53% of samples harboring a mutation. Other chromatin modifiers are frequently altered. For example, *BCOR*, a member of the PRC1.1 complex, is mutated in 18% of all EC samples with a recurrent N1459S mutation in 9% of samples near-unique to EC.

To further investigate the role of chromatin remodelers in EC, we complemented our genomic data with gene editing of normal and malignant endometrial cells by CRISPR-Cas9. We generated chromatin accessibility profiles (ATAC-seq) and RNA-seq data for key chromatin modifiers in the endometrial cancer-derived cell line (Ark2, n=21) and human endometrial epithelial cell line (hEM3, n=6). For example, our analysis reveals that with loss of *ARID1A* in the EC cancer cell line, genes associated with KRAS signaling are upregulated. In contrast, in the hEM3 cell line, genes related to Epithelial-Mesenchymal Transition are upregulated. We further characterize chromatin accessibility profiles and compare the effects of loss of top chromatin modifiers on accessibility of transcription factor binding, enhancer and transcription start sites.

Alterations in chromatin remodelers are highly recurrent in EC, with approx. two-thirds of cases harboring mutations in at least one gene. Our functional analyses by gene editing and ATAC confirm that EC-associated drivers disrupt cellular processes central to oncogenesis; comprehensive interrogation is currently underway. Our study provides the first systematic correlative and functional analyses of chromatin modifiers in EC and provides new insights into EC biology.

COMMON GENETIC VARIATION RELATED TO SCHIZOPHRENIA IS ASSOCIATED WITH COGNITIVE PERFORMANCE IN CHILDREN AND ADOLESCENTS

Gianluca C Kikidis^{1,2}, Alessandra Raio¹, Nora Penzel¹, Leonardo Sportelli¹, Linda A Antonucci¹, Alessandro Bertolino^{1,3}, Qiang Chen², Pierluigi Selvaggi^{1,3}, Antonio Rampino^{1,3}, Giulio Pergola^{1,2}

¹Group of Psychiatric Neuroscience, Department of Translational Biomedicine and Neuroscience, University of Bari Aldo Moro, Bari, Italy,

²Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, ³Psychiatry Unit, Bari University Hospital, Bari, Italy

Genetic risk for schizophrenia (SZ) is associated with cognitive performance before the age of onset. Genetic variants have been identified that promote resilience to SZ. We expect that risk and resilience polygenic variants would exert divergent effects on the same cognitive performance tasks from a young age. We tested our hypothesis in healthy Caucasian adolescents, children and adults.

We included 5120 (49.5% female) adolescents (age 8-22 y, m: 13.8, sd: 3.6) from the Philadelphia Neurodevelopmental Cohort including cognitive tasks in social cognition (SC), executive functions (EF), long-term memory (LTM), attention and processing. Our replication samples included 5,264 (47.3% female) children (age 9-12 y, m: 9.92, sd: .63) from the Adolescent Brain and Cognitive Development cohort, and 995 (45.6% female) adults (age 18-65 y, m: 27.0, sd: 8.1) from Bari, Italy, who underwent neurocognitive tests within the same domains (except SC). We calculated polygenic scores (PGS) based on the latest GWAS from the Psychiatric Genomics Consortium and investigated their association with cognitive performance in terms of reaction time (RT) and accuracy via linear models. We also calculated the distance between risk and resilience effect sizes (Δt) to test divergence on cognitive performance.

In PNC we found higher risk PGS associated with slower RT in SC ($t = 4.5$, PFDR < .001), EF ($t = 3.27$, PFDR = .024), LTM ($t = 2.6$, PFDR = .042), and attention ($t = 2.6$, PFDR = .042). Higher resilience PGS were associated with faster RT in SC ($t = -2.9$, PFDR = .026) and processing ($t = -3.5$, PFDR = .003). We also found negative association of risk PGS with accuracy in EF ($t = -3.5$, PFDR = .002) and LTM ($t = -2.5$, PFDR = .049) in children and adults. Additionally, resilience PGS showed association with better EF ($t = 3.2$, PFDR = .014) and attention ($t = 3.7$, PFDR < .001) accuracy in adults. We found divergent effect sizes of risk and resilience in the domains of attention ($\Delta t_{PNC} = 4.2$, $\Delta t_{ABCD} = 2.0$, $\Delta t_{Bari} = 0.6$), EF ($\Delta t_{PNC} = 3.7$, $\Delta t_{ABCD} = 2.7$, $\Delta t_{Bari} = 4.7$), and LTM ($\Delta t_{PNC} = 2.5$, $\Delta t_{ABCD} = 2.0$, $\Delta t_{Bari} = 3.6$).

Our results suggest that cognitive skills are core aspects of SZ and genetically correlate both with risk and resilience before the onset age, aligning with the neurodevelopmental hypothesis of SZ.

INFERENCE OF ADMIXTURE ORIGINS IN INDIGENOUS AFRICAN CATTLE

Kwondo Kim^{1,2}, Donghee Kim³, Olivier Hanotte^{4,5,6}, Charles Lee¹, Heebal Kim^{2,7,8}, Choongwon Jeong³

¹The Jackson Laboratory for Genomic Medicine, Farmington, CT, ²Seoul National University, Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul, South Korea,

³Seoul National University, School of Biological Sciences, Seoul, South Korea, ⁴International Livestock Research Institute, Addis Ababa, Ethiopia,

⁵The Roslin Institute, The Centre for Tropical Livestock Genetics and Health (CTLGH), Midlothian, United Kingdom, ⁶University of Nottingham, School of Life Sciences, Nottingham, United Kingdom, ⁷Seoul National University, Interdisciplinary Program in Bioinformatics, Seoul, South Korea, ⁸eGnome, Seoul, South Korea

Present-day African cattle retain a unique genetic background compared to the other *Bos taurus* and *Bos indicus* populations across the world. Multiple admixture events in the continent are considered one of the major components that contributed to it, but their origins remain unclear. Here, we infer the source of admixture in the earliest domestic cattle in Africa, the African taurine, and reveal a significant amount of admixture (~50%) from a basal taurine lineage, which may explain why present-day African taurine shows such a significant divergence from the other taurine cattle populations. In addition, we infer the origins of the indicine admixture in African cattle in a fine resolution to distinguish different indicine sources. The results suggest that the indicine source in the African cattle populations is a mixture of two ancestries, present today in North India and South China, respectively. Our findings support the hypothesis of local auroch introgression into African taurine and generate a novel hypothesis regarding the Asian origin of *Bos indicus* on the continent, which may have involved more than one admixture event.

RANDOMIZING THE HUMAN GENOME BY ENGINEERING RECOMBINATION BETWEEN REPEAT ELEMENTS

Jonas Koeppel¹, Raphael Ferreira^{2,3}, Fabio G Liberante⁴, Thomas Vanderstichele¹, Gareth Girling¹, George Church^{2,3}, Tom Ellis^{1,5}, Leopold Parts^{1,6}

¹Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom,
²Harvard Medical School, Department of Genetics, Boston, MA, ³Harvard University, Wyss Institute for Biologically Inspired Engineering, Boston, MA, ⁴EMBL's European Bioinformatics Institute, Elixir, Hinxton, United Kingdom, ⁵Imperial College London, Department of Bioengineering, London, United Kingdom, ⁶University of Tartu, Department of Computer Science, Tartu, Estonia

While our understanding of protein-coding genes has improved considerably, the organization and function of the non-coding genome remain comparatively poorly characterized. This gap is due to a lack of tools for engineering variants at the magnitude needed to interrogate gigabases of the human genome. To enable such experiments, we have generated a toolbox to create deletions, inversions, translocations, and duplications at scale by highly multiplexed targeting of CRISPR prime editors to repetitive LINE1 elements and inserting recombinase recognition sequences. Using this strategy, we derived stable cell lines with up to 332 clonal site-directed insertions, the highest number in a single human genome that we are aware of. Subsequent recombinase induction in these cells generated hundreds of structural variants detected after one day using long-read whole genome sequencing. We observed median rearrangement lengths of 1.4 megabases and retrieved 112 variants spanning more than ten megabases. At high recombinase concentrations, induction resulted in an average of more than thirty simultaneous rearrangements per cell. The scrambling of the human genome by deleting, transposing, inverting, and inserting sequences will for the first time allow the direct interrogation of the human genome's organizing principles at the megabase scale.

SYSTEMATIC MUTAGENESIS REVEALS FUNCTIONAL COMPLEXITY OF HUMAN ENHancers IN VIVO

Michael Kosicki, Stella Tran, Jennifer A Akiyama, Ingrid Plajzer-Frick, Catherine S Novak, Yiwen Zhu, Momoe Kato, Anne Harrington, Riana D Hunter, Kianna von Maydell, Janeth Godoy, Eman M Meky, Sarah Bartonn, Erik Beckman, Diane E Dickel, Axel Visel, Len A Pennacchio

Lawrence Berkeley Lab, EGSB, Berkeley, CA

Distant-acting enhancers are central to development and the list of human disease phenotypes resulting from enhancer dysfunctions continues to grow. However, our understanding of human enhancer grammar remains limited. This represents a major hurdle for the interpretation of enhancer mutations identified in patients through whole-genome sequencing.

Here, we combined systematic mutagenesis of human enhancers *in vivo* with detailed assessment of the resulting activity changes through scaled transgenic mouse reporter assays. We selected seven human enhancers with reproducible *in vivo* activity patterns during development and systematically mutagenized their entire length in 12 bp blocks, approximately corresponding to the size of an individual transcription factor binding site. In total, we assessed the activity in multiple embryonic structures of ~300 mutagenized constructs, including constructs with different combinations of mutated tiles, in >3,000 transgenic mouse embryos.

Across all enhancers, we observed that 42% of individual block mutations caused a major loss of enhancer function, while 8% caused gains of function. Intriguingly, the robustness of individual enhancers to mutagenesis varied widely, with 21-78% of individual sequence block mutations causing substantial changes to function. This robustness measure correlated inversely with the number of gnomAD variants. Further, only two of seven enhancers harbored gain-of-function blocks, in both cases multiple independent ones. To investigate functional redundancy within enhancers, we mutagenized 17 pairs of sequence blocks which in isolation caused partial loss-of-function. The activity of combinatorial constructs was in most cases additive, leading to a more severe reduction, with little evidence of regulatory redundancy or buffering. To determine if single, non-functional blocks could exhibit functional changes in a sensitized background, we combined mutations from blocks that showed evidence of functionality in single-block experiments with ones that did not. We discovered that an additional 12% of blocks showed evidence of activity in this sensitized context. Finally, we resolved loss-of-function blocks to single base pairs critical for their *in vivo* function, validating our block mutagenesis strategy.

Taken together, our data indicate that human enhancers mutagenized at the level of 12 bp blocks *in vivo* are subject to substantial loss of function, with some enhancer robust and others fragile to such change. Through the generation of additional such datasets, we anticipate improved prediction of consequences of sequence changes in human enhancers, which is essential to inform whole human genome sequencing efforts in the clinic.

PREDICTION OF EFFECTOR PROTEIN STRUCTURES FROM FUNGAL PHYTOPATHOGENS ENABLES EVOLUTIONARY ANALYSES

Kyungyong Seong¹, Ksenia V Krasileva^{1,2,3}

¹University of California, Department of Plant and Microbial Biology, Berkeley, CA, ²University of California, Center for Computational Biology, Berkeley, CA, ³University of California, Innovative Genomics Institute, Berkeley, CA

Fungal pathogens encode hundreds of potential virulence factors called effectors. However, understanding their evolution had been difficult due to rapid loss of primary sequence similarity. We developed robust comparative genomic pipelines to identify sequence-unrelated structurally similar (SUSS) fungal effectors within and across fungal species and addressed long-standing questions: the origin and the evolution of the SUSS effectors. Using AlphaFold2, we predicted the structures of 26,653 secreted proteins from 14 fungal pathogens and 6 non-pathogens as well as an oomycete outgroup. About 70% of the secreted proteins could be modeled with confidence even without a large number of homologs. Structure-based clustering suggested that many pathogens preferentially expand specific protein folds, although different pathogens diversify distinct sets of folds. The most drastic example is a powdery mildew encoding about 400 RNase-like folds which represent a half of the pathogen's secretome. By identifying new classes of SUSS effectors characteristic to specific pathogens, we described mechanisms of rapid effector diversification, including point mutation, acquisition of unstructured loops and transcriptional reprogramming. Collectively, our study highlights diverse effector evolution mechanisms and supports divergent evolution as a major force in driving effector evolution from ancestral proteins shared with non-pathogens.

(CONTEXT-) TRANSCRIPTION FACTORS CREATE COOPERATIVE REGULATORY ENVIRONMENTS AND MEDIATE ENHANCER COMMUNICATION

Judith F Kribelbauer^{1,2}, Olga Pushkarev^{1,2}, Julie Russeil¹, Vincent Gardeux^{1,2}, Guido van Mierlo^{1,2}, Bart Deplancke^{1,2}

¹ EPFL, Institute of Bioengineering, School of Life Sciences, Lausanne, Switzerland, ² Swiss Institute of Bioinformatics, Lausanne, Switzerland

Enhancer regulation relies on the complex interplay of transcription factors (TFs) and cofactors. Despite many experimental tools that map epigenetic features in various cell types, we still lack a detailed understanding of enhancer function.

One poorly understood aspect is the role TFs play in enhancer communication. Previous experiments showed that the creation of high concentration environments, or condensates, is an important contributor. However, there is currently no tool that can reliably detect cooperative environments genome-wide. As a result, we lack mechanistic insight into how enhancer sequence instructs cooperative assemblies of TFs and cofactors both locally and across multiple elements.

To address this, we leveraged chromatin accessibility quantitative trait loci (caQTL) in lymphoblastoid cell lines to nominate cooperative enhancer environments. Specifically, we focused on enhancers that contain a causal variant in form of a single nucleotide polymorphism (SNP). SNPs affect the binding of at most a single TF, making it likely that enhancer-wide accessibility changes depend on local cooperativity with other TFs. Comparing caQTL enhancers to those containing ‘no-effect’ SNPs, we identify two classes of TFs: those enriched at the causal SNP, including previously discovered cell type-specific TFs, and those specific to the sequence context of caQTL enhancers. The latter group of TFs is largely uncharacterized and lacks an independent association with DNA accessibility. Hence, we name them ‘context-TFs’.

Using large-scale, autonomous transcription and enhancer activity assays, we find that although context-TFs do not independently drive transcriptional activity, they enhance it when added to cell type-specific motifs. Further, while motifs for both cell type-specific and context-TFs are associated with active enhancer marks, only the latter are predictive of high levels of the cofactor BRD4. Given BRD4’s role in condensate formation, we investigated whether context-TFs may mediate enhancer communication, revealing a significant enrichment of context motifs in enhancers that exhibit molecular coordination with a lead caQTL. We validated that this communication is direct by showing that, similar to Super Enhancers (SEs), genes controlled by coordinated enhancers are particularly sensitive to BRD4 inhibition. Upon stratifying SEs that contain caQTL enhancers into those with or without coordinated neighbors, we show that only the former truly cooperate, suggesting that not all SEs are made equal.

In summary, our analysis describes a new class of TFs, “context-TFs”, that potentiate the effect size of cell type-specific TFs and whose motif strength is associated with BRD4 recruitment. Cooperative environments created in this manner explain how enhancers communicate: they share the same functional grammar.

CLOUD-SCALE TRAINING AND EDUCATION IN THE NHGRI ANALYSIS, VISUALIZATION, AND INFORMATICS LAB-SPACE (ANVIL)

Natalie Kucher¹, Michael C Schatz^{1,2}, Anthony Philippakis³, AnVIL Team^{*1,2,3}

¹Johns Hopkins University, Department of Biology, Baltimore, MD, ²Johns Hopkins University, Department of Computer Science, Baltimore, MD,

³Broad Institute of MIT and Harvard, Cambridge, MA

* The full list of contributors is available at:

<https://anvilproject.org/about/team>.

As stakeholders in biomedical research increasingly adopt cloud computing, educational platforms and materials are needed to support the transition of genomic data science researchers to the cloud, for expert bioinformaticians and the next generation alike. Smoothly managing informatics training with software, data, and participant access as well as new elements relevant to the cloud, such as computing environment and billing are necessary for successful adoption.

The **NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space or AnVIL** (<https://anvilproject.org/>), is a secure, cloud-based platform for storage, management, and analysis of genomic and related datasets that offers many opportunities for supporting genomics research training in the cloud. A major component of AnVIL is Galaxy (<https://galaxyproject.org/>), an interactive analysis platform that empowers users with and without programming experience to perform complex bioinformatics analysis from a user-friendly web interface. Galaxy in AnVIL offers participant and billing management, dedicated computing resources, and a consistent but customizable software environment all within a security perimeter necessary for use of protected and private datasets. These advantages result in faster jobs, reduced data transfer, tracked class activity and spending, and flexibility for in person and virtual training events.

Building on the rich resource of Galaxy training materials (<https://training.galaxyproject.org/>), we have developed scalable training and outreach content for AnVIL (<https://anvilproject.org/learn>). Galaxy's extensive training network offers dozens of tutorials for the life sciences, including for genome assembly and variant analysis, functional genomics, proteomics, and epigenetics as well as other data intensive sciences such as climate research or machine learning. The AnVIL materials guide users in onboarding and explore the necessary building blocks for independent use of the AnVIL platform. In this presentation we will showcase the available AnVIL training resources, how they fit in the larger Galaxy Training Network, and summarize our experiences about lessons learned from onboarding users for using cloud resources.

IDENTIFICATION OF CONSERVED SEQUENCE ELEMENTS ACROSS 242 PRIMATE GENOMES WITH DEEP LEARNING

Sabrina Rashid^{*1}, Lukas F.K. Kuderna^{*1}, Jacob Ullirsch^{*1}, Mo Ameen¹, Laksshman Sundaram¹, Glenn Hickey², Anthony J Cox¹, Hong Gao¹, Arvind Kumar¹, Francois Aguet¹, Primate Conservation Sequencing Initiative³, Benedict Paten², Kerstin Lindblad-Toh^{4,5}, Jeffrey Rogers⁵, Tomas Marques Bonet^{7,8}, Kyle Kai-How Farh¹

¹Illumina, Inc., Artificial Intelligence Laboratory, Foster City, CA,

²University of California, UC Santa Cruz Genomics Institute, Santa Cruz, CA, ³Primate Conservation Sequencing Initiative, Consortium, -, Spain,

⁴Uppsala University, Science for Life Laboratory Department of Medical Biochemistry and Microbiology, Uppsala, Sweden, ⁵Broad Institute of MIT and Harvard, NA, Cambridge, MA, ⁶Baylor College of Medicine, Human Genome Sequencing Center and Department of Molecular and Human Genetics, Houston, TX, ⁷Institute of Evolutionary Biology, (UPF-CSIC), Barcelona, Spain, ⁸CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

Noncoding DNA is central to our understanding of human gene regulation and complex diseases, yet the rapid turnover of noncoding sequence compared to protein-coding sequence has meant that the genomic elements underlying recent adaptations in the primate lineage have remained elusive. Here, we assembled the genomes of 190 primate species de novo and combined them with existing assemblies to construct a multiple sequence alignment of 242 primate species, which represents more than half of all extant species in this lineage. We develop a deep-leaning approach to detect sequence conservation of DNA elements by explicitly modeling mutational processes, with improved sensitivity at the relatively short evolutionary timescales covered by primates. Using this resource, we identify 167,124 DNase I hypersensitivity sites and 560,579 transcription factor binding sites in the human genome that are conserved in primates but not across other placental mammals. We experimentally and computationally validate that these elements drive gene expression in primates, while the orthologous mouse sequences generally lack activity. Finally, we show that primate-specific conserved regulatory elements are enriched for human genetic variants underlying gene regulation and complex traits and diseases. Our results highlight the central role of recent adaptation in regulatory sequence elements underlying the emergence of humans and non-human primates from other placental mammals.

PRIORITIZING CORONARY ARTERY DISEASE RISK VARIANTS IN ATHEROSCLEROSIS USING DEEP LEARNING MODELS OF CHROMATIN ACCESSIBILITY IN MOUSE

Soumya Kundu^{*1}, Robert Wirka^{*2}, Daniel Li¹, Joao Monteiro¹, Disha Sharma¹, Laksshman Sundaram¹, Thomas Quertermous¹, Anshul Kundaje¹

¹Stanford, School of Medicine, Stanford, CA, ²UNC, School of Medicine, Chapel Hill, NC

Arterial smooth muscle cells (SMCs) respond to atherosclerotic lesions in coronary artery disease (CAD) by de-differentiating, proliferating, and migrating in a process known as phenotypic modulation. Recent work has revealed that these SMCs transform into fibroblast-like cells, known as fibromyocytes, which play a protective role by forming fibrous caps that cover these lesions and prevent adverse events, such as arterial thrombosis and myocardial infarction. Furthermore, genes at loci implicated in CAD risk through genome-wide association studies (GWAS), such as TCF21, ZEB2, and SMAD3, play a crucial role in the formation of these protective SMC-derived fibromyocytes. Although the transcriptomic changes underlying this phenotypic modulation of SMCs have recently been studied, the gene regulatory programs driving these transcriptomic changes remain poorly understood. Similarly, the causal variants modulating disease risk at hundreds of CAD GWAS loci remain unidentified.

In order to characterize the changes in the regulatory landscape during phenotypic modulation, we profiled the gene expression and chromatin accessibility of lineage-traced SMCs in aortic atherosclerotic tissue from mice at single cell resolution. We found a trajectory of cell states from quiescent SMCs to fibromyocytes and chondromyocytes, which are similar to endochondral bone-forming cells and are found at the end of the phenotypic modulation trajectory. To identify the transcription factors (TFs) regulating the genes expressed along this trajectory, we trained convolutional neural network models that can map regulatory DNA sequence to base resolution chromatin accessibility profiles in each cell state. Using model interpretation tools, we annotated the putative binding sites for each of these TFs along the trajectory and linked them to their target genes.

Finally, we used these cell-state specific models trained on mouse data to score the functional impact of all non-coding human variants associated with CAD risk. Strikingly, we found more than 500 variants with significant effects on predicted accessibility, spanning the entire phenotypic modulation trajectory. These variants disrupt motifs of key transcription factors, such as SRF, AP-1, MEF2, and ZEB, and have concordant scores from models trained on human cells with matching phenotype, highlighting the power of this approach to identify functional variants across species. Experimental validation of these prioritized variant effects in human coronary artery smooth muscle cells is underway.

LEVERAGING DUPLEX DNA SEQUENCING TO COMPREHENSIVELY INVESTIGATE GERMLINE MUTATIONS IN LONGITUDINALLY SAMPLED BULK SPERM.

Jason Kunisaki¹, Suchita Lulla¹, Michael Goldberg¹, Kenneth Aston², Jim Hotaling², Aaron Quinlan¹

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

Germline de novo mutations (DNMs) drive human evolution and underlie human disease. Therefore, their rates, patterns, and underlying sources are essential to understand. Work from our group and others have genome-sequenced blood from parent-child trios to uncover features of male germline mutagenesis, including estimating the mutation accumulation rate at 1.5 DNMs/year. However, such pedigree approaches study mutations in a single sperm capable of facilitating normal child development: thus, the unbiased analysis of germline mutation in bulk sperm is less explored. To overcome this survivorship bias, we examine bulk sperm, which include both reproductively “fit” and “unfit” gametes, wherein DNMs accumulate with age over the lifecycle of spermatogonial stem cells.

To investigate DNMs in bulk sperm, we apply innovative sequencing and computational strategies to a longitudinal cohort of sperm samples collected an average of ~16 years apart from nine fertile men aged between 31 to 68 years. This cohort is ideal for studying intra-individual germline mutation dynamics over a wide range of reproductive ages. To this end, we used deep (~10,000X) targeted duplex sequencing to distinguish true mutations as those present on complementary DNA strands from errors that arise on single strands, resulting in an error rate of 1 in 100 million. Duplex sequencing is necessary because the error rate of Illumina sequencing is orders of magnitude higher than the germline mutation rate. We also developed novel error modeling strategies for duplex sequencing data to improve our accuracy in detecting low-frequency DNMs in both control genomic regions and in 93 genes involved in spermatogenesis and DNA repair.

Our analysis of longitudinal sperm samples shows that clonal mutations (i.e., those found in multiple sperm derived from a shared common ancestor) accumulate with age at a rate of ~0.8 mutations/year. This is especially the case in paternal age effect genes, which drive clonal spermatogenesis and are involved in autosomal dominant diseases. Pathogenic mutations accumulate at a rate of 0.24 mutations/year - of these, up to 15% are clonal, representing a lifelong risk for recurrent deleterious mutation transmission. We will also present ongoing refinements in high-confidence DNM detection to test whether overall mutation accumulation rates measured from clonal and non-clonal mutations found in bulk sperm are greater than pedigree-derived estimates of 1.5 DNMs/year.

Our study aims to characterize temporal features of germline DNMs, including overall and clonal pathogenic mutation accumulation, directly from longitudinally sampled bulk sperm.

A PROGRAMMED MENDELIAN VIOLATION MAINTAINS HETEROZYGOSITY IN A PARTHENOGENETIC ANT

Kip D Lacy¹, Daniel J Kronauer^{1,2}

¹The Rockefeller University, Laboratory of Social Evolution and Behavior, New York, NY, ²Howard Hughes Medical Institute, New York, NY

Parthenogenesis arises sporadically across the animal phylogeny. The underlying alterations to meiosis are rarely understood, but might provide insights into the evolution of asexuality, and the cytological mechanisms of meiosis. In the clonal raider ant *Ooceraea biroi*, diploid zygotes form via the fusion of two haploid meiotic products. This process should lead to rapid and detrimental loss of heterozygosity under a standard meiosis featuring crossover recombination and random segregation. Using a combination of cytology and whole genome sequencing, we show that crossover recombination is in fact common, but that loss of heterozygosity is avoided due to nonrandom co-inheritance of reciprocally recombined chromatids. This programmed violation of Mendel's second law implies a possibly common cellular "memory" of crossovers that results from the biased orientation of recombined chromatids.

AUTOMATED REFERENCE GENOME ASSEMBLY BY GALAXY AND THE VERTEBRATE GENOME PROJECT

Delphine Lariviere¹, Giulio Formenti², Alex Ostrovsky³, Cristobal Gallardo⁴, Linelle Abueg², Nadolina Brajuka², Marc Palmada-Flores⁵, Anton Nekrutenko¹, Bjorn Grüning⁴, Michael Schatz³

¹Penn State University, Eberly college of Science, University Park, PA,

²Rockefeller University, Vertebrate Genome lab, New York, NY, ³Johns Hopkins University, Department of Computer Science, Baltimore, MD,

⁴University of Freiburg, Inst. of Computer Science, Freiburg, Germany,

⁵Universitat Pompeu Fabra-CSIC, Department of Medicine and Life Sciences, Barcelona, Spain

Recent improvements in genome sequencing and assembly promise to generate high-quality reference genomes for many species. Yet the genome assembly process is still laborious, costly, requires substantial expertise, and is generally not scalable to the goals of multispecies scientific efforts. To democratize the training and assembly process, we implemented the latest version of the Vertebrate Genomes Project assembly pipeline in Galaxy (<https://galaxyproject.org/projects/vgp/>). Galaxy is a framework that offers full workflow functionality and can support analyses of thousands of samples. Public Galaxy instances offer powerful computational resources for free, giving access to state-of-the-art analyses to everyone.

The automated pipeline performs de novo assembly based on PacBio HiFi reads, with optional extended graph phasing using HiC or parental data and scaffolding using Bionano optical mapping data and HiC via modular workflows. The workflows also include quality control throughout the assembly process using GenomeScope, gfastats, Merqury, BUSCO, and Pretext. We will demonstrate how to use these workflows via the Galaxy interface or via command line for generating dozens of assemblies rapidly using the free public resources available.

Within the Vertebrate Genome Project, these workflows have already been applied to de novo assemble the genomes of dozens of species, with the goal of assembling a genome for a representative species from all 260 vertebrate orders in the next year. We will discuss the quality of generated assemblies and how it is impacted by the technology used. The long-term goal is to use these workflows to generate high-quality, complete reference genomes for all of the roughly 70,000 extant vertebrate species and to help to enable a new era of discovery across the life sciences.

This partnership with the Vertebrate Genomes Project has led to several enhancements to Galaxy's ability to be utilized on large-scale projects. In addition to the new Galaxy assembly workflows, we will highlight Galaxy's (1) new capabilities in organizing, retrieving, analyzing, and uploading data to private repositories or AWS buckets and (2) current efforts on the globalization and integration of public computational infrastructure from the EU and US.

MASSIVELY PARALLEL PROFILING OF ANDROGEN RECEPTOR PROTEIN-CODING VARIANTS WITH SCAnnEd

Ceejay Lee¹, Tristan Tay², Hui Si Kwok¹, Simon P Shen¹, Calvin Hu¹, Jason D Buenrostro², Brian B Liau¹

¹Harvard University, Department of Chemistry and Chemical Biology, Cambridge, MA, ²Harvard University, Department of Stem Cell and Regenerative Biology, Cambridge, MA

Despite our ability to rapidly map genomic variants in the era of high-throughput sequencing, functionally characterizing their impacts remains challenging. This limitation not only blunts translational potential of genomic data but also hampers fundamental understanding of genes essential for human health. My research aims to bridge this knowledge gap by integrating CRISPR mutational scanning with single-cell genomics, a platform we term **single-cell annotation of genome editing** (SCAnnEd), to study Androgen Receptor (AR), a key oncogenic transcription factor. Through single-cell gene expression profiling, we interrogate the direct effects of AR variants, installed endogenously through base editing, on downstream transcription when stimulated by distinct classes of ligands. By combining mutagenesis, cellular profiling, and variant genotyping within a single workflow, SCAnnEd demonstrates a new screening paradigm in profiling fine-grained, complex consequences of protein-coding mutations.

SYSTEMATIC INVESTIGATION OF ALLELIC REGULATORY ACTIVITY OF SCHIZOPHRENIA-ASSOCIATED COMMON VARIANTS

Jessica C McAfee^{*1,2,3}, Sool Lee^{*1,2,4}, Jiseok Lee Lee^{1,2}, Jessica L Bell^{1,2}, Oleh Krupa^{1,2}, Jessica Davis^{5,6,7,8}, Kimberly Insigne^{5,6,7,8}, Marielle L Bond^{1,3}, Nanxiang Zhao⁹, Hyejung Won^{1,2}

¹University of North Carolina, Department of Genetics, Chapel Hill, NC,

²University of North Carolina, Neuroscience Center, Chapel Hill, NC,

³University of North Carolina, Curriculum in Genetics and Molecular Biology, Chapel Hill, NC, ⁴University of North Carolina, Curriculum in Bioinformatics and Computational Biology, Chapel Hill, NC, ⁵UCLA, Department of

Chemistry and Biochemistry, Los Angeles, CA, ⁶UCLA, Institute for Genomics and Proteomics, Los Angeles, CA, ⁷UCLA, Molecular Biology Institute, Los Angeles, CA, ⁸UCLA, Quantitative and Computational Biology Institute, Los Angeles, CA, ⁹University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, ¹⁰University of Michigan, Department of Human Genetics, Ann Arbor, MI

Genome-wide association studies (GWAS) have successfully identified 145 genomic regions that contribute to schizophrenia risk, but linkage disequilibrium (LD) makes it challenging to discern causal variants.

Computational finemapping prioritized thousands of credible variants, ~98% of which lie within poorly characterized non-coding regions. To functionally validate their regulatory effects, we performed a massively parallel reporter assay (MPRA) on 5,173 finemapped schizophrenia GWAS variants in primary human neural progenitors (HNPs). We identified 439 variants with allelic regulatory effects (MPRA-positive variants), with 71% of GWAS loci containing at least one MPRA-positive variant. Transcription factor binding had modest predictive power for predicting the allelic activity of MPRA-positive variants, while GWAS association, finemap posterior probability, enhancer overlap, and evolutionary conservation failed to predict MPRA-positive variants. Furthermore, 64% of MPRA-positive variants did not exhibit eQTL signature, suggesting that MPRA could identify yet unexplored variants with regulatory potentials. MPRA-positive variants differed from eQTLs, as they were more frequently located in distal neuronal enhancers. Therefore, we leveraged neuronal 3D chromatin architecture to identify 272 genes that physically interact with MPRA-positive variants. These genes annotated by chromatin interactome displayed higher mutational constraints and regulatory complexity than genes annotated by eQTLs, recapitulating a recent finding that eQTL- and GWAS-detected variants map to genes with different properties. Finally, we propose a model in which allelic activity of multiple variants within a GWAS locus can be aggregated to predict gene expression by taking chromatin contact frequency and accessibility into account. In conclusion, we demonstrate that MPRA can effectively identify functional regulatory variants and delineate previously unknown regulatory principles of schizophrenia.

THE GENOMIC LANDSCAPE ACROSS 624 SURGICALLY ACCESSIBLE EPILEPTOGENIC HUMAN BRAIN LESIONS

Costin Leu¹, Christian M Boßelmann¹, Jean Khoury², Lucas Hoffmann³, Sara Baldassari⁴, Robyn M Busch², Stéphanie Baulac⁴, Peter Nürnberg⁵, Imad Najm², Ingmar Blümcke³, Dennis Lal¹

¹Cleveland Clinic, Genomic Medicine Institute, Cleveland, OH, ²Cleveland Clinic, Neurological Institute, Cleveland, OH, ³University Hospital Erlangen, Department of Neuropathology, Erlangen, Germany, ⁴Sorbonne University, Paris Brain Institute, Paris, France, ⁵University of Cologne, Cologne Center for Genomics, Cologne, Germany

Rationale: Understanding the exact molecular mechanisms involved in the etiology of epileptogenic brain lesions is essential for improving the treatment of drug-resistant focal epilepsy.

Methods: We tested for somatic variant enrichment in lesional brain tissues from 624 individuals with epilepsy that received resective surgery (N=369 with malformations of cortical development, MCD, and N=255 with low-grade epilepsy-associated tumors, LEAT). To screen for low-fraction somatic variants, we used >350x whole-exome sequencing for 431 samples and >1000x targeted sequencing of 120 candidate genes identified in our previous study for 193 samples. To test for somatic variant enrichment, we used the dNdScv model from Martincorena et al. (2017), which tests for the ratio of non-synonymous to synonymous somatic mutations while accounting for variations in the background local mutation rate along the human genome.

Results: Our somatic variant enrichment analysis confirmed eight genes previously reported to cause MCD or LEAT. Specifically, we found 4/8 previously reported genes enriched with somatic variants in individuals with two LEAT subtypes (GG, ganglioglioma and DNET, dysembryoplastic neuroepithelial tumors): *BRAF* in 14.6% of individuals with GG ($q<10^{-16}$), *PTPN11* in 2.2% individuals with GG ($q=3.07\times10^{-4}$), *NFI* in 2.2% individuals with GG ($q=6.61\times10^{-3}$), and *FGFR1* in 18.4% individuals with DNET ($q<10^{-16}$). Three genes were enriched in individuals with MCD and subtypes (FCD2, Focal cortical dysplasia type 2 and MOGHE, mild FCD with oligodendroglial hyperplasia in epilepsy): *SLC35A2* in 22.4% of individuals with MOGHE ($q<10^{-16}$), *MTOR* in 7.4% individuals with FCD2 ($q=9.48\times10^{-14}$), and *RHEB* in 0.8% individuals with MCD ($q=2.53\times10^{-3}$). *PIK3CA* was significantly enriched in all 624 individuals with MCD or LEAT (1.1% carriers, $q=4.43\times10^{-3}$). Finally, we identified somatic variant enrichment in two novel genes associated with LEAT and eight genes with MCD (specifically MOGHE). Based on the number of significantly enriched genes and sample size, our study suggests that MOGHE (15% of the study cohort, five significantly enriched genes) has a higher genetic homogeneity than FCD2 (30% of samples, one significant hit).

Conclusions: The dNdScv-based somatic variant enrichment analysis is a viable model to identify (novel) causes for lesional epilepsy forms without somatic variant data from healthy controls. Such genotype-phenotype analyses will emerge with increasing numbers of genetically delineated lesion subtypes and inform clinical diagnostic screening and care.

IDENTIFICATION OF CELL TYPES AND CELLULAR DYNAMICS GENETICALLY ASSOCIATED WITH BRAIN DISORDERS AND COGNITIVE TRAITS

Ang Li¹, Irina Voineagu², Ryan Lister³, Naomi R Wray^{1,4}, Jian Zeng¹

¹University of Queensland, Institute for Molecular Biosciences, Brisbane, Australia,

²The University of New South Wales, School of Biotechnology and Biomolecular Sciences, Sydney, Australia, ³The University of Western

Australia, The Harry Perkins Institute of Medical Research, Perth, Australia,

⁴University of Queensland, Queensland Brain Institute, Brisbane, Australia

Human brain development plays a fundamental role in the development of brain disorders and cognitive ability, yet the process remains poorly understood. Recent studies have proposed that a more comprehensive understanding of brain development and its relationship to complex traits can be achieved by integrating signals from genome-wide association studies (GWAS) with single-cell RNA-seq. However, the causal cell type in which genetic variants and genes affect trait variation is not known, and there is no gold standard to compare the performance of different methods.

To address this, we first established a set of putatively causal cell type and trait/disease pairs as the ground truth based on general knowledge and empirical evidence from prior studies. We used scRNA-seq data from different cell types across 23 tissues from a murine scRNA-seq dataset to assess the performance of different methods. We found that LD score regression with cell-type specific gene set identified by top 10% expression proportion (EP-LDSC) and single-cell Disease Relevance Score (scDRS) performed better than the MAGMA gene-set approach in maximizing power and minimizing false positive rate.

We then applied EP-LDSC and scDRS to 11 brain disorders and traits, using scRNA-seq data from prefrontal cortex tissue across 6 developmental stages. By combining samples across stages of development, we identified cell types significant in both methods for brain disorders/trait. Our results are consistent with mounting evidence for the association between microglia and Alzheimer's disease (AD). Among the cell type/disease pairs we analyzed, 52.17% (12/23) exhibited significant cellular heterogeneity signals, all of which can be explained by the differences in developmental stages between cells.

Furthermore, we used the 4 “trends” of gene expression throughout development to capture cellular dynamics and tested their associations with brain disorders and traits in each cell type. Our results suggest that the genetic risk of AD is associated with the transiently up trend cellular dynamics in microglia.

In summary, our study enhances our understanding of brain disease aetiology in relation to the dynamics of brain development and highlights the potential of integrating GWAS and scRNA-seq data to better understand the underlying biology of complex traits.

CHANGES IN ASTROCYTES TRANSCRIPTOME AND PROTEOME WITH AGING IN NORMAL AND ALZHEIMER'S DISEASE MICE BRAIN

Jiangtao Li¹, Michelle Olsen²

¹Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, ²School of Neuroscience, Virginia Tech, Blacksburg, VA

Alzheimer's disease (AD) is the most common age-related neurodegenerative disease, with 58 million individuals affected worldwide. AD is characterized by abnormal aggregation of β -amyloid (A β) peptides and neurofibrillary tangles (NFTs) derived from hyperphosphorylated tau (p-tau). The vast majority (>95%) of affected individuals do not have an AD associated gene mutation and it is thought multiple factors may drive this disease. Changes in astrocyte function have been observed in brains from individuals with AD, and using both in vitro and in vivo animal models. Astrocytes play an important role in neuronal and brain health. The influence of astrocytes on the accumulation and clearance of A β and tau in the brains of individuals with AD is poorly understood, although there is growing awareness that these are therapeutically important interactions. Here, we use hAPPJ20 (a commonly used AD animal model, with vascular A β accumulation) animal to evaluate the changes of astrocytes between WT and AD animal during disease progression and aging process in male and female animals. RNA was extracted from isolated astrocytes from 3, 6, 12 and 18 months WT and AD mice cortex using a magnetic isolation approach. RNA sequencing was used to evaluate the transcriptome changes between WT and AD animal. Our data indicate robust changes in gene expression starting as early as 3 months of age, with the largest numbers of DEG's seen between 12 – 18 months. We identified four genes associated with reactive astrocytes (GFAP, Serpinf1, Serpina3n and Hspb6) are changed at all time points examined in females. Gene Ontology (GO) analysis shows that immune responses and apoptotic pathways are the most impacted pathways between WT and AD group, suggesting astrocytes may be dying during disease progression. Gliogenesis, microtubule-based movement and cell communication are also impacted in female AD mice across disease progression. Future studies include a multi-omics comparison of transcriptomic data, and immunohistochemical approaches to evaluate astrocytes in the diseased brain to gain a better understanding of how astrocytes are impacted and contribute to AD disease progression.

Support

This work was supported by NIH R01AG065836

Key work

Alzheimer's disease, astrocytes, transcriptome

ASSOCIATING CANCER AND STROMAL GENOMES WITH TRANSCRIPTOMES BY HIGH-THROUGHPUT SINGLE-CELL SEQUENCING

Siran Li¹, Joan Alexander¹, Jude Kendall¹, Gary Goldberg², Dan Levy¹, Michael Wigler¹

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ²Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, New Hyde Park, NY

Single-cell genomic analyses can provide information on tumor and normal genomes, whereas single-cell transcriptomic analyses can distinguish cell types and states of tumor and stroma. To combine these modalities, we developed a high-throughput capture of both DNA and RNA from single nuclei, and developed algorithms for the separation and clustering of the data. We applied this hybrid protocol to single nuclei extracted from frozen biopsies of five different endometrial cancer patients and clustered the genome and expression data. We also analyzed the same samples using RNA-only and DNA-only protocols, respectively, to verify the clustering. In addition, we developed the “multinomial wheel,” an algorithm that measures the deviation of each single cell from the major clusters.

Working with nuclei from five uterine cancer patients, we observe that tumor expression clusters are highly distinct between patients, whereas the stroma has many identifiable expression clusters that are mostly shared between patients. All five patients contain evidently mutant stroma. We found a significant amount of stromal mutations existed in patients with the worst clinical outcomes, indicating that mutant stroma might be a poor prognostic indicator. Within any given patient, we observed different tumor clones could project into distinct or shared expression states, with nearly all possible genome-transcriptome correlations observed in the cohort, suggesting epigenetic regulation and perhaps active selection for clonal diversity. These observations demonstrate the potential of multi-omic technologies for investigating cancer evolution, host reaction, and stromal mutations.

EMPLOYING LINKED READ SEQUENCING (HAPLOTAGGING) TO PROFILE DE NOVO STRUCTURAL VARIATION IN SPERM THROUGH AGING

Stacy Li^{1,2}, Joana Rocha², Peter H Sudmant^{1,2}

¹UC Berkeley, Computational Biology Graduate Group, Berkeley, CA, ²UC Berkeley, Department of Integrative Biology, Berkeley, CA

Aging is an emergent phenomenon shared across the breadth of life, hallmark by the deterioration of physiological processes over time. DNA repair becomes less efficient with age, shifting both the type and frequency of observed mutations, including that of *de novo* structural variants (dnSVs). This effect is particularly pronounced in the male germline where gametogenesis occurs continuously, yielding a distinct mutational signature of aging. Germline dnSVs arise in relation to parental haplotypes and pose a significant risk to reproductive success. Although haplotype-specific SVs have been well-studied in somatic contexts, their presence and impact within the male germline have yet to be fully explored.

In this study, we assess how haplotype and paternal age influence the emergence of dnSVs during gamete formation. We adapt haplotagging, a linked read sequencing method, to sequence sperm samples from several (100 planned) donors of varying ages, where each sample represents over 100 million haploid genomes. Haplotagging leverages the physical properties of DNA in solution to assign unique molecular identifiers (UMI) to individual DNA molecules. Libraries derived from the same strand of DNA share the same UMI, simultaneously encoding molecular origin alongside sequence data (“molecule-linked reads”).

Molecular linkage enables the reconstruction of physically phased haplotypes: we infer donor haplotypes and identify recombination events by noting “switches” between haplotypes. Linkage-aware mapping yields long-range genomic information and improves power to detect complex variants: we detect dnSVs by searching for significant gaps in linked molecular coverage that segregate alongside continuous sequence data. Altogether, our work presents insight into how haplotype context influences dnSV development, and provides a basis for further research into germline stability and genome evolution over the reproductive lifetime.

MULTI-OMIC BAYESIAN HIERARCHICAL MODELING REVEALS TRAIT-RELEVANT RARE GENETIC VARIATION

Taibo Li¹, Rebecca Keener¹, Rachel Ungar², Nicole Ferraro², Matilde Cimnigliaro³, Stephanie Arteaga³, Bohan Ni¹, Jerome Rotter⁴, Stephen Rich⁵, Dan Arking¹, Daniel Geschwind³, Stephen Montgomery², Alexis Battle¹

¹Johns Hopkins University, Baltimore, MD, ²Stanford University, Stanford, CA,

³Univ of California, Los Angeles, Los Angeles, CA, ⁴Lindquist Institute, Torrance, CA, ⁵Univ of Virginia, Charlottesville, VA

Rare genetic variation is abundant in individual human genomes. Our understanding of rare regulatory variants is limited by a lack of accurate methods to infer their functional impact. We previously developed probabilistic methods to integrate functional data into personal whole genome sequencing (WGS) analysis; however, existing studies are limited to mRNA sampled from blood in individuals of European ancestry. Here, we leveraged multi-omic data from diverse populations to comprehensively annotate the functional impact of rare variation and dissect their regulatory mechanisms underlying complex traits.

First, we extended our Bayesian hierarchical model, Watershed, to integrate multi-omic measurements with genomic annotations of personal WGS. From 1,319 participants in the Multi-Ethnic Study of Atherosclerosis (MESA), we identified those who had extreme (“outlier”) levels of mRNA expression, methylation, splicing, and protein expression, reflecting multiple stages of gene regulation. We prioritized ~200,000 variants out of 30 million with evidence of driving outlier status of nearby genes, thus creating a catalog of potential rare regulatory variants for reference when analyzing new WGS datasets. Notably, our method performed well in diverse genetic ancestries.

Based on Watershed and our prioritized variants, we investigated context-specific and trait-relevant rare variants across multiple datasets. First, using data across 44 tissues in GTEx, we found that sex-specific rare variants are enriched in regulatory regions and discovered sex-specific transcription factor networks anchored by noncoding rare variants. Next, leveraging WGS and transcriptomic data from individuals with autism spectrum disorders (ASD) and controls ($N = 119$), we implicated brain-specific splicing variants from GTEx brain tissues in ASD diagnosis and dissected regional variability of their effects. Lastly, we applied rare variant Watershed posteriors to nominate several known and novel genes associated with polygenic traits including telomere length (TL) and electrocardiogram intervals. From 109,110 TOPMed individuals, we estimated 69% of those with short TL have at least one rare variant prioritized by Watershed from a reference panel of 16 causal genes. We showed that synthesizing rare variant scores by Watershed can improve risk stratification for PR interval over common variant models.

Overall, we present a powerful and flexible framework to prioritize functional rare variants and reveal widespread context-specific effects in the noncoding genome. As whole genome projects identify hundreds of millions of rare variants, we demonstrate an integrated personal-omics approach to identify the most impactful rare variants influencing a range of complex traits.

A NOVEL PATHWAY ANALYSIS METHOD FOR scRNA-SEQ AND SPATIAL TRANSCRIPTOMICS DATA

Qingnan Liang, Ken Chen

The University of Texas, MD Anderson Cancer Center, Department of Bioinformatics and Computational Biology, Houston, TX

Advances in single-cell technology have enabled molecular cellular dissection of heterogeneous biospecimens at unprecedented scales and resolutions. Pathway analysis plays a critical role by connecting newly collected data to curated functional-relevant gene sets. However, pathway analysis could also be challenging for single-cell data due to the high dimensionality, sparsity, and low noise-to-signal ratio of single-cell data. Moreover, cluster-centric approaches have limited power in dissecting and interpreting highly heterogenous, dynamically evolving data. Additionally, computational approaches for identifying spatial-relevant pathways for spatial transcriptomics (ST) data remain limited. Here, we report a novel computational tool, GSDensity, that has the following features: 1) Robust to sparse and noisy single-cell data; 2) Can quantify overall pathway heterogeneity and single-cell pathway activity simultaneously, 3) Independence to clustering or annotation of cells, and 4) Can detect pathways with spatial relevance. We show that GSDensity can not only accurately detect biologically distinct gene sets but also reveal novel cell-pathway associations that are ignored by existing methods. This is particularly evident for samples with continuous state transitions, as we demonstrated in triple-negative breast cancer data and mouse brain development data. We also show that GSDensity can identify spatially relevant pathways in mouse brains including those following a high-order organizational pattern in the ST data. Finally, we performed pan-cancer pathway analysis and identified spatially relevant and recurrently active pathways across six different tumor types. We conclude that GSDensity is a highly sensitive and robust computational tool for pathway-centric single-cell data analysis with many application potentials.

PHENOME-WIDE PGS PORTABILITY IN THE COLORADO CENTER FOR PERSONALIZED MEDICINE BIOBANK SUGGESTS OVERLOOKED CHALLENGES IN DIVERSE POPULATIONS

Meng Lin^{1,2}, Christopher H Arehart¹, Nicholas Rafaels¹, Kristy R Crooks¹, Nikita Pozdeyev^{1,2}, Audrey Hendricks^{1,3}, Sridharan Raghavan^{2,4}, Christopher R Gignoux^{1,2}

¹University of Colorado Anschutz Medical Campus, Colorado Center for Personalized Medicine, Aurora, CO, ²University of Colorado Anschutz Medical Campus, Department of Biomedical Informatics, Aurora, CO,

³University of Colorado Denver, Department of Mathematical and Statistical Sciences, Denver, CO, ⁴Rocky Mountain Regional VA Medical Center, Section of Hospital Medicine, Aurora, CO

Increasingly powered association results are available as an abundant resource for generating polygenic scores (PGSs). We comprehensively examined the landscape of current PGS transportability across a phenome-wide range of conditions by empirically leveraging the PGS Catalog to assess the performance of their predictions for >1k phenotypes based on electronic health records (EHR) in the Colorado Center for Personalized Medicine Biobank ($N=33,863$, or $N_{\text{predictions}}=661,265$). We found that some diseases were predicted reasonably well by PGSs together with demographic covariates, such as type 2 diabetes (T2D) and hypertension ($P=2.7e-170$ and $1.1e-164$, AUC=0.77 and 0.81), with the former showing a homogeneous prediction effect across all ancestry groups. For both of these phenotypes, individuals scoring in the top 3-4% in the biobank had OR>3. However, the majority of predictions had considerable cross-group heterogeneity in performance (average $I^2=0.18$ in phecode~PGS pairs with FDR<0.1). This can greatly affect the potential use of PGS in personalized medicine and bias downstream interpretation in such frameworks as Mendelian Randomization, where we observed flipped directions between instrument and outcome in different ancestry groups. Additionally, we found that the choice of PGS unit of measure in associations, whether taken per SD or stratified at the top decile against the remainder, yielded highly discordant estimates of heterogeneity ($r=0.29$). Using multilevel nested mixed models, we found influences on heterogeneity and model classification included multiple features from both the training and test sets, such as the distribution of disease prevalence between ancestry groups in CCPM, and functional scores of variants. Our results suggest that there are important consequences of applying PGS to downstream applications that assume causality that only stem from homogeneous predictions across ancestry groups. The overlooked attributions are due to both characteristics of the training and test settings and may be determined empirically. This provides a description of hopes and pitfalls in ongoing efforts to apply PGS resources in diverse populations.

SCGRAPH2VEC: A NEW METHOD FOR GENE EMBEDDING AUGMENTED BY GRAPH NEURAL NETWORK AND SINGLE-CELL OMICS DATA

Shiqi Lin^{1,2}, Peilin Jia¹

¹Beijing Institute of Genomics, Chinese Academy of Sciences, Key Laboratory of Genomic and Precision Medicine, Beijing, China, ²University of Chinese Academy of Sciences, College of Life Sciences, Beijing, China

Exploring the cellular processes of genes from the aspects of biological networks is of great interest to understanding the properties of complex diseases and biological systems. Biological networks, such as gene regulatory networks and protein-protein interaction networks, can provide insights into the molecular basis of cellular processes and functions and often form different modules in different tissue and disease environments. By taking advantage of the rapidly accumulated single-cell omics data and the advances of deep learning, we developed a computational method called scGraph2Vec to integrate single-cell gene matrices and gene interaction networks. scGraph2Vec builds on the variational graph autoencoder (VGAE) extensible framework and generates gene embeddings that are highly informative by incorporating single-cell data patterns of neighbor genes and communities. We applied scGraph2Vec to nine representative human tissues and generated tissue-specific gene embeddings ready for multiple downstream applications. We demonstrated that these gene embedding data could be used to reveal functional gene modules that either represented general cellular processes or tissue-specific processes, such as gene subnetworks regulating neuron differentiation in the brain. We also applied these gene embeddings to infer disease-associated genes using GWAS data for COVID-19 and Alzheimer's disease. We showed that these gene embedding data can identify more disease genes and reveal the underlying mechanisms of the diseases. In summary, scGraph2Vec not only reconstructs tissue-specific gene networks with single-cell resolution but also obtains high-dimensional biological information for genes.

DYNAMIC COMPLEXITY OF GENETIC REGULATORY EFFECTS IN RESPONSE TO A HIGH CHOLESTEROL, HIGH FAT DIET IN BABOONS

Wenhe Lin¹, Ge Li³, John VandeBerg⁴, Deborah Newman⁵, Michael Olivier³, Mark Abney¹, Jeff Wall⁶, Laura A Cox³, Yoav Gilad²

¹The University of Chicago, Department of Human Genetics, Chicago, IL, ²The University of Chicago, Department of Medicine, Chicago, IL, ³Wake Forest University School of Medicine, Center for Precision Medicine, Winston-Salem, NC, ⁴University of Texas Rio Grande Valley, Edinburg, TX, ⁵Texas Biomedical Research Institute, San Antonio, TX, ⁶HIBio, San Francisco, CA

Environmental factors are known to play a pivotal role in disease development and prevention. Yet, due to experimental challenges in human studies, gene-by-environment (GxE) interactions have largely been ignored. Model organisms, including non-human primates, offer a way to study GxE interactions in a controlled environment. In particular, baboons, due to their striking genetic and physiological similarity to humans, have been used to model a variety of complex human diseases, including dyslipidemia, obesity, and atherosclerosis. To identify and characterize the molecular effects of diet in living tissues, we obtained liver, muscle, and adipose tissue biopsy samples from 99 captive baboons (56M, 43F) before and after they were fed a high cholesterol, high fat (HCHF) diet for two years. Using RNA-sequencing data from 589 high-quality tissue samples, we obtained transcriptional profiles from each tissue at the two time points. We then used these data to examine differences in gene expression and genetic regulation following the HCHF diet.

Across all three tissues, we discovered 6,378 diet-responsive (DR) genes that were differentially expressed following the HCHF diet ($FDR < 0.01$). We found that the DR genes, in adipose and liver specifically, are enriched in inflammatory responses and epithelial-mesenchymal transition. We computationally inferred cell type enrichment in each sample and found a variety of immune cell types to be significantly more enriched after the HCHF diet. We also observed adipose-specific sex differences in transcriptional responses to the HCHF diet. Next, we discovered 12,251 cis eQTL-gene pairs (eGenes) across all Time x Diet groups ($LFSR < 0.05$). By analyzing regulatory effects jointly, we identified 2,714 dynamic, diet-responsive eQTLs (DR eQTLs), which are significant in only one diet condition, or significant in both conditions but with different effects. Compared to steady-state eQTLs, we observed that DR eQTLs were more tissue-specific. We also found that the sexes can have divergent regulatory patterns, thereby masking diet-responsive effects. Finally, by integrating DR eQTLs with GWAS results from baboons and humans, we found that four DR eQTLs colocalized with GWAS loci for lipid traits in baboons ($PP > 0.5$), and that DR eGenes were enriched among relevant disease-associated genes in humans.

Our results demonstrate the dynamic complexity of genetic regulatory effects and showcase the translational potential of our baboon model system. The diet-responsive effects identified in this study can be used to understand gene-by-diet interactions, their associated mechanisms, and their role in modulating disease risk that cannot be characterized in humans.

AN INTEGRATIVE STUDY TO IDENTIFY THE LINK BETWEEN
DYSREGULATED INTERCELLULAR SIGNALINGS AND GENETIC
VARIANTS IN ALZHEIMER'S DISEASE

Andi Liu^{1,2}, Xiaoyang Li^{1,2}, Yulin Dai², Zhongming Zhao^{1,2,3}

¹The University of Texas Health Science Center at Houston, School of Public Health, Department of Epidemiology, Human Genetics and Environmental Sciences, Houston, TX, ²The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Center for Precision Health, Houston, TX, ³The University of Texas Health Science Center at Houston, School of Public Health, Human Genetics Center, Houston, TX

Alzheimer's disease (AD) is a complex and debilitating neurodegenerative disorder affecting over 40 million people worldwide. The onset and progression of AD are heavily influenced by genetic variants, which have been the focus of extensive research in recent years. Researchers have identified over 75 genetic variants linked to AD through genome-wide association studies (GWASs) and have used single-cell RNA sequencing (scRNAseq) data to identify alterations in gene expression and intercellular signals. However, whether and how AD-associated genetic variants manifest their impacts on intercellular signaling and functional pathways in disease progression remains poorly understood. In this study, we aim to identify dysregulated intercellular signals and their underlying genetic variants by integrating scRNAseq, GWAS, and individual whole genome sequencing data. Firstly, we updated and applied the CellChat tool on single-cell transcriptome profiles covering 48 prefrontal cortex (PFC) samples from the Religious Orders Study and Memory and Aging Project (ROSMAP). This allowed us to identify 229 unique dysregulated ligand-receptor (LR) pairs, primarily observed in astrocytes, microglia, and excitatory neurons. Subsequently, we identified 713 Gene Ontology (GO) functional pathways containing dysregulated LR pairs. We conducted parallel pathway-level analyses by leveraging population (GWAS statistics)- and individual (Whole genome sequencing)-level genetic information to prioritize AD-associated GO pathways. As a result, we highlighted seven key GO terms containing ten dysregulated LR pair genes (*APLP2*, *APOE*, *APP*, *C3*, *GPR37*, *GPR37LI*, *LRP1*, *PSAP*, *SERpine2*, *SORL1*), which are known to be involved in amyloid precursor protein regulation and process among other. To conclude, the enriched genetic risks linked to the cellular cross-talks among astrocytes, microglia, and excitatory neuron were identified, providing new insights into the potential therapeutic targets involved in dysregulated cell-cell communication in AD.

SINGLE-CELL RNA-SEQ LINKS CELL TYPE-SPECIFIC REGULATION OF SPLICING TO AUTOIMMUNE DISEASES

Chi Tian¹, Yuntian Zhang², Yihan Tong¹, Boxiang Liu^{1,2}

¹National University of Singapore, Department of Pharmacy, Singapore, Singapore, ²National University of Singapore, Department of Biomedical Informatics, Singapore, Singapore

Genetic regulation of pre-mRNA splicing (sQTLs) is a fundamental mechanism that affects complex traits and diseases. Existing large-scale sQTL studies conducted with tissue-level RNA-seq cannot resolve the cellular heterogeneity of genetic regulatory mechanisms. Leveraging the first phase of the Asian Immune Diversity Atlas, we provide a detailed dissection of the cell-type specific genetic regulation of pre-mRNA splicing by single-cell RNA sequencing of 1,058,909 peripheral blood mononuclear cells from 503 healthy individuals. We demonstrate robust splicing quantification and reproducible genetic effects on splicing using replicate samples. We identify thousands of independent cis-sQTLs and hundreds of trans-sQTLs, most of which have regulatory effects orthogonal to those on expression (eQTLs). Furthermore, we discovered a substantial number of cell-type specific sQTLs across 19 immune cell subtypes, as well as sex- and ancestry-biased sQTLs for genes known to be involved in autoimmune diseases. We next identified the dynamic usage of introns and changes in sQTL effects across the developmental trajectory from naive to memory B cells. Finally, we observed strong enrichment of sQTL effects in autoimmune GWAS loci and applied complementary colocalization and transcriptome-wide association approaches to pinpoint hundreds of cell-type specific putative causal genes for autoimmune GWAS. This work highlights the feasibility and importance of cell-type specific sQTL and their involvement in complex autoimmune diseases.

MAPPING AND FUNCTIONAL CHARACTERIZATION OF STRUCTURAL VARIATION IN 1,060 PIG GENOMES

Liu Yang¹, Lijing Bai¹, Hongwei Yin¹, Kui Li¹, George E Liu², Lingzhao Fang^{3,4}

¹Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China, ²Animal Genomics and Improvement Laboratory, USDA ARS, Beltsville, MD, ³MRC Human Genetics Unit at the Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, ⁴Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

Structural variations (SVs) have profound consequences on complex phenotypes through rearranging large regions of DNA sequence. Here, we report a comprehensive SV catalog based on the whole genome sequence of 1,060 pigs representing 101 breeds. It covers 9.6% of the pig genome nonredundantly, including 42,487 deletions, 37,913 mobile element insertions, 3,308 duplications, 1,664 inversions, and 45,184 break ends. Estimates of breed ancestry and hybridization between pig breeds using genotyped SVs generated similar results to those using single nucleotide polymorphisms. For example, we observed hundreds of deletions that have been stratified across geographic groups and breeds. We confirmed that the known duplications of the KIT gene were underlying the dominant white coat color in European pigs. We also explored the impacts of SVs on gene expression, functional elements, and complex traits of economic importance by systematically integrating them with expression quantitative trait loci (eQTLs) from 34 major tissues, chromatin states from 14 tissues, and sequence-based genome-wide association study (GWAS) signals of 14 complex traits from the Pig Genotype-Tissue Expression (PigGTEx) project, as part of FarmGTEx project. We further found a recent SINE element (Pre_SS01) insertion in the 3'-UTR of the MYO5A gene mainly in European pigs, distributed according to the geographic locations. We hypothesized that this insertion, overlapping an intron of MYO5A, could affect the alternative splicing pattern of this gene, facilitating coat color changes. A Yorkshire-specific copy number gain within ABCG2 could potentially downregulate its host gene and increase the expression of its long-distance downstream (~79.5 kb) gene SPP1 overlapping its nearby gene PKD2 (~7.0 kb) in multiple tissues by reorganizing chromatin interactions. In summary, this SV catalog is a valuable resource for studying diversity and evolutionary history in pigs and how domestication, trait-based breeding, and adaptive evolution have functionally shaped the pig genome.

[The co-authors also include Wenye Yao, Tan Tao, Qianyi Zhao, Yahui Gao, Jinyan Teng, Zhiting Xu, Qing Lin, Shuqi Diao, The FarmGTEx Consortium, Zhangyuan Pan, Dailu Guan, Bingjie Li, Huaijun Zhou, Zhongyin Zhou, Fuping Zhao, Qishan Wang, Yuchun Pan, and Zhe Zhang]

DNA-SEQUENCE AND EPIGENOMIC DETERMINANTS OF LOCAL RATES OF TRANSCRIPTION ELONGATION

Lingjie Liu^{1,2}, Yixin Zhao¹, Adam Siepel¹

¹Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, ²Stony Brook University, Graduate Program in Genetics, Stony Brook, NY

Transcription in eukaryotes that is carried out by RNA polymerases (RNAPs) can be divided into three phases: initiation, elongation, and termination. Although initiation was traditionally considered to be the most important regulatory step of transcription, precise mapping of RNAPs using advanced Nascent RNA sequencing (NRS) technologies like PRO-seq has indicated the dynamics of RNAPs during transcription elongation is much more complex than we initially thought. Accumulated evidence has suggested that the elongation rates along gene bodies not only determine the efficiency of the mRNA production but also influence the outcome of co-transcriptional activities such as splicing by providing windows of opportunity. Here we define a generalized linear model to describe the dynamics of RNAPs along gene bodies under steady-state conditions by assuming the site-specific elongation rates are an exponentiated linear function of a collection of features. By pooling information across numerous sites, we quantitatively determine how local elongation rates are modulated conjointly by epigenomics features, including histone marks, transcription factors, splicing sites and so on. Besides, we are able to identify a set of sequence determinants that have significant effects on elongation rates. Combining the primary determinants, our model provides a good prediction of the local elongation rates with a high resolution.

CORRECTING AND CLASSIFYING SARS-COV-2 RNA EXPRESSION IN SINGLE CELLS

Wendao Liu^{1,2}, Zhongming Zhao^{1,2}

¹The University of Texas MD Anderson Cancer Center UTHealth Houston, Graduate School of Biomedical Sciences, Houston, TX, ²The University of Texas Health Science Center at Houston, Center for Precision Health, School of Biomedical Informatics, Houston, TX

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection induces heterogeneous immune responses in various host cells. Increasing number of studies have applied single-cell RNA sequencing (scRNA-seq) to profile the expression of both host genes and SARS-CoV-2 genes in the tissue of COVID-19 patients. In this work, we systematically reviewed and examined SARS-CoV-2 infection in various cell types across multiple organs from COVID-19 patients and its potential phenotypic implications. However, the transcription biology of coronaviruses is distinct from most eukaryotes, relying on subgenomic RNAs (sgRNAs) to translate most proteins. This makes it difficult for current bioinformatic methods to accurately quantify the expression level of individual SARS-CoV-2 genes. To address this problem, we developed a tool SC2count to classify SARS-CoV-2 reads and correct SARS-CoV-2 gene expression from alignment files. SC2count uses pair-end 5' library scRNA-seq data, assigns viral reads to corresponding genes, and classifies viral RNAs into several categories. We also trained a multivariate multiple regression model to predict corrected counts from uncorrected counts. The pretrained weights could be used to correct SARS-CoV-2 gene expression in other single-end or 3' libraries. We applied SC2count to COVID-19 bronchoalveolar lavage fluid (BALF) scRNA-seq samples. We found that SARS-CoV-2 canonical sgRNAs with leader-gene body fusion comprised only a small proportion of all SARS-CoV-2 RNAs, while most RNAs contained no leaders. Our analysis revealed that some SARS-CoV-2 genes were selectively transcribed in specific cell populations. With corrected SARS-CoV-2 gene expression, we identified host genes whose expression were strongly correlated with SARS-CoV-2 gene expression and involved in multiple virus-induced pathways. In summary, we introduce a tool to correct biased quantification of SARS-CoV-2 gene expression, which can be used for identifying additional features of SARS-CoV-2 transcription in COVID-19 scRNA-seq data. It may facilitate downstream analyses in understanding cell type-specific viral gene expression and host immune responses.

THE CHROMATIN REGULATORY LANDSCAPE OF MOUSE LIVER REGENERATION

Palmira Llorens-Giralt^{*1}, Marina Ruiz-Romero^{*2}, Macarena Herranz-Itúrbide³, Ramil Nurtdinov², A Silvina Nacht², Guillermo P Vicent⁴, Florenci Serras¹, Isabel Fabregat³, Montserrat Corominas¹

¹Universitat de Barcelona, Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Biomedicina (IBUB), Barcelona, Spain, ²Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology (BIST), Barcelona, Spain, ³Bellvitge Biomedical Research Institute (IDIBELL), Departament de Ciències Fisiològiques, Universitat de Barcelona, Oncology Program, CIBEREHD, Instituto de Salud Carlos III, Barcelona, Spain, ⁴Molecular Biology Institute of Barcelona, Consejo Superior de Investigaciones Científicas, IBMB-CSIC, Barcelona, Spain

The molecular and cellular mechanisms underlying liver regeneration after moderate damage or resection are well described. However, little is known about how chromatin organization, specific enhancers and transcriptional networks are regulated to initiate liver regeneration while sustaining essential metabolic functions. To analyze the relationship between chromatin accessibility and transcriptional changes during mouse early liver regeneration, RNA-seq and ATAC-seq data were obtained at 6, 24 and 48 hours after 2/3 partial hepatectomy. We identified over 17,000 differentially accessible regions, including regions exclusively detected in the regenerating liver (regeneration-specific enhancers). Taking advantage of available Hi-C and ChIP-seq data, we predicted the target genes of these regions and found that chromatin accessibility is increased in regulatory elements of cell signaling and proliferation genes, while regions with lower accessibility correspond to regulatory elements of metabolic genes. We then analyzed footprints within regeneration-activated enhancers and found enrichment for transcription factors that are known to be required for liver regeneration, such as AP-1, and others with unknown or recently found roles, such as XBP1. Our findings suggest the existence of regeneration-specific regulatory elements that become sequentially activated to control the gene expression programs of liver regeneration.

UNRAVELING THE GENETIC BASIS OF RAPID DIVERSIFICATION IN ROCKFISH

Runyang Nicolas Lou¹, Laura Timm², Stacy Li¹, Katie D'Amelio², Kirby Karpan², Nathan Sykes³, Gregory Owens³, Wesley Larson², Peter Sudmant¹

¹UC Berkeley, Department of Integrative Biology, Berkeley, CA, ²National Oceanographic and Atmospheric Administration, Alaska Fisheries Science Center, Juneau, AK, ³University of Victoria, Department of Biology, Victoria, Canada

Rockfish (genus *Sebastodes*) is one of the most speciose group of vertebrates, with more than 100 species described worldwide. Most of this diversity is distributed in the northeast Pacific Ocean, where rapid speciation occurred repeatedly in separate lineages, often along a depth gradient. These speciation events are also accompanied by extreme life history changes, resulting in highly variable lifespan across rockfish species ranging from 11 years to over 200 years. In this project, we selected 17 rockfish species from 7 species pairs/trios and broadly sampled across their geographic distribution. Using whole genome sequencing of 947 individuals, we show different pairs and trios of species are in different stages of speciation with varying degrees of reproductive isolation. Within each species, there is a diversity of population structures, from range-wide panmixia to strong signatures of isolation-by-distance. Strikingly, we uncover several cases of cryptic diversity within species, where distinct populations exist in sympatry or parapatry and are potentially in the process of ongoing speciation. By implementing various population genetic analyses under a comparative genomics framework, we investigate the shared and unique genetic basis of rapid diversification in different groups of rockfish. We further construct a rockfish pangenome reference by assembling chromosome-level genomes for all species included in this study, which allows us to examine the roles of structural variation in speciation with an unprecedented resolution. Together, we demonstrate that rockfish harbor snapshots of speciation at its different stages, thus enabling us to study speciation in real time. Our findings lend unique insight into the mechanisms of ecological speciation and have important implications in lifespan evolution as well as in fisheries management.

CHARACTERIZATION OF HOUSEKEEPING REGULATORY ELEMENTS IN THE HUMAN GENOME

Martin Loza¹, Alexis Vandenbon², Kenta Nakai¹

¹The University of Tokyo, The Institute of Medical Science, Tokyo, Japan,

²Kyoto University, Institute for Life and Medical Sciences, Kyoto, Japan

Within the complex regulatory mechanism behind gene expression, the interaction of so-called enhancers and promoters has become a long-standing topic of research [1], and thanks to the advances in sequencing techniques, now it's possible to identify enhancers and their target genes with unprecedented resolution. However, due to the complexity and specificity of protocols, the experimental validation of enhancer-promoter interactions in many cell types is not easy. Therefore, we rely on multi-omics analysis of a large number of sequencing data to characterize these regulatory elements and predict their interactions [2-3]. In this study, we leverage epigenetics and nucleotide features of interactions predicted by the ABC method [3] to characterize enhancer-promoter interactions in 50 human cell types. We show that even though most of the elements are cell type-specific, around 10,000 "housekeeping regulatory elements" (HKREs) are active in at least 90% of the cell types and have distinctive features that differentiate them from cell type-specific ones. Around 18% of the HKREs are close to or inside the promoter region of most of the housekeeping genes; moreover, around 50% of the core promoters of housekeeping genes are HKREs. HKREs are rich in GC and CpG content and they seem to interact with a considerably large number of genes (a median of 22 target genes) over long distances (around 75 Kbp median distance). The number of HKREs correlates with the number of protein coding genes across chromosomes (Pearson correlation = 0.988) suggesting their important role in gene regulation. Most of the HKREs are close to or inside promoter regions, which implies their possible relation with ePromoters [4], a regulatory element with both promoter and enhancer capabilities. Overall, our work unveils a new key regulatory element active in multiple cell types with distinctive epigenetic characteristics, which will broaden our understanding of the regulation of gene expression.

- [1] K. Nakai and A. Vandenbon. (Chapter 2) "Higher-order chromatin structure and gene regulation." (Chandra Boosani and Ritobrata Goswami eds.) Epigenetics in Organ Specific Disorders. Academic Press, pp.11-32, 2023.
- [2] ENCODE Project Consortium, et al. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." Nature 583 (7818), 2020.
- [3] Fulco, C. P. et al. "Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations". Nature Genet. 51(12), 2019.
- [4] Dao, Lan TM, et al. "Genome-wide characterization of mammalian promoters with distal enhancer functions." Nature genetics 49.7, 2017.

CHARACTERIZATION OF SVS IN THE HUMAN PANGENOME REFERENCE

Shuangjia Lu¹, Wen-wei Liao^{1,2}, Haley J Abel³, Heng Li^{4,5}, Ira Hall^{1,6}

¹Yale University School of Medicine, Department of Genetics, New Haven, CT,

²Washington University School of Medicine, Division of Biology and Biomedical Sciences, St. Louis, MO, ³Washington University School of Medicine, Department of Medicine, St. Louis, MO, ⁴Dana-Farber Cancer Institute, Department of Data Sciences, Boston, MA, ⁵Harvard Medical School, Department of Biomedical Informatics, Boston, MA, ⁶Yale University School of Medicine, Center for Genomic Health, New Haven, CT

The most important use of the human reference genome is short read mapping, which is the first step of many genetic analyses such as variant genotyping and trait mapping. Despite widespread use of the current human reference genome, GRCh38, it possesses inherent limitations, namely a mosaic haplotype structure, 210Mb of sequence gaps, and poor representation of structurally polymorphic haplotypes. These limitations affect the accuracy and sensitivity of read alignment and downstream analyses. To overcome these issues, the Human Pangenome Reference Consortium (HPRC) has generated 94 high-quality haplotype assemblies from diverse populations using PacBio HiFi long-read sequencing technologies, and constructed multiple human pangenome reference graphs that aim to represent all sequences and genetic variants from those assemblies. By representing these sequences and variants, the pangenome reference mitigates short read alignment errors and bias, thus improving the accuracy of variant genotyping and trait mapping.

To characterize the representation of complex loci in the human pangenome graphs, we first identified 620 large multiallelic structural variants (SVs) in the pangenome graph, 44 of which overlap with clinically relevant protein coding genes. We then performed a detailed analysis of 5 clinically relevant multiallelic SV loci: RHD/RHCE, HLA-A, CYP2D6/CYP2D7, C4, and LPA. We visualized subgraph structures of these complex loci in the two independent pangenome graphs constructed by the HPRC, derived from minigraph-cactus (mc) and PGGB methods. After annotating the location of genes and paths of haplotypes in the pangenome subgraphs, we identified copy number variants, gene conversions and insertions within these genes and adjacent regions. In CYP2D6/CYP2D7, C4 and LPA, both pangenome graphs accurately recapitulated previously described haplotypes. For example, 96% of our CYP2D6/CYP2D7 genotyping results of each assembly matched with short read genotyping results called by a published tool, Cyrus. In RHD/RHCE, besides previously described haplotypes, we discovered 5 novel haplotypes, including one RHD duplication, one inversion, and 3 gene conversions. In the HLA-A locus, we found a long novel inserted sequence (~65kb) around the HLA-Y pseudogene, which has low homology with sequences in the GRCh38 reference genome. Our result also demonstrated how the mc and PGGB graphs use different approaches to represent adjacent homologous sequences. The mc graph tends to separate copies of adjacent homologous sequences, while the PGGB graph tends to collapse homologous sequences into a single copy. Moving forward, we will explore the potential of the pangenome reference to improve SV detection and increase power for trait mapping applications.

STRATEGIES FOR IDENTIFYING HIGH-CONFIDENCE DE NOVO MUTATIONS IN SOMATIC AND GERMLINE CELLS THROUGH DUPLEX SEQUENCING OF DIVERSE TISSUE TYPES

Suchita Lulla¹, Jason Kunisaki¹, Laurel Hiatt¹, Michael Goldberg¹, Kenneth Aston², Aaron Quinlan¹

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

Identification of de novo mutations (DNMs) and the rate at which they accumulate have applications in research and clinical work, especially related to cancer and genetic disease risk. Several research efforts have used whole-genome sequencing (WGS) technologies to analyze DNMs.

However, WGS has an error rate on the order of 10^{-3} , which presents challenges when distinguishing low-frequency mutations (signal) from sequencing artifacts (noise) in bulk tissue samples. TwinStrand Duplex Sequencing overcomes the limitations associated with WGS by reducing error rates to the order of 10^{-8} through exclusive reporting of variants that occur on complementary DNA strands, allowing more accurate quantification of DNM rates. Here, we present a computational workflow to quality control and examine DNMs detected with TwinStrand Duplex Sequencing across bulk somatic and germline tissue samples.

We apply duplex sequencing to two different cohorts. First, we analyze DNMs in a cohort of rapid autopsy cadavers where we collect blood, colon, skin, prostate, testicular tissue, and sperm from each individual. With this cohort, we aim to compare mutation rates across somatic and germline tissue types within the same individual and evaluate mutagenic etiologies in cancerous, pre-cancerous, and noncancerous colon. These pilot studies will guide our future investigations of the relationship between increased age and degrees of germline and somatic mutagenesis. Second, we duplex sequence bulk sperm and blood from 35 fertile men and subfertile men aged between 24 to 68 years. With this data, we aim to explain the association between male infertility and negative health outcomes, including reduced lifespans and elevated cancer risk, by testing whether infertile men harbor elevated germline and somatic mutation rates relative to fertile men. We will further conduct mutation signature analyses to nominate biological processes that could illuminate the underpinnings of mutagenesis, reduced sperm counts, and poor health in subfertile men.

Investigating tissue specificity of somatic mosaicism requires us to differentiate between DNMs, germline variants, and sequencing artifacts. We therefore present a computational pipeline to refine duplex sequencing data to high-confidence DNM calls through several filtering strategies. We will present ongoing work identifying erroneous variant calls, categorizing clonal and subclonal mutations, and distinguishing tissue-specific DNMs from inherited variants in order to compare, contrast, and quantify mosaicism between tissues.

NUCLEOTIDE DETERMINANTS OF DERIVED ENHANCER FUNCTION IN HOMININ EVOLUTION

Riley J Mangan, Craig B Lowe

Duke University Medical Center, Molecular Genetics and Microbiology,
Durham, NC

Evolutionary changes in enhancer elements have shaped human-characteristic phenotypes through the adaptive modification of gene regulation. We recently characterized HAQERs, human ancestor quickly evolved regions, as the most divergent regions of the human genome, which rapidly evolved under positive selection and elevated mutation rates to generate new enhancer elements unique to the human lineage. Here, we hypothesized that derived enhancer function in HAQERs was mediated by substitutions in transcription factor binding motif sequences, altering DNA/transcription factor interactions. Our ongoing work combines transcription factor motif scanning with comparative sequence alignments to quantify genome-wide patterns of motif gain, loss, and modification on the human lineage. We also employ deep learning models for sequence regulatory prediction to perform in silico saturated mutagenesis in HAQERs to relate nucleotide mutations in transcription factor motifs observed in evolutionary history to their inferred regulatory significance.

SINGLE-CELL MULTI-OMICS IN FETAL DOWN'S SYNDROME REVEALS THE IMPACT OF ANEUPLOIDY IN CELLULAR DIFFERENTIATION AND GENE REGULATION

Andrew R Marderstein¹, Marco De Zuani², Haoliang Xue², Jon Bezney¹, Shuo Wong², Stephen B Montgomery*¹, Ana Cvejic*²

¹Stanford, Dept Pathology, Stanford, CA, ²Univ of Cambridge, Dept Haematology, Cambridge, United Kingdom

Introduction: Aneuploidy has unique impacts on genome biology. Inborn trisomy of chromosome 21 (Ts21, Down's Syndrome) predisposes individuals to leukemia in a process that is initiated before birth. However, the molecular processes in Ts21 cells preceding any mutational processes toward leukemia are unknown because it has been challenging to study throughout key developmental contexts.

Methods: From 15 Ts21 and 3 healthy fetuses, we combined 1) >1.1 million scRNA-seq cells using matched human fetal liver and bone marrow; 2) 10X Visium spatial transcriptomics; and 3) 10X multiome of 56,890 cells, to examine the molecular impact of trisomy 21 in blood development.

Results: We found that expression of non-chr21 genes were cell type- and environment-dependent, resulting in markedly different cell type composition and spatial organization between Ts21 and healthy samples. Using non-coding fine-mapped GWAS SNPs to tag enhancers relevant to blood differentiation, we found that GWAS-harboring peaks for red blood cell count were more accessible in a subpopulation of Ts21 hematopoietic stem cells (HSCs).

To understand underlying regulatory mechanisms driving lineage bias in Ts21 and implicate causal genes, we correlated enhancer accessibility with gene expression. We identified 4.1-times more significant peak-gene links in Ts21 HSCs compared to healthy HSCs. 11.2% of Ts21-specific peak-gene links were due to trisomy modifying the effect of peak accessibility on gene expression, while the remaining peak-gene links were significantly more accessible (2.1-fold enrichment) and upregulated (1.4-fold enrichment) in Ts21 compared to healthy. Thus, while some enhancer-gene links were perturbed by trisomy, greater activation of regulatory elements and more widespread transcription in Ts21 HSCs allowed many more peak-gene links to be discovered.

Finally, we intersected fine-mapped GWAS SNPs for blood cell traits with enhancer-gene links to prioritize likely causal mechanisms. GWAS-harboring peaks were enriched for association with gene expression.

Furthermore, target genes were more often differentially expressed between Ts21 and healthy HSCs ($P < 0.05$), suggesting a key role for these enhancers and their target genes in driving differentiation of Ts21 HSCs towards the erythroid lineage, which we subsequently confirmed *in vitro*.

Conclusion: Our results demonstrate how integrating single-cell multi-omics with blood cell trait GWAS illuminates the role of genetic background (aneuploidy) in gene regulation early on in development.

CROSS-ANCESTRY, CELL-TYPE-INFORMED ATLAS OF GENE, ISOFORM, AND SPLICING REGULATION IN THE DEVELOPING HUMAN BRAIN

Cindy Wen^{1,2,3}, Michael Margolis^{2,3}, Rujia Dai⁴, Pan Zhang^{2,3}, Paweł Przytycki⁵, Daniel Vo^{2,3,4}, Bogdan Pasaniuc^{1,3}, Jason Stein⁶, Michael Love⁶, Katherine Pollard⁵, Chunyu Liu⁴, Michael Gandal^{1,2,3,7}

¹University of California, Los Angeles, Interdepartmental Program in Bioinformatics, Los Angeles, CA, ²David Geffen School of Medicine, University of California, Los Angeles, Department of Psychiatry, Los Angeles, CA, ³David Geffen School of Medicine, University of California, Los Angeles, Department of Human Genetics, Los Angeles, CA, ⁴SUNY Upstate Medical University, Department of Psychiatry, Syracuse, NY, ⁵Gladstone Institute of Data Science and Biotechnology, Gladstone Institute of Data Science and Biotechnology, San Francisco, CA, ⁶University of North Carolina at Chapel Hill, Department of Genetics, Chapel Hill, NC, ⁷The Children's Hospital of Philadelphia, Lifespan Brain Institute, Philadelphia, PA

Genomic regulatory elements active in the developing human brain are notably enriched for genetic risk for neuropsychiatric disorders, including autism, schizophrenia, and bipolar disorder. However, prioritizing the specific risk genes and candidate molecular mechanisms underlying these genetic enrichments has been hindered by the lack of a single unified large-scale gene regulatory atlas of human brain development. Here, we build a comprehensive set of developmentally regulated gene and isoform co-expression networks across 672 fetal brain samples, and overlay enrichments for cell markers, common variation, and rare variation in order to contextualize fetal gene regulation into disease-enriched gene modules. Of note, we identify a group of gene-level modules, with hub genes including EP300, EP400, ARID1A, and POGZ, that enrich for chromatin remodeling pathways, converge with neuronal marker genes, and strongly enrich for rare variation associated with autism and developmental delay.

THE HIGH-COVERAGE GENOME OF A MALE NEANDERTAL

Diyendo Massilani¹, Stéphane Peyrégne², Cesare De Filippo², Leonardo N Iasi², Alba Bossoms Mesa², Divyaratna Popli², Arev Pelin Sümer², Christian Heide², Maxim B Kozlikin³, Michael V Shunkov³, Anatoly P Derevianko³, Samantha Brown⁴, Thomas Higham⁵, Katerina Douka⁵, Matthias Meyer², Hugo Zeberg², Janet Kelso², Svante Pääbo²

¹Yale School of Medicine, Department of Genetics, New Haven, CT, ²Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, ³Siberian Branch of the Russian Academy of Sciences, Institute of Archeology and Ethnography, Novosibirsk, Russia, ⁴University of Tübingen, Institute for Scientific Archaeology, Tübingen, Germany, ⁵Faculty of Life Sciences, University of Vienna, Department of Evolutionary Anthropology, Vienna, Austria

To date only three high-coverage Neandertal genome sequences are available. Using collagen fingerprinting, we identified an undiagnostic bone fragment from Denisova Cave (Layer 12, East Chamber) in the Altai Mountains (Russia) as being of hominin origin. We named this individual Denisova 17. Sequencing of DNA libraries prepared from the specimen revealed that ~73% of the DNA fragments sequenced map to the human genome and that less than 1% of these fragments are present-day human DNA contamination. The exceptional ancient DNA preservation allowed the genome to be sequenced to 35-fold average genomic coverage. Analyses of the genome reveals that Denisova 17 was a Neandertal male individual. This is the first male Neandertal from whom a genome has been sequenced to high coverage, providing the first insight into Neandertal Y chromosome structure. We estimated his age to be ~110,000 years before present based on the accumulation of nucleotide substitutions in his mitochondrial and nuclear genomes. This individual belonged to a population related to a ~120,000-year-old female individual from the same cave (Denisova 5 or Altai Neandertal) and is distinct from later Neandertals in the same region and in Western Eurasia. Similarly, to Denisova 5, the genome of Denisova 17 shows a high level of inbreeding, confirming that mating among close relatives was common among these early East Eurasian Neandertals. The availability of a fourth high-coverage Neandertal genome will allow genomic changes unique to Neandertals to be more reliably identified and allows us to describe additional haplotypes that were introgressed from Neandertals to present-day humans. Results of these analyses will be presented, including a haplotype unique to Denisova 17 that affects female reproductive physiology and was introgressed from Neandertals to present-day Asian populations.

LT-FREE: A NOVEL METHOD FOR LEVERAGING FAMILY HISTORY IN GENETIC ASSOCIATION STUDIES OF ARBITRARILY COMPLEX DISEASES.

Jamie Matthews¹, Mike Thompson², Noah Zaitlen^{3,4}

¹UCLA, Bioinformatics Interdepartmental Program, Los Angeles, CA, ²The Barcelona Institute of Science and Technology, Center for Genomic Regulation, Barcelona, Spain, ³UCLA, Department of Neurology, Los Angeles, CA, ⁴UCAL, Department of Computational Medicine, Los Angeles, CA

Rich phenotypes collected in biobanks have provided opportunities to improve genetic association study power via alternative analysis methods¹. For example, the LT-FH method leveraged family history data to estimate genetic liability for a given trait, thereby improving study power². However, this method imposes assumptions about underlying trait genetic architectures and requires external estimates of heritability and generational prevalence. Here we introduce LT-Free as a freer and interpretable model that does not require external information and does not assume any underlying model of genetic architecture. Briefly, LT-Free allows an arbitrary latent phenotype value for each combination of case-control status, sex, and family history data observed in the data. LT-Free then alters these phenotypes to optimize association test statistic values at SNPs known to be associated with the target phenotype. In data simulated under the LT-FH model, we find that LT-Free needs around 50 known GWAS SNPs to train on before it achieves comparable association statistics at held-out SNPs compared to LT-FH. Further, once the simulated genetic architecture deviates from the LT-FH model, fewer SNPs are needed. For example, under a strong sex effect, as few as 5 SNPs are needed before LT-Free obtains higher association statistics than LT-FH. We then apply LT-Free to 9 diseases in the UKBB and similarly meet or exceed the performance of LT-FH in 8 of the 9 diseases with significant improvement in 6 diseases: heart disease ($p = 6.2e-5$), COPD ($p = 1.2e-6$), bowel cancer ($p = 7.0e-5$), depression ($p = 2.5e-4$), hypertension ($p = 9.4e-4$), and lung cancer ($p = 2.2e-4$). In the one disease in which LT-Free did not perform as well as LT-FH, type II diabetes, the difference in association statistics between methods was not significant. In heart disease, we observe evidence of different latent phenotypes estimated across sexes and parental origin. Family history data is a powerful way to improve genetic association studies. By modelling this data with more general models, we can both improve study power and gain insights into genetics architectures.

References

1. <https://doi.org/10.1038/ng.3975>
2. <https://doi.org/10.1038/s41588-020-0613-6>

ANALYSIS OF ANCIENT BONE DNA SAMPLES FROM
EXCAVATIONS AT ST PETER'S BURIAL GROUND, BLACKBURN

Shakhawan Mawlood, Catriona Pickard, Benjamin Pickard

SIPBS, Glasgow, United Kingdom

In summer 2015 the remains of 800 children are among 1,967 bodies were exhumed by archaeologists at St Peter's Burial Ground in Blackburn, Lancashire. One hundred samples from these 19th century ancient bones were selected for DNA analysis. These comprised samples biased for those which prior osteological evidence indicated a potential for microbial infection by *Mycobacterium tuberculosis* (causing tuberculosis, TB) or *Treponema pallidum* (causing Syphilis) species, as well a random selection of other bones for which visual inspection suggested good preservation (and, therefore, likely DNA retrieval). They were subject to polymerase chain reaction (PCR) assays aimed at detecting traces of DNA from infecting mycobacteria, with the purpose both of confirming the palaeopathological diagnosis of tuberculosis and determining in individual cases whether disease and death was due to *M. tuberculosis* or other reasons. Our secondary goal was to determine sex determination and age prediction. The results demonstrated that extraction of vast majority ancient bones DNA samples succeeded.

SOCIAL ENVIRONMENTAL EFFECTS ON GENE REGULATION AND AGING IN A LARGE COHORT OF COMPANION DOGS.

Brianah M McCoy^{1,2}, Layla Brassington^{1,2}, Beth Slickas^{1,2}, The Dog Aging Project Consortium³, Noah Snyder-Mackler^{1,2}

¹Arizona State University, School of Life Sciences, Tempe, AZ, ²Arizona State University, Center for Evolution and Medicine, Tempe, AZ,

³University of Washington, Department of Biology, Seattle, WA

Positive socio-economic, built, and natural environmental factors confer environmental stability that plays a positive role in overall health and disease risk in both early and late life. Yet precisely how these effects impact health at the molecular level and may differentially act across the lifespan has been difficult to study in humans due to complex lifestyle factors and long lifespans. Here, we leveraged a relatively new and powerful model for human aging, the companion dog (The Dog Aging Project), to quantify how environmental factors (i.e., income, social relationships, environmental stability) alter aging and health at the organismal and molecular level. In a cross-sectional sample of 21,410 dogs, we found that poor environments are associated with accelerated age-related health decline and disease incidence. In an effort to understand how these environmental factors transduce into molecular and immunological health effects that could lead to disease, we performed reduced representation bisulfite sequencing (RRBS) to quantify environmental effects on CpG methylation changes in 165 dogs. We found that a myriad environmental factors were associated with DNA methylation at tens of thousands of CpG sites in the dog genome. We also saw that of our social environmental factors, environmental stability was significantly associated with methylation at the most CpG sites ($n= 20,202$) with 16,263 showing increased and 3,939 with decreased methylation as environmental stability increased. Currently, we are generating DNA methylation data for an additional 500 dogs, which will substantially improve our power to detect how variation in the environment might affect age-related immune function and inflammation. Overall, this study provides preliminary molecular evidence for the link between the social environmental stability, molecular changes, and health outcomes, in a novel model for human aging – the companion dog. These data suggest the importance of further investigating the effects of the social environment on dog health while also providing a framework for better understanding human aging through the lens of methylation changes that may persist for many years.

BEYOND THE EXOME: A GENOMICS-BASED UNDIAGNOSED GENETIC DISEASE RESEARCH PROGRAM

Stephen Meyn^{1,2}, Bryn D Webb^{1,2}, Derek Pavelec³, Heather Motiff¹, Jadin Heilmann¹, Xiang Qiang Shao^{2,4}, Vanessa Horner^{1,5}, April Hall^{1,2}

¹University of Wisconsin - Madison, Center for Human Genomics and Precision Medicine, Madison, WI, ²University of Wisconsin - Madison, Department of Pediatrics, Madison, WI, ³University of Wisconsin - Madison, Biotechnology Center, Madison, WI, ⁴University of Wisconsin - Madison, State Laboratory of Hygiene, Madison, WI, ⁵University of Wisconsin - Madison, Department of Pathology and Laboratory Medicine, Madison, WI

Background: Just over half of 9,000+ rare genetic disorders have a known cause and most patients who undergo clinical WES fail to obtain a diagnosis. To address these issues we created the University of Wisconsin Undiagnosed Disease Program (UW UDP), which takes a "beyond the exome" approach to evaluating genetics patients.

Objectives of the UW UDP are: 1) discover new disease genes; 2) improve our understanding of genetic disorders; 3) provide patients with actionable diagnoses; and 4) evaluate novel technologies. Our workflow begins with clinical WES reanalysis, followed by trio short read genome sequencing. Long read sequencing, RNA-Seq, and epigenomic profiling are utilized ad hoc.

Results: To date, the UW UDP has enrolled 53 probands and 108 relatives. >90% of probands had prior clinical WES. We identified candidate causal variants for 5 of the first 10 patients. Short and long read WGS, WES reanalysis, and RNA-Seq each played a role in finding a deletion and an instance of chromoplexy missed by clinical testing; three new candidate disease genes; and a patient whose novel phenotype is the likely result of synergy between two rare disorders. Additional analyses are on-going.

Conclusion: Our initial results suggest that clinical WES may be a suboptimal test, as a significant fraction of our clinical WES-negative patients were diagnosed using combinations of short and long-read WGS, supplemented by RNA-Seq. An undiagnosed genetic disease program can serve as an important component of a comprehensive center for rare diseases, as it offers patients access to emerging technologies and facilitate the discovery of new disease genes while advancing our understanding of rare genetic disorders.

HYPOXIA LEADS TO UNIQUE mtDNA TRANSCRIPTIONAL PATTERNS AND AFFECTS MITO-NUCLEAR REGULATORY COORDINATION

Noam Shtolz, Sara Dadon, Dan Mishmar

Ben-Gurion University of the Negev, Department of Life Sciences, Beer-Sheva, Israel

Mitochondria is a central player in cellular metabolism, cell life, and death. Unlike all cellular functions, the mitochondria are operated by a unique bi-genomics system that involves cooperation between factors encoded by the mitochondrial genome (mtDNA) and by the nuclear genome (nDNA).

Recently, we showed that such coordination is perturbed in human diseases, such as COVID19. Notably, these diseases are aggravated by exposure to environmental stresses, especially hypoxia. Therefore, we asked whether mitochondrial transcription and mito-nuclear co-expression, respond to environmental stresses. To address this question, we have grown three human cell lines (HeLa, U87, and D407) in two types of mitochondria-related stresses: either carbon source change (glucose vs galactose) or hypoxia (24hr 1% oxygen). These cells were subjected to in-vivo assessments of nascent transcription (PRO-seq), mitochondrial copy number (qPCR), and membrane potential. Our results revealed that whereas growth under galactose-based media did not result in consistent alteration of mitochondrial transcription, hypoxia did. Specifically, although mtDNA light strand transcriptional initiation was elevated in hypoxia, transcriptional elongation levels of the same strand dropped around mtDNA position ~8700 below the control in HeLa and U87 cells, but not in D407 cells. Additionally, the altered mtDNA transcription was associated with reduced transcription in nDNA-encoded OXPHOS genes, which was even more pronounced in RNA-seq. Interestingly, such changes did not associate with altered mtDNA copy number, hence suggesting a transcriptional regulatory change rather than an overall impact on mitochondrial biogenesis. The unexpected similarity in mtDNA transcriptional pattern alteration in hypoxia across cell lines suggests similarity in the underlying mechanism.

DOMAIN-ADAPTIVE NEURAL NETWORKS IMPROVE
SUPERVISED MACHINE LEARNING BASED ON SIMULATED
POPULATION GENETIC DATA

Ziyi Mo^{1,2}, Adam Siepel¹

¹Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology,
Cold Spring Harbor, NY, ²Cold Spring Harbor Laboratory, School of
Biological Sciences, Cold Spring Harbor, NY

Investigators have recently introduced powerful methods for population genetic inference that rely on supervised machine learning from simulated data. Despite their performance advantages, these methods can fail when the simulated training data does not adequately resemble data from the real world. Here, we show that this “simulation mis-specification” problem can be framed as a “domain adaptation” problem, where a model learned from one data distribution (the “source domain”, here consisting of simulated data) is applied to a dataset drawn from a different distribution (the “target domain”, here real data). By applying an established technique based on a gradient reversal layer (GRL), originally introduced for image classification, we show that the effects of simulation mis-specification can be significantly mitigated. We focus our analysis on two state-of-the-art deep-learning population genetic methods — SIA, which infers positive selection from features of the ancestral recombination graph (ARG), and ReLERNN, which infers recombination rates from genotype matrices. In the case of SIA, the domain adaptive framework also compensates for ARG inference error. Using the domain-adaptive SIA model, we estimate improved selection coefficients at selected loci in the 1000 Genomes CEU population. We anticipate that domain adaptation will prove to be widely applicable in the growing use of supervised machine learning in population genetics.

A UNIFIED COMPUTING ENVIRONMENT FOR GENOMICS DATA STORAGE, MANAGEMENT, AND ANALYSIS: NHGRI GENOMIC DATA SCIENCE ANALYSIS, VISUALIZATION, AND INFORMATICS LAB-SPACE (AnVIL)

Stephen L Mosher¹, Michael C Schatz^{1,2}, Anthony Philippakis³, AnVIL Team⁴

¹Johns Hopkins University, Biology, Baltimore, MD, ²Johns Hopkins University, Computer Science, Baltimore, MD, ³Broad Institute of MIT and Harvard, Data Sciences Platform, Cambridge, MA, ⁴The full list of contributors is available at: <https://anvilproject.org/about/team>, Baltimore, MD

Recent years have seen astronomical growth in human genomics. Together with single-cell and functional genomics, electronic medical records and other biomedical data, the field is well-positioned to make great advances in human health. However, the complexity of genomic data sharing, where data is downloaded from centralized datastores for local analysis, is unsustainable and cost prohibitive. Furthermore, housing genomic data across redundant institutional compute infrastructures makes assuring data security and compliant usage of protected data a massive challenge.

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space, or AnVIL (<https://anvilproject.org/>) was developed to address these and other concerns by providing a unified cloud-based computing environment for genomics data storage, management and analysis. The AnVIL platform inverts the genomics data sharing model by eliminating the need for data movement, which in turn allows for active threat detection and monitoring and provides scalable, shared computing resources for researchers as needed.

AnVIL currently provides harmonized access to more than 600,000 genomes from several key NHGRI projects, such as as the CCDG (Centers for Common Disease Genomics), CMG (Centers for Mendelian Genomics), eMERGE (Electronic Medical Records and Genomics), GTEx (Genotype-Tissue Expression Project), T2T (Telomere-to-Telomere), and many more.

The platform is built on a set of established components that have been used in a number of flagship scientific projects. The Terra platform provides a compute environment with secure data and analysis sharing capabilities. Dockstore provides standards based sharing of containerized tools and workflows. Jupyter, R/Bioconductor and Galaxy provide analysis environments for users at all skill levels to interactively explore and understand data with thousands of tools available. The Gen3 data commons framework provides data and metadata ingest, querying, and organization. Together, AnVIL provides a collaborative environment for creating, analyzing, and sharing data and analysis workflows for even the largest projects.

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.

ACCURATE *DE NOVO* DETECTION OF SOMATIC MUTATIONS IN SINGLE-CELL GENOMICS AND TRANSCRIPTOMICS DATA

Francesc Muyas¹, Ruoyan Li², Thomas J Mitchell^{2,3,4}, Sahand Hormoz^{5,6,7}, Isidro Cortés-Ciriano¹

¹European Molecular Biology Laboratory, European Bioinformatics Institute, Cancer genomics, Hinxton, United Kingdom, ²Wellcome Trust Sanger Institute, Cancer, Ageing and Somatic Mutation, Hinxton, United Kingdom, ³Cambridge University Hospitals NHS Foundation Trust, NIHR Cambridge Biomedical Research Centre, Cambridge, United Kingdom,

⁴University of Cambridge, Department of Surgery, Cambridge, United Kingdom, ⁵Harvard Medical School, Department of Systems Biology, Boston, MA, ⁶Dana-Farber Cancer Institute, Department of Data Science, Boston, MA, ⁷Broad Institute of MIT and Harvard, Affiliate Faculty, Cambridge, MA

Detecting somatic mutations at single-cell resolution is essential to study genetic heterogeneity, clonal mosaicism in non-neoplastic tissues, and to identify the mutational processes operative in malignant and phenotypically normal cells. However, the identification of mutations in individual cells is still challenging from a technical and algorithmic standpoint. Here, we present *SComatic*, an algorithm designed to detect somatic mutations *de novo* in single-cell transcriptomic and ATAC-seq data sets. Using more than 2 million single cells from 420 single-cell RNA-seq and ATAC-seq data sets, we show that *SComatic* can detect somatic mutations not only in tumour samples, but also in differentiated cells from polyclonal tissues not amenable to mutation detection using existing methods. In addition, *SComatic* allows the accurate estimation of mutational burdens and *de novo* mutational signature analysis at cell-type resolution. Using matched DNA sequencing and single-cell RNAseq data, we show that *SComatic* has higher precision (> 10-fold) than existing algorithms for detecting somatic mutations without compromising sensitivity. Overall, *SComatic* permits the study of somatic mutagenesis at unprecedented scale and resolution using high-throughput single-cell profiling data sets.

INTRASPECIFIC VARIATION OF TRANSPOSABLE ELEMENT DYNAMICS AND PROTEIN FAMILIES IN A FUNGAL PHYTOPATHOGEN REVEAL DIFFERENCES IN THE EVOLUTIONARY HISTORY OF ITS VARIOUS PATHOTYPES

Anne A Nakamoto^{*1}, Pierre M Joubert^{*1}, Daniil M Prigozhin², Ksenia V Krasileva¹

¹University of California, Berkeley, Department of Plant and Microbial Biology, Berkeley, CA, ²Lawrence Berkeley National Laboratory, Berkeley Center for Structural Biology, Molecular Biophysics and Integrated Bioimaging Division, Berkeley, CA

*These authors contributed equally

Fungi adapt to a wide range of lifestyles and environments. Transposable elements (TEs) are known to play important roles in fungal genome evolution, however, our understanding of the processes that shape TE landscapes is limited, as is our understanding of the relationship between TE content, population structure, and evolutionary history of fungal species. Fungal plant pathogens, which often have host-specific populations within the same species, are useful systems in which to study intraspecific TE content diversity. Here, we completed the annotation and phylogenetic classification of TEs in five host-specific lineages of *Magnaporthe oryzae*, the fungus that causes blast disease of rice, wheat, and many other grasses. We identified differences in TE content between these lineages, and showed that recent lineage-specific expansions of certain TEs have contributed to greater TE content in rice-infecting and *Setaria*-infecting lineages. We reconstructed the histories of LTR-retrotransposon expansions and found that they were caused by complex proliferation dynamics of one element, or by multiple elements from an older population of TEs proliferating in parallel. Additionally, we found evidence suggesting the recent transfer of a DNA transposon between rice and wheat-infecting *M. oryzae* lineages, and a region showing evidence of homologous recombination between those lineages, which could have facilitated such a transfer. By investigating intraspecific TE content variation, we uncovered key differences in the proliferation dynamics of TEs in various pathotypes of a fungal plant pathogen. These differences additionally gave us a better understanding of the evolutionary history of the pathogen itself.

Genes involved in innate immunity are also thought to play a role in the adaptation of fungi to their environment, and may have similar evolutionary characteristics and genomic contexts to TEs. We are currently extending this study to include the characterization of fungal proteins that may be involved in innate immunity, which are currently largely uncharacterized in fungal systems. By investigating innate immune receptor diversity and evolution across fungal plant pathogen lineages, we hope to gain a better understanding of how their genomes evolve with respect to both pathogenesis and defense against their environment.

USING SNAKES REARRANGEMENT DISPLAY TO VISUALIZE PAIRWISE ALIGNMENTS ON THE UCSC GENOME BROWSER

Luis R Nassar, Brian J Raney, Mark Diekhans, Maximilian Haeussler, William J Kent, Galt P Barber, Jonathan Casper, Hiram Clawson, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Christopher M Lee, Gerardo Perez

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

For almost two decades the UCSC Genome Browser has shown genome-genome alignments using a display mode called "chains" which presents the alignable portions of a second genome sequence on a reference genome sequence, with alignable consecutive segments connected by intron-like lines. Other genome browsers approach the problem using similar display types, recycling drawing code largely from gene displays. When inversions or translocations occur, the annotations are moved to the next row of the display. The problem with this display is that it does not show the order of the alignable segments in the other genome, only the segments which are alignable and consecutive.

Snakes display expands on the chains by connecting segments with lines in the order that they appear on the other genome. This more clearly shows inversions and translocations, as well as duplications. Mismatching bases within the alignment are also marked with colors. Clicking into specific alignments allows one to see a dot plot representation of the alignment and an option to see the alignment using the other genome as a reference.

In an age of more and more sequenced genomes, especially human ones, and an increased interest in structural variants, we believe that this way to show genome rearrangements is a good compromise between display types that break the paradigm of a linear sequence ("Tube maps", "Circos plots") and the existing simple consecutive-segments based displays of genome browsers.

MODELING THE IMPACT OF RARE STRUCTURAL VARIANTS ON GENE EXPRESSION IN RARE DISEASE CASES

Bohan Ni¹, Tanner Jensen², Pagé Goddard², Rachel Ungar², Benjamin Strober³, Nicole Ersaro², Taibo Li⁴, Euan A Ashley⁵, Matthew Wheeler⁵, Stephen B Montgomery², Michael C Schatz¹, Alexis Battle^{1,4}

¹Johns Hopkins University, Computer Science, Baltimore, MD, ²Stanford University, Genetics, Stanford, CA, ³Harvard University, Epidemiology, Cambridge, MA, ⁴Johns Hopkins University, Biomedical Engineering, Baltimore, MD, ⁵Stanford University, Medicine, Stanford, CA

Rare and private variants have been shown to contribute to rare diseases. While rare Single Nucleotide Variants (SNVs) are routinely examined for functional effects, rare Structural Variants (SVs) are often not considered due to challenges in SV calling. Further, predicting the functional impact of noncoding SVs remains challenging, especially for SVs disrupting gene regulation. To prioritize rare and functional SVs, we developed Watershed-SV, a probabilistic graphical model that integrates transcriptomics outlier signals with SV-specific annotations such as length, type, and summary scores of SV-impacted regions. Trained using the GTEx project v8 SV calls and gene expression, Watershed-SV produces substantial improvements over baseline models in both tissue-aggregate and tissue-specific models. Watershed-SV models more consistently explain expression outliers (eOutliers) than Watershed-SNV models, confirming the increased impact of rare SVs on gene expression.

We validated our model's performance using the NHGRI Inherited Muscular Disease (MD) dataset and the Undiagnosed Disease Network datasets. We generated an AnVIL workflow using Jasmine and Paragraph to ascertain rare variants, especially for novel SVs in UDN long-read calls where the majority were unseen in short-read data. To call eOutliers, we combined patient RNA-seq with tissue-matched data from GTEx and RECOUNT3. Our model correctly prioritized known SV-gene pairs implicated in MD. We also identified novel disease-related SVs in patients that affect coding and non-coding regions. For example, Watershed-SV prioritized a deletion in TANGO2, a gene with a strong link to arrhythmia, in a patient with undefined cardiomyopathy (Watershed posterior=0.97). We also prioritized a deletion of a muscle-specific enhancer near TLN1, a gene with a strong link to cardiomyopathy, in a patient with congenital myopathy (Watershed posterior=0.95) while baseline models failed to prioritize the variant.

In summary, Watershed-SV prioritizes functional rare SVs using SV-relevant annotations and transcriptomic data. It performs well in held-out validation and is able to prioritize rare coding and noncoding disease SVs. Finally, we provide a reusable workflow to practically select rare SVs for evaluation and calling eOutliers, making Watershed-SV accessible to rare disease diagnoses.

BABOONXCAN, A FRAMEWORK TO IDENTIFY GENES ASSOCIATED WITH PHENOTYPES IN BABOONS

Festus Nyasimi¹, Wenhe Lin², Ellen Quillen³, John VandeBerg⁴, Deborah Newman⁵, Michael Olivier³, Jeff Wall⁶, Laura A Cox³, Yoav Gilad^{1,2}, Hae Kyung Im¹

¹The University of Chicago, Department of Genetic Medicine, Chicago, IL,

²The University of Chicago, Department of Human Genetics, Chicago, IL,

³Wake Forest University School of Medicine, Center for Precision

Medicine, Winston-Salem, NC, ⁴University of Texas Rio Grande Valley,

Edinburg, TX, ⁵Texas Biomedical Research Institute, San Antonio, TX,

⁶HIBio, San Francisco, CA

Genetic predictors of gene expression traits are commonly used to identify genes associated with complex diseases and traits in humans, using methods such as PrediXcan. Baboons have been shown to be a valuable model system for studying human diseases due to their physiological and genetic similarity to humans. However, a framework for predicting gene expression from genotype data does not currently exist for baboons. To address this, we set out to build such a framework, leveraging data from a tissue-specific eQTL study that included whole-genome sequencing and RNA sequencing data from 99 baboons.

The data included RNA sequencing in Adipose, Liver, and Muscle tissues before and after two years of a high-fat diet. We trained genetic predictors using the elastic net pipeline for each time point and tissue. We found no substantial differences between the prediction models for the two time points, so we used the average of the two, which yielded more predictors than any one time point. Using a stringent nested cross-validated prediction performance correlation greater than 10%, we obtained prediction models for 4344, 3392, and 3560 genes for Adipose, Liver, and Muscle tissues, respectively. The median prediction performance was 0.384, 0.391, and 0.386. Interestingly, we found that genes that are well-predicted in baboons tend to be well-predicted in humans as well.

To explore the utility of our framework, we conducted a preliminary association analysis between predicted gene expression levels and body weight in 277 baboons. This analysis identified several interesting genes, including ZCRB1, NVL, FAM189B, GRM4, P3H4, and PRRT3, which were associated with various body size and impedance phenotypes in the PhenomeXcan database based on the UK Biobank and other large scale human studies. Overall, our results suggest that our framework could be a valuable tool for studying complex diseases and traits in baboons, which could ultimately shed light on similar traits in humans.

Acknowledgement

Minor wordsmithing with chatGPT + last sentence added by chatGPT

IDENTIFICATION OF HOST GENES ASSOCIATED WITH COVID-19 RISK AND SEVERITY BY ANCESTRY-AWARE TRANS-LAYER MULTI-OMIC ANALYSIS

Meritxell Oliva, Justyna A Resztak, Sabah Kadri, Jacob Degner

AbbVie Inc., Computational Genomics, Genomics Research Center, Chicago, IL

Since the beginning of 2020, SARS-CoV-2 infection and its disease, COVID-19, have caused the largest contemporary pandemic to date. While many efforts are being devoted to the characterization of the genetic architecture of COVID-19 effects on human host, its underlying molecular basis has not been exhaustively explored across multiple molecular layers. To understand host response and to prioritize treatment targets, we sought to identify human genes influencing genetically-driven disease risk and severity. To this end, we performed ancestry-aware, trans-layer, multi-omic analyses by integrating recent (April 8, 2022) COVID-19 Host Genetics Initiative GWAS data from six ancestry endpoints - African, Amerindian, South Asian, East Asian, European and meta-ancestry - with functional maps and QTL catalogs.

We explored 91 GWAS hits ($P < 5e-7$), 28% of which were identified in a single ancestry. We analyzed a comprehensive set of 300 cis QTL maps from ~100 biotype sources for colocalization, including disease-relevant biotypes and contexts; blood of COVID-19 patients, large airway epithelium, and lung cell contexts. Across all GWASs, QTL maps and molecular phenotypes, we identified thousands of colocalizations ($PP4 > 0.75$) involving >100 genes. For 48 loci, molecular links were derived from at least two molecular phenotypes (m/eQTLs, m/pQTLs, e/pQTLs); including 5 loci colocalized with eQTLs and pQTLs for OAS1, ABO, CSF3, NPNT and NSF proteins. For 7 loci, molecular links were identified exclusively by ancestry-aware colocalization. For 8 loci, we identified gene associations in lung cell-type specific contexts, including DSP and GPX4 in lung epithelial and myeloid cells, respectively. DSP is specifically expressed in lung epithelial cells and is associated with idiopathic pulmonary fibrosis by limiting the ability of lung cells to repair injuries. GPX4 is involved in ferroptosis, feature linked to COVID-19, and the corresponding genomic locus is associated to blood hemolysis.

Overall, by performing an ancestry-aware and pulmonary cell-aware GWAS-QTL colocalization approach, we provide orthogonal, robust evidence of molecular links to COVID-19 associated genomic loci that contribute to our understanding of host response and prioritization of treatment targets.

Disclosures:

All authors are employees of AbbVie. The design, study conduct, and financial support for this research were provided by AbbVie. AbbVie participated in the interpretation of data, review, and approval of the publication.

SYSTEMATIC CHARACTERIZATION OF REGULATORY VARIANTS OF BLOOD PRESSURE GENES

Winona Oliveros^{*1}, Kate Delfosse^{*2}, Daniella F Lato², Katerina Kiriakopulos^{2,3}, Milad MokhtariDoust², Abdelrahman Said², Brandon J McMurray², Jared W Browning^{2,3}, Kaia Mattioli⁴, Guoliang Meng⁵, James Ellis^{5,3}, Seema Mital^{2,6,7}, Marta Melé^{*1}, Philipp G Maass^{*2,3}

¹Barcelona Supercomputing Center, Life Sciences Department, Barcelona, Spain, ²The Hospital for Sick Children, Genetics & Genome Biology Program, Toronto, Canada, ³University of Toronto, Department of Molecular Genetics, Toronto, Canada, ⁴Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, ⁵The Hospital for Sick Children, Developmental and Stem Cell Biology Program, Toronto, Canada, ⁶Ted Rogers Centre for Heart Research, Ted Rogers Centre for Heart Research, Toronto, Canada, ⁷The Hospital for Sick Children, Department of Pediatrics, Toronto, Canada

High blood pressure (BP) is the major risk factor for cardiovascular disease. Genome-wide association studies have identified genetic variants for BP, but functional insights into causality and related molecular mechanisms lag behind. We functionally characterize 4608 genetic variants in linkage with 135 BP loci in vascular smooth muscle cells and cardiomyocytes by massively parallel reporter assays. High densities of regulatory variants at BP loci (i.e. ULK4, MAP4, CFDP1, PDE5A) indicate that multiple variants drive genetic association. Regulatory variants are enriched in repeats, alter cardiovascular-related transcription factor motifs, and spatially converge with genes controlling specific cardiovascular pathways. Using heuristic scoring, we define likely causal variants for BP genes and CRISPR prime editing finally determines causal variants for KCNK9, SFXN2, and PCGF6, which are candidates for developing high BP. Our systems-level approach provides a catalogue of functionally relevant variants and their genomic architecture for a better understanding of blood pressure gene regulation.

THE ALPHA-2A ADRENERGIC RECEPTOR (ADRA2A) MODULATES SUSCEPTIBILITY TO DYSAUTONOMIA AND RAYNAUD'S DISEASE

Anniina Tervi¹, Markus Räsänen², Samuel E Jones¹, Caroline Heckman¹, Erik Abner³, Tonus Esko³, Jacqueline M Lane⁴, Matthew Maher⁵, FinnGen FinnGen¹, Estonian biobank research team³, Richa Saxena^{5,6}, Thomas Quertermous², Hanna M Ollila^{1,5,6}

¹University of Helsinki, FIMM, HiLIFE, Helsinki, Finland, ²Stanford University, Division of Cardiovascular Medicine, Palo Alto, CA,

³University of Tartu, Institute of Genomics, Tartu, Estonia, ⁴Brigham and Women's Hospital and Harvard Medical School, Division of Sleep and Circadian Disorders, Boston, MA, ⁵Massachusetts General Hospital, Center for Genomic Medicine, Boston, MA, ⁶Massachusetts General Hospital, Anesthesia, Critical Care, and Pain Medicine, Boston, MA

Raynaud's disease is a dysautonomia where exposure to cold increases the vascular tone of distal arteries causing vasoconstriction and hypoxia particularly in fingers and toes. Using genetic and electronic health record data from the UK Biobank, the Mass-General Brigham Biobank, the Estonian Biobank and from the FinnGen study we identified 10,680 individuals with a diagnosis for Raynaud's disease and 1,088,877 population controls. We found six loci suggesting endothelial nitric oxide (NOS3) and immune function (HLA) and a notable association at the alpha-2A adrenergic receptor (ADRA2A) locus (rs7090046, P = 1.51 x 10-47) implicating adrenergic signaling as a major risk factor with Raynaud's disease. Functional follow-up analysis revealed an eQTL that colocalized and increased ADRA2A gene expression in tissue specific manner in the distal arteries. Staining with RNA scope further clarified the specificity of ADRA2A expression in small vessels. Finally, functional contraction assay in ADRA2A deficient smooth muscle cells had lower contraction than smooth muscle cells with ADRA2A and the effect was temperature dependent having larger contraction in cold than in normal body temperature. Our results indicate that Raynaud's disease is related to immunity and endothelial function mediated by nitric oxide and adrenergic signaling where ADRA2A modulates vascular tone in Raynaud's disease in temperature dependent fashion.

RETROSPECTIVE LINEAGE TRACING AND PHENOTYPIC PROFILING IN HUMAN TISSUES BY DROPLET SINGLE CELL MICROSATELLITE SEQUENCING

Nathaniel D Omans^{1,2}, Tamara Prieto^{1,2}, John Zinno^{1,2}, Jake Qiu^{1,2}, Shu Wang³, Lucy A Godley⁴, Dan A Landau^{1,2}

¹Weill Cornell Medicine, The Meyer Cancer Center, New York, NY, ²The New York Genome Center, New York, NY, ³Mission Bio, Inc, South San Francisco, CA, ⁴The University of Chicago, Depts. of Medicine and Human Genetics, Chicago, IL

Joint single cell lineage tracing and phenotyping is a powerful approach to study developmental and somatic evolution processes, but existing methods rely on prospective tracking of synthetic DNA barcodes, precluding use in primary human samples. Phylogenetic modeling of somatic mutations offers the ability to infer lineages retrospectively in humans and reveal historic population dynamics, including timing of driver events, population sizes, and growth rates. However, due to the sparsity of somatic mutations in non-cancerous tissues, previous studies have used whole genome sequencing of *in vitro* expanded clones, which restricts application to dividing cells and limits co-profiling of cell phenotype.

To overcome these limitations, we hypothesized that mutable microsatellites can serve to build high resolution phylogenies in primary human cells. We therefore developed a scalable method ('Phylocity') for the targeted sequencing of 1121 highly mutable microsatellites in single cells. Using simulated evolution of published microsatellite mutation rates, we show that the somatic variation of 1121 microsatellites is sufficient to reconstruct lineage histories of thousands of cells after 100 cell divisions accurately (quartet similarity=97%). Additionally, we used a cell culture model of somatic evolution wherein sequential single cell cloning and expansion produced a set of clones with known pedigrees. Phylocity data collected on 5617 cells admixed from these clones allowed us to compare cladic genotypes to known clone-specific microsatellite variants, thus validating the Phylocity approach. Furthermore, we demonstrate that coalescent models on the Phylocity tree accurately estimate the experimental timing between cloning and expansion of *in vitro* clones.

Next, we applied Phylocity to investigate hematopoietic cells from an individual with *DNMT3A* mutated clonal hematopoiesis, combining the microsatellite panel with DNA barcoded antibodies for cell type-specific markers. Phylogenetic branch lengths of hematopoietic stem cells were shorter than more differentiated cells, consistent with their relatively small number of lifetime cell divisions. To further validate the tree topology, we annotated cells with *DNMT3A* genotypes, and observed two major clades concordant with the mutation status (F1 Score=0.87). Finally, joint cell surface profiling of progenitors revealed the known myeloid differentiation bias in the *DNMT3A* mutated clade.

In summary, we report on the development of Phylocity, a microsatellite-based strategy for joint single cell lineage tracing and phenotyping in human tissues. The throughput of thousands of cells allows for high resolution study of cellular phenotypes in light of their phylogenetic history. This high throughput, tissue agnostic technology will enable the study of key biological processes such as developmental trajectories, differentiation biases, and phylodynamic analyses directly in human tissues.

FROM "ALASKAN THUNDERF*CK" TO "MAUI WOWIE:" THE GENETIC ARCHITECTURE OF CANNABINOID CONCENTRATION IN 500 STRAINS OF POT.

Sara J Oppenheim¹, Armin Scheben², Dean M Bobo³, Robert DeSalle¹

¹American Museum of Natural History, Invertebrate Zoology, NYC, NY, ²Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ³American Museum of Natural History, Institute for Comparative Genomics, NYC, NY

Cannabis sativa has been under human cultivation for 5,000+ years, and present-day strains reflect this history of selection for traits related to *Cannabis*'s value as fiber, food, medicine, and mind-altering substance. Drug-type *Cannabis* is now legal in several states. As *Cannabis* becomes a mainstream crop, consumers expect products that predictably deliver specific effects (e.g., nausea reduction in chemotherapy patients). Instead, the emergence of *Cannabis* from the shadowy world of illegal drugs to the realm of big business has been accompanied by a host of issues not found in any other commercial crop.

The phenotype of interest to most growers and consumers of *Cannabis* is the relative concentration of the psychoactive metabolite Δ9-THC and the non-psychoactive CBD. This concentration determines the effect of *Cannabis* on human users because Δ9-THC and CBD compete for the same receptors in the human brain. Precise control of CBD:THC ratios is the holy grail of *Cannabis* cultivation, and the biosynthesis pathway that leads to the production of CBD and THC is now well understood. The genomic location of the component genes and the activity of the encoded enzymes are solved questions.

And yet, the combined action of known candidate genes explains only a fraction of the observed variation in THC:CBD ratios, and effect sizes vary across studies and between strains. *Cannabis* contains hundreds of different terpene and cannabinoid metabolites, so the failure of a small set of candidate genes to fully explain the phenotype of interest is not shocking. But the consequence of this incomplete understanding of the very traits that *Cannabis* is being marketed for is that there is no standardization of the product being offered to consumers and no predictability in the outcome of growers' efforts.

To avoid the biases inherent in a candidate-based approach, we generated amplicon sequences for several hundred loci in 300 strains of commercially grown *Cannabis* with a broad range of cannabinoid phenotypes. We combined these data with existing SRA libraries to examine almost 2,000 individual samples representing 500 different strains of *Cannabis*, each of which has reliable phenotyping data from Washington State testing labs. We have mapped these sequences to ten different reference genomes and are using a GWAS approach to examine genotype-phenotype relationships for several cannabinoid-related traits.

To date, we have identified many significant SNPs that do not coincide with any known candidate gene. Stated explicitly, most putatively causal SNPs do not map to the known CBDA synthase and THCA synthase genes. Instead, we find a variety of less-expected genes and non-coding regions implicated. We discuss the identity of the new candidate loci and suggest some possible reasons for the lack of association between genes in the CBD and THC production pathway with the phenotypic variation in THC:CBD ratio within and between cultivars of commercially available *Cannabis*.

METABOLIC AND BEHAVIORAL EFFECTS OF A MODERN HUMAN-SPECIFIC AMINO ACID SUBSTITUTION IN ADENYLOSUCCINATE LYASE

Xiangchun Ju¹, ShinYu Lee¹, Chika Azama¹, Tomomi Miyamoto¹, Agnieszka Kubik-Zahorodna², Ronald Naumann³, Victor Wiebe⁴, Jeanette Frommolt⁴, Rowina Voigtlaender⁴, Michael C Roy¹, Wulf Hevers⁴, Izumi Fukunaga⁵, Svante Pääbo^{1,4}

¹Okinawa Institute of Science and Technology, Human Evolutionary Genomics Unit, Onna-son, Japan, ²Czech Centre for Phenogenomics, Phenotyping Module, Vestec, Czech Republic, ³Max Planck Institute of Molecular Cell Biology and Genetics, Transgenic Core Facility, Dresden, Germany, ⁴Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, ⁵Okinawa Institute of Science and Technology, Sensory and Behavioural Neuroscience Unit, Onna-son, Japan

High-quality genome sequences from Neandertals and Denisovans, archaic hominins who shared a common ancestor with present-day humans about half a million years ago, allow genomic features that appeared and rose to high frequencies in modern humans since their divergence from a common ancestor shared with the archaic hominins to be identified. Among these are amino acid substitutions in the proteins ADSL, GLDC and SLTRK1. Pathogenic mutations in the genes encoding these proteins have numerous effects that include social and behavioral symptoms.

We introduced the modern human-specific substitutions in the corresponding mouse genes and analyzed the effects on dominance. In tube tests, where “humanized” and wild type mice enter a narrow tube from opposite ends and the mouse that pushes the other out is considered “dominant”, the mice carrying the “humanized” forms of the three genes were more dominant. We then analyzed each substitution individually in a system that allows several aspects of behavior in a more complex social environment to be automatically recorded (IntelliCageTM). The mice carrying substitutions in Gldc and Slitrk1 showed no dominance behavioral effects. In contrast, the mice carrying two adjacent substitutions in Adsl that change a mouse-specific amino acid to the common mammalian one and introduce the human-specific amino acid substitution tended to be more socially dominant in terms of gaining access to water when it is restricted and how other mice avoided water ports previously visited by the “humanized” mice.

Furthermore, the “humanized” mice with substitutions in ADSL displayed reduced purine de novo biosynthesis, a metabolic pathway in which ADSL functions is an essential catalytic enzyme. The reduction in purine de novo biosynthesis mimics metabolic differences seen between humans and apes as well as in human cells compared to cells “ancestralized” with respect to the amino acid in ADSL. These results suggest the modern human-specific amino acid change in ADSL may contribute to metabolic and behavioral differences between modern and archaic hominins.

UTILIZING TEMPORAL PROTEOMICS OF iPSC-DERIVED NEURONAL CELL STATES FOR STUDY OF DISEASE-SPECIFIC PATHWAYS IN MENTAL DISORDERS

Petra Páleníková^{1,2}, Greta Pintacuda^{1,2}, Yu-Han Hsu^{1,2}, Julia Biagini^{1,2}, Daya Mena^{1,2}, Joshua Ching^{1,2}, Travis Botts^{1,2}, Nadine Fornelos^{1,2}, Kasper Lage^{1,2}

¹Broad Institute of MIT and Harvard, Stanley Center for Psychiatric Research, Cambridge, MA, ²Broad Institute of MIT and Harvard, Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Cambridge, MA

Mental disorders, such as schizophrenia (SCZ), autism spectrum disorders (ASDs) or severe developmental disorders (DD), are complex and heterogeneous conditions with severe impact on the quality of a person's life. Recent genetic studies identified high-confidence protein-coding risk genes that confer substantial risk for SCZ, ASDs or DD. However, these studies provide limited insight into the function of risk genes in disease pathogenesis, and importantly, they do not inform about the functional convergence of risk genes or their relationships. To address this problem, we studied risk gene co-expression patterns throughout neuronal differentiation. We analyzed whole cell proteomes of six differentiating cell states, spanning iPSCs, neural progenitors (NPC) and neurons (iN). We evaluated common expression patterns between all detected proteins by k-means clustering and identified clusters of proteins with peak abundances in iPSCs, NPCs and iNs, respectively. Proteins conferring ASD-predominant risk were enriched in the iN specific cluster and DD-predominant proteins in the NPC cluster, supporting the importance of temporal proteomics in pinpointing disease-relevant time points. However, the clusters identified exclusively based on protein expression contained large numbers of proteins, ranging from 379 to 2874, likely encompassing both disease-specific and general cell maintenance pathways. Therefore, the integration of proteomic data with orthogonal datasets seems necessary to identify biological pathways specifically involved in disease. To this end, we compared proteomic clusters to protein interaction partners of SCZ risk genes identified by IP-MS in the same cell states. We observed enrichment of iN specific interactors in the iN specific cluster, highlighting the value of proteomic data in interpreting cell state specific protein-protein interactions (PPIs). To gain insights into disease using this approach, we analyzed whole proteomes of iPSC derived neurons containing the 22q11.2 deletion, which has been linked to SCZ and NDDs. This allowed us to investigate how this deletion perturbs the composition of PPI networks, and thereby to infer how the networks can be dysregulated in disease. In the future, we plan to integrate whole cell and PPI proteomic datasets with genetic data to prioritize pathways for follow-up functional investigation.

THE IMPACT OF REPAIR CONTEXT ON MUTATIONS GENERATED BY CAS9

Ananth Pallaseni¹, Elin Madli Peets¹, Özdemirhan Serçin², Balca Mardin², Michael Kosicki³, Leopold Parts^{1,4}

¹Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom,
²BioMedX, Oncology, Heidelberg, Germany, ³Lawrence Berkeley National Laboratory, Biosciences, Berkeley, CA, ⁴University of Tartu, Computer Science, Tartu, Estonia

The repair of double stranded breaks is dependent on the combined action of several repair genes whose exact function and behaviours in the repair process are not well characterised. Here we measure how the absence of repair genes changes the mutational profile of Cas9-generated cuts across a variety of targeted regions. We performed self-targeting CRISPR-Cas9 screens in 21 mouse embryonic stem cell lines with single repair genes knocked out, and measured the mutational profiles generated at 5,760 loci. We find that knockouts modulate mutation profiles at each locus in a structured manner, consistent with the action of the repair pathway they belong to. We catalogue that certain classes of mutation were preferentially generated in certain knockouts, such as complex combined insertion and deletions in the absence of Xrcc6. Knockouts were also found to divergently modulate sets of superficially similar mutations, suggesting that they were produced under multiple mechanisms. We use knowledge of this modulation to build predictive models of Cas9 repair outcomes in each of these knockout backgrounds and find that they perform as well as replicate measurements. This work improves our understanding of repair gene function and the determinants of double-strand break repair.

CLONALLY SELECTED LINES AFTER CRISPR/CAS EDITING ARE NOT ISOGENIC

Arijit Panda¹, Milovan Suvakov¹, Jessica Mariani^{3,4}, Kristen L Drucker², Yohan Park⁵, Yeongjun Jang¹, Thomas M Kollmeyer², Gobinda Sarkar², Taejeong Bae¹, Jean J Kim⁶, Wan Hee Yoon⁵, Robert B Jenkins², Flora M Vaccarino^{3,4}, Alexej Abyzov¹

¹Mayo Clinic, Quantitative Health Sciences, Rochester, MN, ²Mayo Clinic, Lab Medicine and Pathology, Rochester, MN, ³Yale University, Child Study Center, New Haven, CT, ⁴Yale University, Neuroscience, New Haven, CT, ⁵Oklahoma Medical Research Foundation, Aging and Metabolism Research Program, Oklahoma, OK, ⁶Baylor College of Medicine, Molecular & Cellular Biology, Houston, TX

The CRISPR-Cas9 system has enabled researchers to precisely modify/edit the sequence of a genome. A typical editing experiment consists of two steps: (i) editing cultured cells; (ii) cell cloning and selection of clones with and without intended edit, presumed to be isogenic. The application of CRISPR-Cas9 system may result in off-target edits, while cloning will reveal culture-acquired mutations. We analyzed the extent of the former and the latter by whole genome sequencing in three experiments involving separate genomic loci and conducted by three independent laboratories. In all experiments we hardly found any off-target edits, while detecting hundreds to thousands of single nucleotide mutations unique to each clone after relatively short culture of 10-20 passages. Notably, clones also differed in copy number alterations that were several kb to several mb in size and represented the largest source of genomic divergence among clones. We suggest that screening of clones for mutations and copy number alterations acquired in culture is a necessary step to allow correct interpretation of DNA editing experiments. Furthermore, since culture associated mutations are inevitable, we propose that experiments involving derivation of clonal lines should compare a mix of multiple unedited lines and a mix of multiple edited lines.

SYSTEMATIC IDENTIFICATION OF SILENCERS IN THE HUMAN GENOME

Baoxu Pang¹, Michael Snyder²

¹Leiden University Medical Center, Cell and Chemical Biology, Leiden, Netherlands, ²Stanford University, Genetics, Stanford, CA

The majority of the human genome contains non-coding sequences that do not encode proteins. Many of these non-coding regions are expected to contain important regulatory sequences that control gene expression. To date, most studies have focused on activators such as enhancers, but regions that repress gene expression—silencers—have not been systematically studied. We developed a system that identifies silencer regions in a genome-wide fashion. We found that silencers are widely distributed across the human genome and often function in a tissue specific fashion. These silencers harbor unique, tissue-specific epigenetic signatures and are recognized by different transcription factors. Deletion of silencer regions linked to the drug transporter genes ABCC2 and ABCG2, caused up-regulation of these drug transporters and chemo-resistance. 3D chromosome conformation analysis suggests that silencers act at multiple genes, and at the level of chromosomal domains and long-range interactions. Overall, our study demonstrates that tissue-specific silencing is widespread throughout the human genome and likely contributes significantly to the regulation of gene expression.

ALTERNATIVE SPLICING OF RPS24 GENE IS A PROGNOSTIC BIOMARKER IN KIDNEY RENAL CLEAR CELL CARCINOMA

Jiyeon Park^{1,2,3}, Yeun-Jun Chung^{1,2,4}

¹The Catholic University of Korea, Precision Medicine Research Center, College of Medicine, Seoul, South Korea, ²The Catholic University of Korea, Integrated Research Center for Genome Polymorphism, Seoul, South Korea, ³The Catholic University of Korea, Department of Biomedicine & Health Sciences, Graduate School, Seoul, South Korea,

⁴The Catholic University of Korea, Department of Microbiology, College of Medicine, Seoul, South Korea

Ribosomes have long been considered homogenous cellular organelles that synthesize proteins, but recent studies have reported heterogeneous composition of ribosomal proteins across tissues and diseases. Alternative splicing (AS) is a posttranscriptional regulation that allows a single cell to encode multiple transcripts. Recent reports indicate that AS is responsible for generating diverse transcripts of the RPS24 gene, which encodes a ribosomal protein. However, detailed observation is not easy because of several short-length exons that are targets of AS. Therefore, we conducted a comprehensive analysis of splicing junction reads in the RPS24 gene using large-scale omics data, and provided a detailed view of tissue-specific and cancer-specific regulation. From a pan-cancer perspective of RPS24 AS, we highlighted the potential of RPS24 AS as a valuable molecular marker in kidney clear cell renal carcinoma (KIRC). The cancer-related AS changes were validated in five KIRC datasets generated from independent experiments. To elucidate the cellular source of the cancer-specific changes, we analyzed immune cells in KIRC and found that RPS24 AS reflects the level of infiltrated immune cells, specifically macrophages. Taken together, these results suggest that RPS24 AS could serve as a prognostic marker for KIRC and provide information about the immune environment.

EXPANDING CANCER THERAPY OPTIONS THROUGH GENOME-SCALE IDENTIFICATION OF SYNTHETIC LETHAL PARALOG PAIRS

Phoebe C Parrish^{1,2,3}, Austin M Gabel^{2,3,4}, Daniel Groso^{1,2}, Shriya Kamlapurkar^{1,2}, James D Thomas^{2,4}, Robert K Bradley^{2,3,4}, Alice H Berger^{1,2,3}

¹Fred Hutchinson Cancer Center, Human Biology Division, Seattle, WA, ²Fred Hutchinson Cancer Center, Computational Biology Program, Seattle, WA,

³University of Washington, Department of Genome Sciences, Seattle, WA,

⁴Fred Hutchinson Cancer Center, Basic Sciences Division, Seattle, WA

Synthetic lethal therapies have the potential to expand therapeutic options for cancer patients. Synthetic lethality occurs when dual inactivation of a gene pair leads to cell death but single gene inactivation has little effect on viability.

When one synthetic lethal gene is inactivated in cancer, targeting the remaining gene with an inhibitor will result in tumor cell death while normal cells remain viable. One successful synthetic lethal therapy is the use of *PARP* inhibitors in *BRCA1/2*-mutant breast and ovarian cancer. To extend the benefits of synthetic lethal therapies to other cancer types, we developed a dual knockout CRISPR method and analysis pipeline to systematically identify synthetic lethal human gene pairs.

Since genetic interactions are rare we chose to target paralogs, duplicated genes that frequently exhibit functional redundancy. To find synthetic lethal paralogs that could serve as lung cancer drug targets we developed paired guide RNAs for paralog genetic interaction mapping (pgPEN). pgPEN is a pooled dual-targeting CRISPR-Cas9 library that targets over 2,000 paralogs, a quarter of which are druggable. We applied pgPEN to cancer cell lines derived from lung and cervical tissue and found that 12% of paralogs exhibited synthetic lethality in at least one context.

Additionally, we have developed two software packages to enable identification of genetic interactions in human CRISPR screening data. The first package is a reproducible Snakemake pipeline that allows users to analyze data from our custom pgPEN sequencing method, facilitating the application of our approach to other model systems. The second package enables interactive analysis of dual-targeting CRISPR screen data across multiple cell lines and drug treatment conditions. Specifically, this software pipeline will empower users with a limited coding background to perform quality control analyses, quantify the growth effects of genetic perturbations and drug treatments, and identify buffering and synthetic lethal genetic interactions.

This work reveals synthetic lethal paralog interactions in human cancer cells that can be further tested and translated to the clinic. Additionally, the pgPEN CRISPR library and analysis pipeline will enable other researchers to identify targetable synthetic lethal interactions in their model systems of interest.

Paralog synthetic lethal therapy provides a relatively low-toxicity therapeutic approach that can expand targeted therapy options and improve patient outcomes.

RECONSTRUCTING THE *CIS*-REGULATORY LANDSCAPE OF ARCHAIC HOMINIDS USING DEEP LEARNING

Aman Patel¹, Georgi Marinov², Anshul Kundaje^{1,2}

¹Stanford University, Computer Science, Stanford, CA, ²Stanford University, Genetics, Stanford, CA

The ability to sequence genomes from samples dating back many thousands of years has transformed our understanding of human evolution. However, the wealth of information contained in these sequences has not yet been fully utilized, as the impact of the variation observed in these genomes relative to that in modern humans is still poorly understood. In particular, despite the crucial role of *cis*-regulatory elements in the regulation of gene expression, variants associated with these regions have so far received very little attention. In this study, we utilized deep learning to examine the functional effects of over 50,000 variants unique to Neanderthals and Denisovans, the two archaic hominid species most closely related to humans. Through *in silico* mutagenesis experiments, we applied over 1,500 cell type-specific deep learning models, designed to predict chromatin accessibility from sequence, to quantify and score the cell type-specific regulatory effects of each variant.

When analyzing the set of genes located nearby to the highest-scoring variants, we detected enrichment for several highly relevant body systems, including the nervous system (in particular, brain development), metabolism, and skeletal morphology. We also performed hierarchical clustering on all cross-cell type variant effect vectors, allowing us to locate clusters of variants with heightened activity in brain tissue, immune cells, and multiple other important cell types. These findings align with known differences between archaic and modern humans as determined by previous studies and the skeletal record.

Through state-of-the-art feature attribution methods, we identified the binding motifs containing notable variants, and we observed several examples of motif creation and disruption near relevant genes. In one example, an archaic allele close to two important brain-related genes drives a predicted accessibility peak through the creation of a Leucine zipper motif, thus indicating a possible role for this locus in nervous system regulation in archaic humans. We also observe multiple striking cases of synergistic variant effects, where predicted accessibility changes are observed if a string of archaic variants in close proximity are substituted together but not detected if the variants are substituted individually.

Overall, this study has quantified regulatory differences between humans and our closest neighbors, provided a comprehensive dataset of cell type-specific archaic variant effects, and demonstrated the utility of deep learning in the study of archaic genomics and epigenomics.

CATE: AN ACCELERATED AND SCALABLE SOLUTION FOR LARGE-SCALE GENOMIC DATA PROCESSING THROUGH GPU AND CPU-BASED PARALLELIZATION

Deshan Perera¹, Elsa Reisenhofer¹, Said Hussain¹, Eve Higgins¹, Christian D Huber², Quan Long^{1,3}

¹University of Calgary, Department of Biochemistry & Molecular Biology, Cumming School of Medicine, Calgary, Canada, ²The Pennsylvania State University, Department of Biology, Pennsylvania, PA, ³University of Calgary, Hotchkiss Brain Institute, Department of Medical Genetics, Department of Mathematics and Statistics, Alberta Children's Hospital Research Institute, Calgary, Canada

BACKGROUND: The selection pressures that govern the evolution of a genome are quantified by statistical tests on evolution. The power of these statistical tests is strengthened by increased sample sizes. However, this results in taxation on a computer's processing capabilities resulting in extended processing times.

Parallelization of these statistical tests can reduce computational time. The Graphical Processing Unit (GPU) housing thousands of cores enables the potential of large-scale parallelization. NVIDIA's CUDA-based GPUs are becoming commonplace in both general purpose and specialized computing. Presently, the use of parallel processing to solve genetic algorithms with the aim of reducing computation time has gained traction. So far, such potential of high parallelization has not been realized in molecular evolution analyses.

METHODS: CATE (CUDA Accelerated Testing of Evolution) is such a software solution. It is a scalable program built using NVIDIA's Compute Unified Device Architecture (CUDA) platform together with an exclusive file hierarchy to process six different tests frequently used in molecular evolution, namely: Tajima's D, Fu and Li's D, D*, F and F*, Fay and Wu's H and E, McDonald–Kreitman test, Fixation Index, and Extended Haplotype Homozygosity.

CATE is composed of two main innovations. A file organization system coupled with a novel multithreaded search algorithm and the large-scale parallelization of the algorithms using the GPU and CPU.

First, a single VCF file can contain over a few million polymorphisms spanning thousands of individuals making them time-consuming to read and process sequentially. Therefore, we have split these files based on the number of SNPs and positions. This enables random accessing of the genomic data. To navigate this file structure, we have designed our "Compound Interpolation Search" algorithm. It is a synergistic implementation of the interpolated search algorithm coupled with a multithreaded form of sequential search. This greatly increases the speed with which information is collected while minimizing the demand for RAM. Second, we developed bespoke algorithms for each test. These algorithms are designed to harness the full potential of the CUDA core architecture to better process the polymorphic data and conduct the required computations.

RESULTS: Powered by these implementations CATE is magnitudes faster than standard tools, being on average over 180 times faster. For instance, CATE processes all 54,849 human genes for all 22 autosomal chromosomes across the five super populations present in the 1000 Genomes Project in less than thirty minutes while counterpart software took 3.62 days. This proven framework has the potential to be adapted for GPU-accelerated large-scale parallel computations of many evolutionary and genomic analyses.

GitHub repository: <https://github.com/theLongLab/CATE>

THE EPIGENETIC LOGIC OF GENE ACTIVATION

Beatrice Borsari¹, Amaya Abad¹, Cecilia C Klein¹, Ramil Nurtdinov¹, Vasilis F Ntasis¹, Alexandre Esteban¹, Emilio Palumbo¹, Marina Ruiz-Romero¹, Raül G Veiga¹, Maria Sanz¹, Bruna R Correa¹, Rory Johnson¹, Sílvia Pérez-Lluch¹, Roderic Guigó^{1,2}

¹Centre de Regulació Genòmica, CRG, Barcelona, Spain, ²Universitat Pompeu Fabra, UPF, Barcelona, Spain

Histone modifications are widely accepted to play a causal role in the regulation of gene expression. This role has been recently challenged by reports showing that gene expression can occur in the absence of histone modifications. To address this controversy, we have generated densely-spaced transcriptomic and epigenomic maps during the transdifferentiation of human pre-B cells into macrophages, a time-course cell homogeneous process that occurs with massive transcriptional changes. We found that the relationship between histone modifications and gene expression is weaker than previously reported, and that it can even run contrary to established assumptions, such as for H3K9me3. Our data suggest a model that reconciles the seemingly contradictory observations in the field: histone modifications are associated with expression only at the time of initial gene activation. Further changes in gene expression, even larger than those occurring at gene activation, are essentially uncoupled from changes in histone modifications. Our dense time-course has allowed us to decipher the general order in which histone modifications are deposited upon gene activation, revealing that only H3K4me1/me2 precede gene expression at promoter level. These results are supported by experimental validations performed for a subset of genes specifically activated during transdifferentiation. In addition, this precise order of the deposition of histone marks is not specific to our human transdifferentiation system: data available during mouse development largely recapitulate this model, albeit with less resolution. We also observed that marking at promoters generally precedes marking at enhancers, suggesting that histone marks at these regions may partially be a consequence of the deposition of histone modifications at promoters. Overall, our work provides a sketch of the epigenetic logic underlying gene activation in eukaryotic cells.

SLOWER B CELL TRANSDIFFERENTIATION IN HUMAN THAN IN MOUSE IS THE BY-PRODUCT OF THE HUMAN SPECIFIC ALU-REPEAT EXPANSION

Ramil Nurtdinov¹, Maria Sanz¹, Amaya Abad¹, Carme Arnan¹, Alexandre Esteban¹, Sebastian Ullrich¹, Rory Johnson¹, Sílvia Pérez-Lluch¹, Roderic Guigó^{1,2}

¹Centre de Regulació Genòmica, CRG, Barcelona, Spain, ²Universitat Pompeu Fabra, UPF, Barcelona, Spain

Many physiological processes occur at a faster pace in mouse than in human, even though mammalian cells use similar mechanisms to regulate growth and differentiation. Differences in duration of physiological processes may, in turn, underline the different longevity of these species. We have specifically focused on the transdifferentiation from B cells to macrophages. This is a unique, biologically simple, paired cell system that occurs with massive transcriptional changes. It is artificially induced by CEBPA, but it takes seven days in human and three in mouse. We collected a time-course of densely-spaced, high-depth RNA-Seq and CEBPA ChIP-Seq during human and mouse transdifferentiation. We found that CEBPA genomic binding is weaker in human compared to mouse along transdifferentiation. Reporter analyses using the promoter of the *MYC* gene, known to impact the efficiency of transdifferentiation in mouse, showed that mouse cells are capable of binding CEBPA to weak binding sites while human cells are not, indicating that mouse cells recruit CEBPA in a more efficient manner than human ones. Weak CEPBA binding in human is not specific of *MYC* during transdifferentiation, but a constitutive trait of the human genome across biological conditions. We traced back weak CEBPA binding in the human genome to the primate-specific Alu repeat expansion. A large number of Alu sequences carry strong instances of the CEBPA binding motif, they compete with CEBPA targets for CEBPA binding and, as a result, CEBPA binding is overall attenuated in the human genome. Blocking CEBPA binding in thousands of Alu repeats carrying CEBPA binding motifs, using the CRISPR-dCas9 system, results in increased CEBPA binding to its canonical binding sites. Global attenuation of CEBPA contributes, in turn, to the longer duration of the transdifferentiation process under the control of this factor in human. Consistently, increasing the CEBPA dose in human cells promotes the acceleration of the transdifferentiation in a dose-dependent manner, while it has almost null effect in mouse. While the link between the duration of biological processes and longevity is unclear, we found that Alu repeats containing strong CEBPA instances have accumulated in apes compared to other primates and this accumulation, therefore, is associated to ape specific phenotypes, such as longevity. Our work highlights the highly complex mode in which biological information is encoded in genome sequences, evolutionarily connecting lineage-specific transposable element expansions to species-specific developmental tempos.

CRISPR-CLEAR - IN-SITU INVESTIGATION OF GENOTYPE-TO-PHENOTYPE RELATIONSHIP WITH NUCLEOTIDE LEVEL RESOLUTION CRISPR SATURATION MUTAGENESIS SCREENS

Becerra Basheer^{1,2,3,4}, Martin Jankowiak², Sandra Wittibschlager⁵, Anzhelika Karjalainen⁵, Ana Patricia Kutschat⁵, Ting Wu^{2,3}, Marlena Starrs^{2,3}, Zain Patel^{1,2,4}, Daniel Bauer^{2,3,4}, Davide Seruggia⁵, Luca Pinello^{1,2,4}

¹MGH, Pathology, Boston, MA, ²Harvard Medical School, Pathology, Boston, MA, ³Boston Children's Hospital, Pediatrics, Boston, MA,
⁴BROAD, Boston, MA, ⁵CCRI and CeMM, Vienna, Austria

In this study, we present the development of a genome editing tool called CRISPR-CLEAR for the in-situ investigation of genotype-phenotype relationships at nucleotide and variant-level resolution. This tool combines CRISPR-Cas9 technologies with novel computational strategies to enable high-resolution interrogation of sequence variants, to ultimately link phenotypes to their causal mutations.

As a proof of concept, we investigated a regulatory element upstream of CD19 using NALM-6 cells, a model of B-cell leukemia. We designed a library of 200 sgRNAs that tiled a candidate enhancer and performed four screens using nucleases and base editors with extended PAM recognition. We used a fluorescently conjugated antibody to stain cells and sort them into CD19+ or CD19- populations according to their CD19 expression levels. For each sample, we sequenced both the sgRNA (for perturbation counts) and the endogenous targeted region (for direct allele readout) in tandem to compare each approach. By comparing the sgRNA count or alleles in both sorted populations, we are able to identify functional sub-regions relevant for CD19 regulation.

To examine the relationship between the genotype and phenotype at nucleotide and variant level resolution based on the target locus sequencing, we developed a Bayesian linear regression model based on a recent statistical method called millipede, which assigns importance scores to each mutation based on its abundance in either sorted population. Our model shows superior signal detection and functional resolution as compared to current count-based methods that only assess sgRNA enrichment. Using our approach, we uncovered and validated a hotspot of mutations at a 20 bp region corresponding to relevant transcription factor binding sites such as PAX5, which were consistent across biological replicates and perturbation used.

In conclusion, CRISPR-CLEAR offers a powerful tool for investigating genotype-phenotype relationships by directly observing edits at the endogenous locus with higher resolution compared to count-based methods. We envision that this framework will provide a comprehensive solution for the classification of non-coding variants of uncertain significance and the discovery of causal regulatory elements.

AN ACCESSIBLE CHROMATIN ATLAS OF THE HUMAN INTESTINE IDENTIFIES DISEASE SPECIFIC REGULATORY FEATURES AND DELINEATES THE IMPACT OF NON-CODING VARIANTS IN CROHN'S DISEASE

Yu Zhao^{*3}, Ran Zhou^{*1}, Bingqing Xie¹, Candace M Cham¹, Jason Koval¹,
Xin He², Eugene B Chang¹, Anindita Basu¹, Sebastian Pott²

¹University of Chicago, Department of Medicine, Chicago, IL, ²University of Chicago, Department of Human Genetics, Chicago, IL, ³University of Chicago, Pritzker School of Molecular Engineering, Chicago, IL

Crohn's disease (CD) is a chronic inflammatory bowel disease (IBD) that arises from complex interplay between genetic susceptibility, microbial, and environmental factors. Genome-wide association studies (GWAS) in IBD have identified over 200 susceptibility loci, many of which point to non-coding risk variants possibly affecting gene regulatory mechanisms.

However, the lack of cell-type-resolved regulatory features in the gut makes it difficult to understand the contribution of epigenetic mechanisms to CD and to predict the functional consequences of such non-coding variants. To address this roadblock, we generated single-cell chromatin accessibility profiles (scATAC-seq) of 146,595 cells taken from 24 CD-affected and 28 healthy biopsy samples collected from terminal ileum or ascending colon. We identified over 500,000 accessible regions across epithelial, immune and stromal lineages in both ileum and colon, providing the first cell-type-specific, CD-associated atlas of cis-regulatory elements (CREs) and their candidate target genes. Multi-omic analysis revealed cell-type-specific transcription factors, including GATA4 in enterocytes of CD patients. Furthermore, we nominated putative non-coding CD risk variants associated with lineage specific and shared CREs using this atlas. This single-cell accessible chromatin atlas provides a critical step towards interpreting non-coding genetic and epigenetic mechanisms in the etiology of IBD, offering insights into the pathogenesis of this complex disorder.

IT TAKES TWO (STRANDS) TO MAKE A THING GO RIGHT. RIGHT?

Aaron Quinlan^{1,2,3}, Michael Goldberg¹, Brent Pedersen¹, Jason Kunisaki¹, Suchita Lulla¹, Laurel Hiatt¹

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

Illumina DNA sequencing errs roughly once per thousand nucleotides. This accuracy is ideal for the detection of inherited genetic variation. However, such a seemingly low error rate causes a calamitous signal-to-noise problem when studying rare mutations, as mutations arise 1,000 to 1,000,000 times less frequently than errors, depending on the cell type and disease context.

"Duplex" DNA sequencing is heralded as a much more accurate solution. "Duplex-seq" begins by barcoding the two strands of a double-stranded DNA molecule; in turn, PCR amplicons of each strand can be traced back to both the original ssDNA and dsDNA molecule. True mutations emerge as alleles that are corroborated by reads from each strand of the original dsDNA. In contrast, damage, PCR errors, and sequencing errors can frequently be detected as they typically appear in only a subset of the reads originating from a single molecule.

The theoretical error rate of duplex-seq is between 1×10^{-7} and 1×10^{-9} , making it an ideal substrate for rare mutation detection. So, what is the actual error rate in duplex sequencing? No empirical measurements have been published. Also, current bioinformatic workflows do not integrate well-calibrated quality scores to differentiate true mutations from jackpot PCR events or DNA damage arising during library preparation.

We've applied deep (>10,000 fold coverage) duplex sequencing to longitudinal bulk sperm and blood DNA from a cohort of sperm donors. From this extensive dataset, we have created empirical and deep-learning models that estimate errors arising in duplex-seq. While the error rates we measure largely align with theoretical estimates, there are substantial, yet unaddressed caveats. Transversions are orders of magnitude less erroneous than transitions. Furthermore, both sequence context and read position correlate with substantial differences in error rates. Based on these findings, we propose new tools to provide quality scores for duplex sequencing that will improve its utility in diverse studies of genome mutation.

PLASMA PROTEOMIC DETERMINANTS OF COMMON CAUSES OF MORTALITY

Anurag Sethi, Anil Raj, Kevin Wright, Eugene Melamud

Calico Life Sciences LLC, South San Francisco, CA

In complex physiological systems, causal associations between protein concentration and outcomes can be difficult to determine, as proteins can be both effectors of outcomes and responders. This problem becomes especially difficult when the outcome of interest is death because this is often preceded by the failure of multiple physiological systems. Our aim in this study was to determine the extent to which variation in plasma protein abundance can predict all-cause and disease-specific mortality, and to assess the potential causal nature of any observed associations. To identify biomarkers for mortality, we analyzed the largest human plasma proteome dataset compiled to date: 1,459 proteins measured in approximately 54,306 individuals with over 12 years of prospective follow-up data from the UK Biobank (during which 5,032 mortality events occurred). We carried out a multivariate stability selection analysis of all biomarkers and identified 14 proteins that were robustly predictive of all-cause mortality (stability score greater than 0.9). Additionally, we identified 14 proteins that were robustly predictive of cause-specific mortality (stability score greater than 0.9), across six disease categories with the highest rates of mortality. These 28 biomarkers remained significant after accounting for pre-existing conditions and medication usage. Participants in the highest quintile of all-cause mortality risk prediction had ~25 times higher number of death events as compared to participants in the lowest quintile. Moreover, these proteins were more robustly associated with mortality than established diagnostic biomarkers such as glycated hemoglobin (HbA1c) and cholesterol. To complement our epidemiological analyses, we quantified the genetic determinants of plasma protein abundance and identified a median of 14 GWAS loci per protein. We assessed causality via Mendelian randomization and mediation analysis using polygenic scores computed on the entire UK Biobank cohort. We identified 6 biomarkers with putative sex-specific causal effects on all-cause mortality and 7 additional biomarkers with effects on cause-specific mortality. In conclusion, by integrating multiple lines of genetic and epidemiological evidence, we identified plasma protein biomarkers causally associated with human lifespan.

UNCOVERING HIGHER-ORDER MOTIF INTERACTIONS WITH BIOLOGICALLY INTERPRETABLE NEURAL NETWORKS

Chandana Rajesh, Rohan Ghotra, Steven Yu, Peter Koo

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

Deep neural networks (DNNs) have proven to be effective tools for biological sequence analysis, taking DNA sequences as input and predicting a variety of molecular functions. However, their high degree of overparameterization makes them difficult to interpret, impeding efforts to uncover the syntax of *cis*-regulatory element (CRE) motifs. Post-hoc interpretability methods, most commonly attribution maps, are typically used to gain insights into the importance of each nucleotide in a given sequence. However, it remains difficult to decipher motifs from spurious importance scores when interpreting attribution maps. Moreover, attribution maps fundamentally do not inform higher-order interactions within and across motifs, key features that are central to gene regulation. Hence, it may be beneficial to develop an alternative approach to model interpretability through the principled design of inherently interpretable architectures, wherein the learned parameters have direct biological significance. Previous research has mainly focused on design considerations where first-layer filters capture motif representations. Recently, multihead self-attention (MHA) has been shown to build upon the motifs learned in the first layer to capture spatially distant motif interactions. However, as convolutional kernels model DNA sequences as an additive effects model, the MHA is tasked with learning both intra and inter-motif interactions. The extent to which design considerations affect the efficacy of interpreting MHA layers is not clear. Here, we introduce a novel network architecture, HOMINID (Higher-Order Motif INteractions with Interpretable Deep neural networks), which incorporates higher-order convolutional kernels in the first convolutional layer to fully capture motif binding sites, followed by MHA to learn interactions between motifs and with sequence context across the entire input sequence. We perform a systematic evaluation of HOMINID to capture motif interactions across a variety of prominent regulatory genomics prediction tasks which encompass varying levels of complexity. Our results demonstrate that HOMINID can facilitate the biological interpretation of DNNs, as direct visualization of the parameters, revealing motifs and their interaction partners, obviating the need for post-hoc analysis of more opaque models.

THE LANDSCAPE OF TRANSCRIPTOMIC AND EPIGENETIC VARIATION ACROSS HUMAN TRAITS

Raquel García-Pérez¹, Jose Miguel Ramirez¹, Aida Ripoll-Cladellas¹, Ruben Chazarra-Gil¹, Winona Oliveros¹, Oleksandra Soldatkina¹, Mattia Bosio¹, Paul Joris Rognon^{2,3}, Salvador Capella¹, Miquel Calvo⁴, Ferran Reverter⁴, Roderic Guigó⁵, François Aguet^{6,7}, Pedro G Ferreira^{8,9}, Kristin G Ardlie⁶, Marta Mele^{1,6}

¹Barcelona Supercomputing Center, Life Sciences, Barcelona, Spain,

²Universitat Pompeu Fabra, Economics and Business, Barcelona, Spain,

³Universitat Politècnica de Catalunya, Statistics and Operations Research, Barcelona, Spain, ⁴Universitat de Barcelona, Statistics, Barcelona, Spain,

⁵Center for Genomic Regulation, Bioinformatics and Genomics, Barcelona, Spain, ⁶Harvard and MIT, Broad Institute, Boston, MA, ⁷Illumina, Artificial Intelligence Laboratory, San Diego, CA, ⁸University of Porto, Computer Science, Porto, Portugal, ⁹INESC TEC, Artificial Intelligence and Decision SupportPorto, Portugal

Understanding the consequences of individual transcriptome and epigenome variation is fundamental to deciphering human biology and disease.

Demographic traits such as ancestry, sex, age and BMI, and clinical traits such as diabetes, simultaneously affect gene expression, alternative splicing and DNA methylation variation. However, how these variables

mechanistically interplay to ultimately define an individual's phenotype is not well understood. Here, we first implement a statistical framework to quantify the joint contribution of 4 demographic and 17 clinical traits as drivers of gene expression and alternative splicing variation across 46 human tissues and 781 individuals from the Genotype-Tissue Expression project. We demonstrate that demographic traits have additive and tissue-specific contributions to expression variability, but trait interactions are rare. Variation in splicing is dominated by ancestry and is under genetic control in most tissues, with ribosomal proteins showing many tissue-shared splicing events, raising the possibility that specific aspects of the ribosome machinery might differ between human populations across most tissues. We then expand this framework to study DNA methylation changes using Infinium EPIC array data on 9 GTEx tissues. We find that epigenetic variation is mainly driven by age, followed by ancestry and then sex, with few exceptions, consistent with the patterns observed in gene expression. Finally, our analyses reveal important contributions of clinical traits to tissue transcriptome variation. Type 1 and 2 diabetes affect multiple tissues, particularly the tibial nerve, where we further show that the nerve damage induced by diabetes shares similarities with that of biological ageing. Using machine learning techniques trained on histological images to classify diabetic tibial nerve samples, and, importantly, identify new genes related to diabetic neuropathy. Overall, our multi-tissue and multi-trait approach provides an extensive characterization of the main drivers of human transcriptome and epigenome variation in health and disease.

IDENTIFICATION OF FUNCTION VARIANTS ASSOCIATED WITH ADOLESCENT IDIOPATHIC SCOLIOSIS

Darius Ramkhalawan, Gloria Montoya-Vazquez, Kayla Ernst, Nadja Makki

University of Florida, Anatomy & Cell Biology, Gainesville, FL

Adolescent Idiopathic Scoliosis (AIS) is a progressive sideways curvature of the spine that manifests during the adolescent growth spurt in otherwise healthy patients. It is the most common pediatric musculoskeletal disorder, affecting ~3% of children worldwide. Currently, treatment of AIS is limited to restrictive bracing and corrective surgery; without treatment patients may develop chronic pain, restrictive lung disease, and severe deformity later in life. Genome-wide association studies (GWAS) have identified a number of AIS-susceptibility loci, however, the mechanisms by which these loci contribute to AIS pathogenesis remains unclear and few functional variants in linkage disequilibrium have been identified. The majority of these variants are located in non-coding genomic regions, indicating that they are likely affecting gene regulatory elements. By integrating several functional genomic data sets with GWAS data, we have identified several tissue-specific gene regulatory elements that are perturbed in AIS patients, the regulatory activity of which is dependent on transcription factor binding sites that are disrupted by AIS-associated single-nucleotide polymorphisms. These preliminary findings are a starting point to understanding the genetic and molecular mechanisms underlying this common childhood disease.

A TEST OF MOTHER'S CURSE WITH DEEP mtDNA DIVERGENCE AND OUTBRED NUCLEAR BACKGROUNDS IN DROSOPHILA

David M Rand, Faye A Lemieux, Kenneth Bradley, Lindsay Marmor

Brown University, Ecology, Evolution and Organismal Biology,
Providence, RI

Maternal inheritance allows selection to act on mtDNA-encoded effects in females, but prevents direct selection on mtDNA in males. Mutations that are deleterious in males but neutral or beneficial in females can persist in populations. This predicts that mtDNA-based disease or phenotypic variation should be more common in males, while haploid selection in females will purge mtDNA-based variation (Frank and Hurst (1996); repackaged as the 'Mother's Curse' by Gemmell et al. (2004)). There is conflicting evidence for this pattern in the literature. A key assumption in Mother's Curse is that mtDNA phenotypes must be sex limited with different effects, even different signs, in males and females. Extreme Mother's Curse scenarios invoke mtDNA mutations that are beneficial in females and deleterious in males and sweep through populations leading to extinction from male unfitness. Comparisons of sex-specific mtDNA phenotypic effects from different populations and species are needed to evaluate the evolutionary significance of Mother's Curse.

Most Mother's Curse analyses use alternative mtDNAs placed on one or more homozygous nuclear chromosomal backgrounds. Since most organisms are heterozygous at many loci, we sought to perform experiments in several different heterozygous backgrounds. MtDNAs from *Drosophila melanogaster* (OreR and Zimbabwe), *D. simulans* (siI and siII) and *D. yakuba* were each placed on a common *D. melanogaster* w1118 nuclear background. Virgin females from these strains were crossed to males from each of several deficiency stocks carrying a hemizygous segment of chromosome 2L. F1 female and male flies carrying the deficiency chromosome and the w1118 chromosomes were tested for starvation, climbing and flight performance. For all three traits in the majority of chromosomal backgrounds, the variance among mtDNA genotypes was greater in females than in males. This result is the opposite of the Mother's Curse prediction. Moreover, the impact of the foreign *D. yakuba* mtDNA was equally neutral or beneficial in both males and females, suggesting some form of phylogenetic heterosis. The mitonuclear epistatic interactions across the different heterozygous backgrounds and the five mtDNA haplotypes are more pronounced in females than males. This suggests that mtDNA interactions with regional hemizygosity or dominance effects are more pronounced in females than males, overshadowing any effect of Mother's Curse or even hidden Y chromosome variation.

SINGLE-CELL PROFILING OF TRANSCRIPTIONAL CHANGES ASSOCIATED WITH NEIGHBORHOOD STRESS IN IMMUNE CELLS OF ADOLESCENTS WITH ASTHMA

A. Ranjbaran¹, J. Wei¹, J. Resztak¹, S. Nirmalan¹, A. Alazizi¹, H. Mair-Meijers¹, J. Bruinsma², X. Wen⁴, R. Slatcher³, S. Zilioli², R. Pique-Regi¹, F. Luca^{1,5}

¹Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI,

²Wayne State University, Department of Psychology, Detroit, MI,

³University of Georgia, Department of Psychology, Athens, GA, ⁴University of Michigan, Department of Biostatistics, Ann Arbor, MI, ⁵University of Tor Vergata, Rome, Italy

Neighborhood stress is defined as the exposure of residents of a community to frequent violence and disorder. Exposure to community violence is associated with increased wheezing, hospitalization, and morbidity in children with asthma. Adverse community environments are associated with elevated proinflammatory markers in blood. Previous research has demonstrated the impact of adverse psychosocial experiences on asthma-associated gene expression in children. However, we still lack a precise understanding of the impact of these adverse experiences on specific immune cell types. The present study considers parental-reported neighborhood stress for 200 children with asthma in Detroit. scRNA-seq of peripheral blood mononuclear cells from these patients was performed to identify if exposure to neighborhood stress is associated with cell-type specific gene expression changes. Our preliminary results on a subset of 56 patients identified changes in the expression of 136 genes, 81 in B-cells, 38 in natural killer cells, 14 in T-cells, and 3 in monocytes (FDR<10%). High levels of neighborhood stress were associated with increased expression of 81 genes and decreased expression of 54 genes. These differentially expressed genes are enriched for lymphocyte-mediated positive regulation of innate immune response pathways, including IFN- γ signaling. IFN- γ signaling plays a regulatory role in asthma pathophysiology by suppressing Th2 response and its cytokines, IL-13 and IL-5. In T cells, we found a positive correlation between gene expression changes associated with neighborhood stress and those associated with IL-13/IL-5 levels. In contrast, this correlation is negative between neighborhood stress-associated expression changes and those associated with IFN- γ . This suggests that exposure to high neighborhood stress leads to the downregulation of genes involved in IFN- γ -mediated suppression of Th2 cells, while it upregulates genes involved in IL-13/IL-5 signaling. Three genes associated with asthma risk in Transcriptome Wide Association Studies were also differentially expressed in children exposed to neighborhood stress. In particular, CD52, which is involved in immune cell infiltration and anti-adhesion, was upregulated in NK cells from children exposed to high levels of neighborhood stress. Our preliminary results demonstrate that exposure to neighborhood stress is associated with cell-type specific gene expression changes in blood lymphocytes, and serves as a potential risk factor for asthma in children.

DIRECT LONG-READ RNA SEQUENCING UNCOVERS FUNCTIONAL GENETIC VARIATION AFFECTING TRANSCRIPTS EXPRESSION.

Aline Réal^{1,2,3}, Andrew Brown⁴, Gisella Puga Yung², Christelle Borel¹,
Nikolaos Lykoskoufis¹, Jörg D Seebach², Emmanouil T Dermitzakis¹, Anna
Ramisch^{1,5}, Ana Viñuela⁶

¹University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva, Switzerland, ²University Hospitals and Medical Faculty, Division of Immunology and Allergology, Geneva, Switzerland,

³New York Genome Center, NYGC, New York, NY, ⁴University of Dundee, Population Health and Genomics, Dundee, United Kingdom,

⁵University of Geneva Medical School, Department of Basic Neuroscience, Geneva, Switzerland, ⁶Newcastle University, Biosciences Institute, International Centre for Life, Newcastle upon Tyne, United Kingdom

Our knowledge of the impact of genetic variants on gene expression is limited by the short-read RNA-seq technologies currently used, as these do not characterize transcripts in their full-length form. To directly measure the impact of genetic variation on transcript abundance, we produced long-read native poly(A) RNA-seq data for 60 genetically different lymphoblastoid cell lines (LCLs) from the 1000 Genomes/GEUVADIS project. We identified 11,929 protein-coding genes and lncRNAs expressed in at least 50% of the samples, in agreement with published expression data using the same samples based on short-read sequencing. We identified 44,993 transcripts, of which 61% were novel. Of the annotated transcripts, 35% were expressed in all of the 60 LCL samples while this was only true for 14% of novel transcripts.

A genome-wide QTL analysis on transcripts identified 105 variants associated with specific transcripts (trQTLs; FDR 5%) and 34 variants associated with the total expression of the gene (eQTLs). Of the 105 trQTLs, 92 were not identified as eQTLs using a larger published short-read dataset (317 samples). Genes with eQTLs detected with long-reads had a significantly lower number of annotated transcripts than genes with trQTLs (p -value = 5.23e-05), suggesting that genetic effects on genes with higher transcript diversity were missed using gene-level quantifications. The conditional analysis identified secondary trQTLs per gene for 52 genes with 57.7% showing an opposite direction of effect on different transcripts, potentially explaining the inability to detect eQTL effects using gene-level quantifications. For example, rs4796398 was significantly associated with 5 transcripts of the *EIF5A1* gene, while no gene level cis-eQTLs were identified with long or short-read sequencing data. Overall, we were able to identify new trQTLs based on a small number of samples, whose effects on expression were missed when using short-read technology. Current work is focused on identifying RNA modifications and associated QTLs.

THE IGVF “8-CUBED” SINGLE NUCLEUS RNA-SEQ DATASET OF TRANSCRIPTIONAL VARIATION ACROSS MOUSE GENOTYPES

Elisabeth Rebboah^{1,2}, Sina Booeshaghi³, Heidi Y Liang¹, Delaney Sullivan³, Diane Trout³, Maria Carilli³, Ghassan Filimban¹, Parvin Mahdipoor¹, Jasmine Sakr^{1,4}, Fairlie Reese^{1,2}, Brian Williams³, Ingileif Hallgrimsdottir³, Shimako Kawauchi⁵, Grant McGregor⁵, Kim Green⁶, Lior Patcher³, Barbara J Wold³, Ali Mortazavi^{1,2}

¹University of California, Irvine, Department of Developmental and Cell Biology, Irvine, CA, ²University of California, Irvine, Center for Complex Biological Systems, Irvine, CA, ³California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA, ⁴University of California, Irvine, Department of Pharmaceutical Sciences, Irvine, CA, ⁵University of California, Irvine, Transgenic Mouse Facility, University Laboratory Animal Resources, Office of Research, Irvine, CA, ⁶University of California, Irvine, Department of Neurobiology and Behavior, Irvine, CA

The impact of genomic variation on gene expression is a crucial aspect of understanding the molecular basis of phenotypic traits. Identifying the genomic loci associated with regulation of gene expression, or eQTLs, is a major focus of human genetics but is limited by factors such as sample size, environmental factors, and cell type specificity. As part of the IGVF consortium's goal of cataloging cell type specific eQTLs, we first characterize the transcriptional landscape of the mouse Collaborative Cross (CC) line founders. The 8 CC founders, made up of 5 classical inbred strains and 3 wild-derived strains, are both genetically and phenotypically well-characterized. Our “8-cubed” dataset consists of 512 samples from the 8 founders with 4 female and 4 male replicates per genotype across 8 tissues: cortex and hippocampus, hypothalamus and pituitary, adrenal gland, heart, kidney, liver, skeletal muscle, and gonads. We apply single-nucleus RNA sequencing to generate high-resolution transcriptomic profiles using the Parse Biosciences combinatorial barcoding platform. Our data analysis will include gene expression quantification and differential gene expression analysis between the 8 founder strains in specific cell types, with the expectation that genotype impacts gene expression in some cell types but has no effect in others. Overall, this study aims to contribute to our understanding of human genetics and disease by first providing insight into the impact of genotype on gene expression and its relationship with phenotypic variation in mice at the single-nucleus level.

THE ENCODE4 LONG-READ RNA-SEQ DATASET REVEALS DISTINCT CLASSES OF TRANSCRIPT STRUCTURE DIVERSITY

Fairlie Reese^{*1}, Brian Williams^{*2}, Elisabeth Rebboah¹, Narges Rezaie¹, Diane Trout², Heidi Liang¹, The ENCODE RNA Working Group³, Barbara J Wold², Ali Mortazavi¹

¹University of California, Irvine, Center for Complex Biological Systems, Irvine, CA, ²California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA, ³ENCODE, RNAWG, N/A, CA

*These authors contributed equally

The majority of mammalian genes encode multiple transcript isoforms that result from differential promoter use, changes in internal splicing, and alternative 3' end choice. The comprehensive characterization of transcript structure diversity across tissues, cell types, and species is challenging because transcripts are much longer than the short reads normally used for RNA-seq. Long-read RNA-seq (LR-RNA-seq) allows for identification of the complete structure of each transcript. As part of the final phase of the ENCODE Consortium, we sequenced 264 LR-RNA-seq PacBio libraries totaling over 1 billion circular consensus reads (CCS) for 81 unique human and mouse samples. We detected and quantified at least one transcript from 89.7% of annotated human GENCODE protein coding genes for a total of ~200,000 full-length transcripts, ~40% of which have novel intron chains.

We use a new gene and transcript annotation framework using triplets based on the transcript start site, intron chain, and transcript end site used in each transcript. We analyze the resulting transcriptional structure diversity of each gene using the corresponding gene structure simplex and find that half of the genes with more than one expressed transcript show a clear bias toward one of the three mechanisms of transcript structure diversity. We identify a predominant transcript for each sample, which is the most highly expressed transcript in a given gene, and find that 73.8% of expressed protein coding genes have more than one predominant transcript across our dataset, with a median of 3 per gene. Furthermore, we show that 42.3% of genes with a MANE transcript have a different predominant transcript in 67.7% of the samples that they are expressed in. We also find that while human and mouse have globally similar transcriptomes, more than half of orthologous genes show substantial changes in transcript structure diversity in matching samples. Our results represent the first comprehensive survey of both human and mouse transcriptomes using full-length long reads and will serve as a foundation for further analyses of alternative transcript usage.

A ROBUST STATISTICAL FRAMEWORK FOR GENE-WISE SINGLE-CELL DIFFERENTIAL EXPRESSION META-ANALYSIS IN THE CONTEXT OF POPULATION-BASED SINGLE-CELL STUDIES.

Aida Ripoll-Cladellas^{*1}, Monique G.P. van der Wijst², Marc Jan Bonder^{2,3,4},
Lude Franke², Marta Melé¹

¹Barcelona Supercomputing Center, Life Sciences Department, Barcelona, Spain, ²University Medical Center Groningen, Department of Genetics, Groningen, Netherlands, ³German Cancer Research Center, Division of Computational Genomics and Systems Genetics, Heidelberg, Germany, ⁴European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany

Single-cell RNA sequencing (scRNA-seq) has enabled deciphering the human transcriptome at an unprecedented resolution. Its popularity for studying how inter-individual variation affects expression has grown tremendously now that scale, cost, and sensitivity have significantly improved. To fully leverage these emerging population-based scRNA-seq data resources, we have founded the single-cell eQTLGen consortium (sc-eQTLGen), aimed at pinpointing the cellular contexts in which disease-causing genetic variants affect gene expression. Our consortium builds on a federated structure that ‘brings the algorithm to the data’, thereby overcoming the necessity to share privacy-sensitive data, while concurrently reducing the computational load needed for processing all datasets together. Here, we expand the sc-eQTLGen meta-analysis setup to pinpoint the cellular contexts in which specific individual traits (such as age, sex, or ethnicity) or different environmental conditions (such as immune stimulation) affect gene expression. To this end, we have developed a statistical framework to conduct a cell-type-specific gene-wise single-cell differential expression meta-analysis (SiGMeta-DE). As a proof of concept, we applied this framework to three peripheral blood mononuclear cell datasets spanning 809,353 cells from 173 different donors to study how sex (discrete) and age (continuous) affect gene expression. Our approach overcomes several limitations. First, we have used MAST, a two-part Hurdle generalized linear model with random effects for individuals to account for both zero inflation and pseudoreplication bias. Second, by using a meta-analysis approach, we avoid sharing privacy-sensitive and re-analyzing previously processed data, which can often be difficult and cumbersome. We show that our meta-analysis methodology substantially increased the statistical power to detect differentially expressed (DE) genes. Sex-DE genes detected in the individual datasets were all located in the sex chromosomes as these have generally large effect sizes. Our meta-analysis identified additional sex-DE genes, including many with smaller effect sizes located in autosomal chromosomes. In addition, the meta-analysis considerably increased the number of age-DE genes up to hundreds. Newly identified age-DE genes overlapped previously identified age-related genes but many are novel. Together, our meta-analysis framework overcomes the difficulties that have previously hampered studying inter-individual variation in gene expression at single-cell resolution and allows the identification of DE effects that would otherwise remain hidden. Our approach provides a solid framework to associate single-cell molecular phenotypes (such as single-cell chromatin accessibility or single-cell DNA methylation) with demographic traits or environmental conditions in future studies.

SYSTEMATIC COMPUTATIONAL SEARCH FOR ORF GENE FUNCTION

Ellen Tsai^{1,2}, Bin Ye², Suganthi Balasubramanian², Alan R Shuldiner², Juan L Rodriguez-Flores²

¹University of California at Los Angeles, Computational and Systems Biology, Los Angeles, CA, ²Regeneron Pharmaceuticals Inc, Regeneron Genetics Center, Tarrytown, NY

Over 5% of nearly 20 thousand genes in the human genome have no known function, do not belong to a gene family, and hence have generic names such as “C15orf54”, meaning “chromosome 15 open reading frame 54”. We refer to them as “ORF” genes. While ORF genes do not have any homologs in the human genome, they do produce mRNA detectable in bulk transcriptome sequence data, and their predicted amino acid sequences are highly conserved across species. In this study, we sought to further characterize ORF genes with the objective of systematically determining their function and role in human disease (if any).

As a first question regarding these genes, we sought to determine if ORF genes are under greater or lesser evolutionary constraint compared to other genes. Using GnomAD data that includes pLI (probability of loss of function intolerance) and OE (observed/expected ratio of variant counts per gene/transcript) scores of evolutionary constraints on 19,704 human genes, we separated these genes into two sets with similar distributions of gene length and coding sequence length. The first set consists of 597 ORF genes, with names matching the case-insensitive regular expression ‘C[0-9]+ORF[0-9]+’, and the second set consisted of 13,318 “known” genes, after excluding genes with lengths longer than 1 standard deviation above the mean. The size length distributions were confirmed to be similar between the two sets by t test (p value > 0.05). The two sets of genes had significantly different distributions of pLI, exAC pLI, LoF OE, missense OE, and PolyPhen missense OE. In general, the pLI scores were lower for ORF genes, suggesting lower probability of haploinsufficiency. OE scores were higher for ORF genes, suggesting higher tolerance to genetic variability. Overall, the trend was consistent across both scores, and we concluded that ORF genes have less evolutionary constraint than known genes.

Given the lower evolutionary constraint of the ORF genes, we sought to determine if ORF genes are less likely to be mapped to a phenotype in public GWAS catalog data. The GWAS catalog data was filtered such that the mapped gene was one of the genes in the two sets described above with similar gene and coding sequence lengths. A frequency distribution of GWAS associations mapped to each gene was computed, and these distributions were compared using a t test. There were 1,601 associations mapped to 161 ORF genes, and 103,088 associations mapped to 7,310 known genes of similar length. Known genes had a significantly higher number of associations on average (14.1 +/- 53.0) than ORF genes (9.9 +/- 21.0) p value = 0.019. This result suggested a correlation between higher evolutionary constraint and more GWAS associations for a gene.

In conclusion, our data suggest ORF genes are less constrained than known genes with lower likelihood of playing a role in phenotypic variation.

BACTERIAL PANGENOMES SHAPE ESSENTIALITY AND GENE-PHENOTYPE ASSOCIATIONS AND THEREBY DRIVE GENOME-EVOLUTION.

Federico Rosconi¹, Tim van Opijnen²

¹Boston College, Biology Department, Chestnut Hill, MA, ²Broad Institute of MIT and Harvard, Cambridge, MA

Whole-genome sequencing has led to the astonishing discovery that bacterial isolates from the same species can significantly differ in their genetic content. Consequently, instead of only focusing on a single strain, actual microbial genomics studies must consider the "pangenome," i.e., the sum of all genes present across all strains in a species. For instance, a strain of the bacterial pathogen *Streptococcus pneumoniae* contains ~2100 genes, while the entire species harbors >4400 genes, which means that two random strains may differ by the presence and absence of hundreds of different genes. Considering that even small genetic changes can have far-reaching phenotypic consequences, we study how *S. pneumoniae*'s pangenome influences two crucial aspects of its biology: gene essentiality and mechanisms required for host infection. By using Tn-Seq, whole-genome sequencing, and RNA-Seq on a set of 36 *S. pneumoniae* clinical strains representing >68% of its pangenome, we have identified the ESSENTIALOME of the pathogen, e.g., the set of genes essential for growth in at least one strain of the collection. These data highlight that some genes are essential in every strain, while for a subset, essentiality is strain-dependent. We show through different genomics approaches, including experimental evolution, that this strain dependency relies on different mechanisms, including (1) toxic intermediates accumulation, (2) functional redundancy, (3) metabolite recycling, and (4) rewiring of critical pathways. Moreover, by *in vivo* Tn-Seq in a mouse pneumonia model, we have characterized, with a pan-genome perspective, the *in vivo* phenotype of the same strain collection. We find that 338 genes are required by at least two strains to invade the lungs, and within these, only 77 genes are required by every strain. An in-depth analysis of the strain-dependent required genes reveals that: 1) the requirement of some genes depends on how virulent a strain is; and 2) the requirements of central cell processes correlate with the presence/absence of a poorly characterized bacteriocin. These results show that a simple mechanism, e.g., the gain or loss of a single bacteriocin-coding gene, could dictate the strategies required by *S. pneumoniae* to survive inside the host. In conclusion, we demonstrate that bacterial pangomes make gene-essentiality and specific gene-phenotype associations a fluid concept. Bacterial pan-genomes are thereby not only under selection by their environment, but the "ebb and flow" of genetic material between genomes of the same and different species generates intragenomic interactions that shape the genome and create additional selective pressures that we are, through the development of new genomics tools, starting to be able to map-out and predict. Interactions born from the existence of pangomes may thus have profound implications on for instance niche specialization, or even speciation, and thereby drive the evolution of microbial genomes.

A HUMAN ATLAS OF IMPRINTING AND ALLELE-SPECIFIC METHYLATION

Jonathan Rosenski¹, Ayelet Peretz², Judith Magenheimer², Ruth Shemer², Yuval Dor², Benjamin Glaser³, Tommy Kaplan^{1,2}

¹The Hebrew University of Jerusalem, School of Computer Science and Engineering, Jerusalem, Israel, ²Institute for Medical Research Israel-Canada, Hadassah Medical Center and Faculty of Medicine, The Hebrew University of, Dept. of Developmental Biology and Cancer Research, Jerusalem, Israel, ³Hadassah Medical Center and Faculty of Medicine, The Hebrew University of Jerusalem, Dept. of Endocrinology and Metabolism, Jerusalem, Israel

Allele-specific methylation plays a critical role in embryonic development and gene regulation. Imprinted genes, where only the maternally or paternally inherited allele is expressed, are associated with allele-specific methylation at imprinting control regions. eQTLs are also associated with meQTLs, which in turn leads to allele specific methylation. While these regions are of great importance to understanding the biology of early development, gene regulation mechanisms, and genetic disease, they have rarely been studied in humans outside of blood samples.

In this study, we generated a whole-genome human cell type-specific DNA methylation atlas, and developed computational algorithms for the joint analysis of genetic and epigenetic patterns, across >200 samples. We identified 28,000 genomic regions with allele-specific DNA methylation patterns. Many of these show sequence-dependent methylation, elucidating the relationship between genotype and phenotype, in a tissue-specific manner.

We also identified 275 “imprinted” regions, with allelic-specific parent-of-origin differential methylation. These include most known ICRs and >100 novel regions, many of which show cell type-specific “escape” from imprinting in a limited number of cell types.

These regions shed light on the molecular mechanisms underlying uniparental expression of imprinted genes, and include differential methylation of regulatory regions, allele-specific methylation of CTCF sites which alter the 3D folding of the genome, and recruitment of distal, cell type-specific enhancers. In addition, this human cell type-specific allele-specific methylation atlas and associated methods offer a rich resource for the study of genomic imprinting biology and provide insight in the pathogenesis of genetic diseases.

SPATIAL TRANSCRIPTOMICS IN THE ASTROPATH PLATFORM

Jeffrey S Roskes^{1,2}, Elizabeth L Engle^{3,4,5}, Long Yuan⁵, Atul Deshpande⁶,
Joel C Sunshine⁵, Kellie Smith^{4,6}, Drew M Pardoll^{4,6}, Janis M Taube^{*3,4,5,7},
Alexander S Szalay^{*1,2,3}

¹Johns Hopkins University, Department of Physics and Astronomy, Baltimore, MD, ²Johns Hopkins University, Institute for Data Intensive Engineering and Science, Baltimore, MD, ³Johns Hopkins University, The Mark Foundation Center for Advanced Genomics and Imaging, Baltimore, MD, ⁴Johns Hopkins University, Bloomberg-Kimmel Institute for Cancer Immunotherapy and Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD, ⁵Johns Hopkins University School of Medicine, Department of Dermatology, Baltimore, MD, ⁶Johns Hopkins University School of Medicine, Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD, ⁷Johns Hopkins University School of Medicine, Department of Pathology, Baltimore, MD

The AstroPath platform facilitates the study of the tumor microenvironment (TME) through universal standards to ensure high quality and uniform data. The data are stored in a database, with precomputed quantities that enable fast data processing and analysis. The project is built using decades of experience in astronomy and is modeled after the Sloan Digital Sky Survey (SDSS). To date, the database contains hundreds samples with a total of over half a billion cells and is several times larger than the entire SDSS dataset. Analyses using AstroPath have facilitated discoveries of biomarkers that can predict response to treatment in melanoma [1] and lung cancer [2].

Although the platform's initial development focused on multiplex immunofluorescence (mIF) data, the project's vision is to include many different data modalities in a single, cohesive database. This talk presents a new application of AstroPath to spatial transcriptomics data scanned using H&E staining and sequenced using the 10X Genomics platform. The images are calibrated using an adapted version of the AstroPath algorithms used for mIF data, and the transcriptomic data are loaded into the database alongside the images, facilitating fast, precise, reproducible analysis of genomic and image data simultaneously.

[1] Berry, et al. Science 2021.

[2] Emily Cohen, Daphne Wang, Elizabeth Engle, et al. CD8+FoxP3+ cells represent early, effector T-cells and predict outcomes in patients with resectable non-small cell lung carcinoma (NSCLC) receiving neoadjuvant anti-PD-1-based therapy. Abstract #57; SITC 2022 annual meeting.

*authors contributed equally

ENHANCING GENE EXPRESSION PREDICTION IN MAJOR PSYCHIATRIC DISORDERS VIA CO-EXPRESSION MODELS

Fabiana Rossi^{1,2}, Madhur Parihar¹, Leonardo Sportelli², Loredana Bellantuono², Nora Penzel², Joel E Kleinman^{1,3}, Joo Heon Shin¹, Thomas M Hyde¹, Daniel R Weinberger¹, Giulio Pergola^{1,2,3}

¹Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, ²Group of Psychiatric Neuroscience, Department of Translational Biomedicine and Neuroscience, University of Bari Aldo Moro, Bari, Italy, ³Johns Hopkins University School of Medicine, Department of Psychiatry and Behavioral Sciences, Baltimore, MD

Genetic risk for major psychiatric disorders is mainly attributed to non-coding single nucleotide polymorphisms (SNPs), likely affecting gene expression. Predicting gene expression from postmortem brain genetic data generally relies on *cis*-eQTLs. Although, *trans*-eQTLs may explain a large portion of the variance, small effect sizes hinder their identification. In this study, we leveraged co-expression to capture the joint effect of weakly associated *trans*-eQTLs and provide additional expression predictions for *cis*-eQTLs-depleted genes.

We trained *trans*-eQTLs-derived prediction models using the elastic-net algorithm via nested 10-fold cross-validation (cv) on genotype and gene expression data from tissue homogenate RNA-sequencing in the dorsolateral prefrontal cortex from 428 European Ancestry (EA) (Common Mind Consortium Project). We retained only significant models (average $r > 0.1$ and Z-score $p \leq 0.05$ across cv-folds) and we tested performances on a testing dataset of 145 EA cohort (Lieber Institute for Brain Development). We applied the published EpiXcan model to generate *cis*-imputed gene expression, and we compared its predictive performances with our generated *trans*-models: i) "Imputed Network Gene-Expression *trans*-eQTL" (INGENE), exploiting the imputed co-expression partners to predict the expression of each target gene, and ii) "Module QTL Eigengene" (MODULE), that combines SNPs associated with the co-expression module eigenvalue.

As expected, EpiXcan explained more variance than *trans*-eQTLs (5% across 8,137 genes) compared to ~1% (1,790 genes, INGENE) and 2% (3,838 genes, MODULE). Intriguingly, the models showed limited overlap of predicted genes. Only 916 and 1916 genes were predicted by both EpiXcan/INGENE and by both EpiXcan/MODULE. Only 862 genes overlapped between INGENE/MODULE. Our results highlight the high synergistic potential of stacking the *cis*- and *trans*-predictions.

These preliminary results provide a foundation for further work to improve gene prediction outcomes and expand the number of imputed heritable genes through co-expression models. Incorporating predictions from multiple brain regions and larger samples is expected to enhance prediction accuracy further.

A COMPREHENSIVE rRNA VARIATION ATLAS IN HEALTH AND DISEASE

Daphna Rothschild*¹, Teodorus T Susanto*¹, Jeffrey P Spence¹, Naomi R Genuth^{1,2}, Nasa Sinnott-Armstrong¹, Jonathan K Pritchard§^{1,2}, Maria Barna§¹

¹Stanford University, Genetics, Stanford, CA, ²Stanford University, Biology, Stanford, CA

* Co-first author

§ Co-senior author

Repetitive DNA sequences constitute a significant portion of genomes across species, making up over 70% of the human genome. One large class of repetitive DNA elements are ribosomal DNA (rDNA) genes present in hundreds of copies spread across multiple chromosomal loci. Despite their high numbers, sequence variations among rDNA copies are vastly ignored and it remains an outstanding question whether different ribosomal RNA (rRNA) subtypes exist and if variations impact human physiology and disease. Here, we generated the first complete atlas of rRNA variants and show their association with development and cancer. For this, we optimized long-read sequencing of full-length 18S and 28S rRNA from actively translating ribosomes, and developed an efficient novel computational algorithm to detect all variations between related sequences. We discovered hundreds of variants that are not silent but are incorporated into translating ribosomes. These include tens of abundant variants within functionally important domains of the ribosome. Strikingly, variants assemble into distinct ribosome subtypes encoded on different chromosomes. We further examined whether rRNA variants can be regulated, and found differential variant expression across human tissues. Specifically, we analyzed rRNA variant expression in normal tissues in the Genotype-Tissue Expression (GTEx) dataset and discovered tissue-specific variant expression in endoderm/ectoderm derived tissues, especially prevalent between brain and digestive system tissues. In cancer, we analyzed The Cancer Genome Atlas (TCGA) dataset and observed multiple cancer-specific rRNA variants that are normally lowly abundant to be elevated in cancer, with significant relative abundance differences between malignant and control biopsies. Therefore, even lowly expressed rRNA variants could potentially be transcribed in the context of disease, revealing the importance of rRNA variation in human health. Together, this study provides a curated atlas for exploring rRNA variation and our findings functionally link ribosome variation to development and cancer.

GENETIC VARIANTS ASSOCIATED WITH IMMUNE CELL POPULATION ABUNDANCES IN SINGLE-CELL DATA

Laurie Rumker^{1,2,3,4}, Saori Sakaue^{1,2,3,4}, Joyce B Kang^{1,2,3,4}, Yakir Reshef^{1,2,3,4}, Cristian Valencia^{1,2,3,4}, Seyhan Yazar⁵, José Alquicira-Hernandez⁵, Joseph Powell⁵, Soumya Raychaudhuri^{1,2,3,4}

¹Brigham and Women's Hospital, Center for Data Sciences, Boston, MA,

²Harvard Medical School, Division of Genetics, Boston, MA, ³Harvard Medical School, Department of Biomedical Informatics, Boston, MA,

⁴Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Boston, MA, ⁵Garvan Institute of Medical Research, Cellular Science, Sydney, Australia

Genome-wide association studies have shown that an individual's genetic background alters their risk for many autoimmune diseases, yet the physiologic mechanisms mediating these associations largely remain unknown. In blood samples, disease-linked genetic variants have been found to influence the expression of individual genes or the abundances of individual proteins. Such effects may act in aggregate over time to shift the relative abundances of functional cell populations and precipitate disease. Previously, genetic associations to blood cell population abundances have been assessed in flow cytometry data, where specific hypothesized cell populations to test for association are defined *a priori* and distinguished by select protein markers. By contrast, we have developed a novel approach to identify fine-grained immune cell states in single-cell transcriptomic data that change in abundance with allelic dose of a genetic variant. Our approach avoids pre-specifying candidate cell types to test so that we can characterize genetically-associated cell states with more flexibility and statistical power.

When applied in a genome-wide survey of previously published data from 969 Australians of European ancestry—the OneK1K cohort—our method reveals seven independent genome-wide significant loci (lead P-values < 2e-8) associated with changes in immune cell populations. For example, one locus on chromosome 19 (lead SNP rs10743893, P-value 1.7e-12) is a known expression quantitative trait locus (eQTL) for the NK cell lectin-like receptor gene KLRC1, key to innate immune destruction of malignant and virally-infected cells. We find that this SNP is associated with simultaneous shifts in the relative abundances of T, NK, and myeloid cell states, including increased abundance of activated NK cells and CD8+ effector memory T cells. Because we detect these shifts in a population cohort without acute disease, characterizing these effects may illuminate how our genetic background can set the immunological stage for disease risk.

THE CZ CELLXGENE DISCOVER SUITE IS AN ANALYTICAL PLATFORM AND THE LARGEST REPOSITORY OF STANDARDIZED SINGLE-CELL DATA

Erica M Rutherford¹, Jim Chaffer¹, Jenny Chien¹, Lian Morales¹, Jennifer L Zamanian¹, Jason Hilton¹, Michael Cherry¹, CZI Single Cell Biology Team²

¹Stanford University, Genetics, Palo Alto, CA, ²Chan Zuckerberg Initiative, Single-Cell Biology, Redwood City, CA

CZ CELLxGENE Discover (cellxgene.cziscience.com) is a free-to-use online data portal hosting a growing corpus of more than 700 single-cell datasets comprising over 38 million unique cells from the major human and mouse tissues. As of today the portal hosts single-cell data from modalities that include gene expression, chromatin accessibility, DNA methylation, and spatial transcriptomics. All data have been standardized to include raw counts, gene IDs and symbols, as well as cell metadata, such as cell type, tissue, and donor age and disease, each standardized to a community ontology. All data are easily searchable and can be downloaded in both AnnData and Seurat formats via web or by programmatic API calls.

UI-based tools allow for exploration of single datasets without requiring download. New features for the analysis of the entirety of the data are available and under active development. The CELLxGENE explorer displays an interactive 2-dimensional representation of cells in a dataset and allows users to color cells by metadata (e.g. cell type, disease, metadata features etc.) or gene activity. Users can also subset and analyze subgroups of cells, perform differential gene expression and create scatter plots of gene expression. The Gene Expression feature allows querying the expression of any gene across all human and mouse cell types available in the portal, and enables lookup of differentially enriched genes for any cell type. Finally, APIs enable download of individual datasets and cell-based slicing and download of data subsets. CELLxGENE is continuously improving usability and adding new features tailored to the needs of cell and computational biologists.

CELLxGENE is a tool intended for community use and contributions. By supporting multiple modalities and data generated by labs around the world, the CELLxGENE suite of tools and data aims to maximize rapid reuse of high quality data describing the phenotypes of cells and tissues. To date, CELLxGENE supports data sharing from labs around the world as well as consortia such as the CZ Biohub Tabula projects, LungMap, BICCN, Allen Institute for Brain Science, KPMP, HTAN and the Human Cell Atlas. New collaborations and contributions are welcome, and the CELLxGENE team actively supports curation of data to ensure an efficient submission process. Groups interested in submitting their own data can inquire about the inclusion of your data and the submission process by contacting the CELLxGENE team at cellxgene@chanzuckerberg.com.

AGE AND SOCIAL STATUS ARE ASSOCIATED WITH TH1 AND TH2 IMMUNE GENE REGULATORY RESPONSES IN RHESUS MACAQUES

Mitchell Sanchez Rosado¹, Marina M Watowich², Laura Newman³, Melissa A Pavez-Fox⁴, Erin R Siracusa⁴, Macaela Skelton⁴, Josue E Negron-Del Valle⁵, Daniel Phillips⁵, James P Higham³, Lauren Brent⁴, Amanda J Lea², Carlos Sariol⁵, Noah Snyder-Mackler¹

¹University of Puerto Rico-Medical Sciences, Microbiology & Medical Zoology, San Juan, PR, ²Vanderbilt University, Biological Sciences, Nashville, TN, ³New York University, Anthropology, New York, NY,

⁴University of Exeter, Centre for Research in Animal Behaviour, Exeter, United Kingdom, ⁵Arizona State University, School of Life Sciences, Tempe, AZ

Aging is accompanied by immune function decline and increased inflammation. These signatures of immunosenescence are also found in individuals who experience poor health outcomes attributed to social adversity. One possible mechanism through which social adversity may impact health outcomes is by altering immune gene regulation, thus altering immune homeostasis. Here, we investigated how age and social adversity (operationalized as low social status) relate to inflammatory Th1 and anti-inflammatory Th2 immune responses in free-ranging rhesus macaques (n=180). We stimulated peripheral blood mononuclear cells (PBMCs) in vitro with Lipopolysaccharide (LPS) and Dexamethasone (DEX) to stimulate Th1 and Th2 responses respectively, and used mRNA-seq to measure how these responses were associated with variation in age and social status. We detected 3,233 and 2,923 differentially expressed genes (FDR<0.1) in response to LPS and DEX, respectively. Known LPS responsive genes - such as IL1 β and TNF α - and glucocorticoid (GC) responsive genes - such as FKBP5 and TSCD22 – were highly induced. However, the LPS response varied with age: younger individuals had increased immune activation and signaling, including genes in the TLR-4 signaling pathway such as MYD88 and NF κ B2 (FDR<.1), compared to older individuals. Additionally, older age was associated with lower mean expression of genes in the inflammatory and TNF α /NF κ B (p=5x10⁻⁵) pathways. DEX also had an age-associated relationship with gene expression, with older individuals having lower expression of GC responsive genes such as DUSP1 and HSP90 family members (FDR<0.1), pointing to a diminished GC response. Low social status males had significantly higher average expression of inflammatory pathway genes (p=1.2x10⁻²⁰) compared to high status males regardless of age. Moreover, low status males had higher mean expression (p=1.9x10⁻⁴) of genes in the GC resistance-associated c-Myc/MAPK pathway. Together, these findings identify the molecular pathways and arms of the immune system through which social adversity may recapitulate age-related differences in immunity and, ultimately, age-related disease and death.

APPLICATION OF NANOPORE SEQUENCING FOR LIQUID BIOPSY ANALYSIS IN CHILDREN WITH CANCER

Carolin M Sauer¹, Nicholas Tovey², Debbie Hughes³, Marwane Bourdim¹, Reda Stankunaite³, Joanne Stockton², Claire Lynn³, Harvey Che³, Michael Hubank³, John Anderson⁴, Andrew D Beggs², Louis Chesler³, Isidro Cortés-Ciriano¹

¹EMBL, EBI, Cambridge, United Kingdom, ²University of Birmingham, Institute of Cancer and Genomics Sciences, Birmingham, United Kingdom,

³Institute of Cancer Research, Molecular Pathology, London, United Kingdom, ⁴University College London, Great Ormond Street Hospital, London, United Kingdom

Pediatric cancers are the leading cause of death in children post infancy in the Western world. Comprehensive, high-throughput molecular profiling is essential to elucidate the molecular basis of evolving, treatment resistant disease and to more effectively guide clinical decision making. Access to high-quality tumor material for genomic profiling is a challenge in children where tissue biopsies are small. The analysis of cell-free DNA (cfDNA) from liquid biopsies for the detection of circulating tumor-derived DNA (ctDNA) offers a powerful, minimally invasive alternative to tumor profiling. However, at present, ctDNA analysis is limited in sensitivity, specificity, scalability, turnaround time and cost, hindering its implementation into standard clinical care.

Emerging Nanopore sequencing platforms can report on native DNA in either whole-genome or focused format, and are rapid and scalable at low cost, making this platform highly attractive in the clinical setting. Here, we exploit Nanopore sequencing for the multi-modal analysis of cfDNA in pediatric cancer patients, including the analysis of cancer-specific mutations and epigenetic alterations.

Using matched liquid biopsies and tumor tissues obtained from patients with relapsing cancer enrolled in the *UK Stratified Medicine Paediatrics* study, we demonstrate the utility of Nanopore sequencing to detect clinically relevant somatic aberrations, such as *ALK* and *MYCN* amplifications from low volume blood draws in pediatric cancer patients. Overall, copy number aberrations detected using Nanopore sequencing were highly concordant with those detected using Illumina whole-genome sequencing. Using the ability of Nanopore sequencing to read out epigenetic modifications, we show novel methylation deconvolution algorithms to accurately specify tissue-of-origin and oncotype. Finally, integration of copy number and methylation data enables monitoring of disease burden, progression, and detection of disease relapse in longitudinal plasma samples. Taken together, our results suggest that Nanopore-based multi-modal liquid biopsy analysis of ctDNA may present a powerful tool to significantly improve treatment in children with cancer by facilitating early detection, accurate diagnosis, and efficient serial monitoring of disease progression.

ISOFORM-SPECIFIC FUNCTIONS OF RNA-BINDING PROTEINS

Megan D Schertzer^{1,2}, Stella H Park^{1,3}, Erin D Jeffery⁴, Gloria Sheynkman⁴, David A Knowles^{1,5}

¹New York Genome Center, NYGC, New York, NY, ²Columbia University, Computer Science Department, New York, NY, ³Columbia University, Department of Biomedical Engineering, New York, NY, ⁴University of Virginia, Department of Molecular Physiology and Biological Physics, Charlottesville, VA, ⁵Columbia University, Systems Biology Department, New York, NY

RNA binding proteins (RBPs) play outsized roles in cells by interacting with all RNAs to regulate essential aspects of RNA metabolism such as splicing, localization, degradation, export, and translation. Therefore, defects in RBPs can be highly disruptive to basic cellular activities and often lead to disease, especially neurological and muscular disorders, and cancer. Importantly, RBPs are subject to their own regulation, especially splicing, so that multiple isoforms of a given RBP are expressed in the same cell and differentially across cell types. RBP isoforms that produce diverse protein products have many potential downstream effects, but studies of isoform-specific RBP functions are limited.

In this study, we aim to identify annotated and novel RBP isoforms across different tissues, predict downstream functional effects, and validate effects in human cell lines. First, to identify high-confidence RBP isoforms, we developed a customized analysis pipeline that adapts FLAIR and SQANTI3 tools to process 96 PacBio long-read RNA-seq datasets from the ENCODE consortium. From this, we identified 11,173 isoforms for 1,413 RBPs, with an average of 8 isoforms per RBP. Next, to predict protein localization, we used NLSdb to identify nuclear localization signals (NLS) across all isoforms. Based on differential inclusion/exclusion of NLSs, 479 RBPs had isoform-specific protein localization patterns, including RBFOX2 and QKI. To validate these predictions and to test the downstream functional effects of these differences, we expressed FLAG-tagged versions of each RBFOX2 and QKI isoforms in human stem cells. Using a FLAG antibody, we have performed isoform-specific RNA-immunoprecipitation, IP-mass spectrometry, and western blot to investigate differences in RNA binding, co-factor binding, and protein localization, respectively. In addition, we plan to similarly predict, validate, and test downstream effects of RBP isoforms with differential inclusion/exclusion of other protein domains, such as RNA-binding domains. Overall, we provide a strategy to identify and mechanistically investigate isoform-specific functions of RBPs.

DEFINING ANCESTRY, HERITABILITY AND PLASTICITY OF CELLULAR PHENOTYPES IN SOMATIC EVOLUTION

Joshua S Schiffman^{*1,2}, Andrew R D'Avino^{*1,2}, Tamara Prieto^{*1,2}, Yakun Pang³, Yilin Fan⁴, Srinivas Rajagopalan^{1,2}, Catherine Potenski^{1,2}, Toshiro Hara⁴, Mario L Suva⁴, Charles Gawad³, Dan A Landau^{1,2}

¹New York Genome Center, Somatic Evolution, New York, NY, ²Weill Cornell Medicine, Physiology and Biophysics, New York, NY, ³Stanford University School of Medicine, Department of Pediatrics, Stanford, CA, ⁴Massachusetts General Hospital and Harvard Medical School, Department of Pathology and Center for Cancer Research, Boston, MA

The broad application of single-cell RNA sequencing has revealed transcriptional cell state heterogeneity across diverse healthy and malignant somatic tissues. Recent advances in lineage tracing technologies have further enabled the simultaneous capture of cell transcriptional state along with cellular ancestry thus enabling the study of somatic evolution at an unprecedented resolution; however, new analytical approaches are needed to fully harness these data. Here we introduce **PATH** (the Phylogenetic Analysis of Transcriptional Heritability), an analytical framework, which draws upon classic approaches in species evolution, to quantify heritability and plasticity of somatic phenotypes, including transcriptional states. The PATH framework further allows for the inference of cell state transition dynamics by linking a model of cellular evolutionary dynamics with our measure of heritability versus plasticity. We evaluate the robustness of this approach by testing a range of biological and technical features in simulations of somatic evolution. We then apply PATH to characterize previously published and newly generated single-cell phylogenies, reconstructed from either native or artificial lineage markers, with matching cellular state profiling. PATH recovered developmental relationships in mouse embryogenesis, and revealed how anatomic proximity influences neural relatedness in the developing zebrafish brain. In cancer, PATH dissected the heritability of the epithelial-to-mesenchymal transition in a mouse model of pancreatic cancer, and the heritability versus plasticity of transcriptionally-defined cell states in human glioblastoma. Finally, PATH revealed phenotypic heritability patterns in a phylogeny reconstructed from single-cell whole genome sequencing of a B-cell acute lymphoblastic leukemia patient sample. Altogether, by bringing together perspectives from evolutionary biology and emerging single-cell technologies, PATH formally connects the analysis of cell state diversity and somatic evolution, providing quantification of critical aspects of these processes and replacing qualitative conceptions of “plasticity” with quantitative measures of cell state transitions and heritability.

* authors contributed equally.

CGC1, A NEW GAP-FREE AND TELOMERE-TO-TELOMERE REFERENCE GENOME FOR *CAENORHABDITIS ELEGANS*

Kazuki Ichikawa¹, Massa J Shoura², Karen L Artiles², Chie Owa¹, Haruka Kobayashi¹, Manami Kanamori³, Yu Toyoshima³, Yuichi Iino³, Ann E Rougvie⁴, Andrew Z Fire^{2,5}, Erich M Schwarz⁶, Shinichi Morishita¹

¹The University of Tokyo, Department of Computational Biology and Medical Sciences, Chiba, Japan, ²Stanford University, Department of Pathology, Stanford, CA, ³The University of Tokyo, Department of Biological Sciences, Tokyo, Japan, ⁴University of Minnesota, Department of Genetics, Cell Biology, and Development, Minneapolis, MN, ⁵Stanford University, Department of Genetics, Stanford, CA, ⁶Cornell University, Department of Molecular Biology and Genetics, Ithaca, NY

HiFi PacBio with ultra-long Oxford Nanopore sequencing has made it possible to achieve gap-free and telomere-to-telomere reference genome assemblies for complex eukaryotes. So far, only three such genomes have been published: human, watermelon, and the nematode *Oscheius tipulae*. After five decades of use, the standard wild type strain used in *C. elegans* laboratories (N2) has accumulated genetic polymorphisms leading to observable phenotypic differences between different laboratories. We therefore generated and mass-cryopreserved CGC1, an isogenic strain derived from N2. We then used HiFi PacBio and ultra-long Nanopore to assemble 61 contigs with hiCanu, order them chromosomally via a preexisting reference, and close 11 gaps of <1 kb with the consensuses of Nanopore reads. This left 43 gaps filled with long tandem repeats of ≥10 kb in size that we closed through localized hybrid assembly, first closing the 43 gaps with Nanopore reads and then mapping, aligning, and generating consensuses from HiFi reads. Our resulting CGC1 genome has six nuclear chromosomes totaling 105.9 Mb. It encompasses 100.3 Mb of the previous N2 reference assembly, along with an extra 5.6 Mb containing all 12 telomeres, at least 50 protein-coding genes, and 5.3 Mb of tandemly repeated sequences, including repeated regions such as pSX1 (153 kb) that span all but the longest Nanopore reads. Structural variations in outgroup *C. elegans* strains indicate that most disagreements between the older N2 and the newer CGC1 assemblies are due to erroneous omissions in N2, and all but the most compact nematode genomes are likely to contain similarly elusive tandemly repeated regions. The CGC1 reference and its matched isogenic strain will enable maximum precision for future genomic analyses of the model organism *C. elegans*.

improving precision cancer treatment selection by integrating deep omic tumor characterization and patient-specific drug screening

Casey Sederman, Tonya Di Sera, Gabor T Marth

University of Utah School of Medicine, Department of Human Genetics,
Salt Lake City, UT

Background: Despite a burgeoning repertoire of anticancer therapeutics, a lack of rational guidance for precision treatment selection remains a critical barrier to precision oncology practice. Most tumors lack identifiable, clinically actionable molecular alterations, limiting personalized treatment selection to a minority of patients. Further, tumor heterogeneity manifests in highly variable drug responses, even amongst theoretically targetable patient cohorts. Recently, deep learning (DL) models have emerged as a promising approach for precision treatment selection, achieving state of the art performance in cancer cell lines. These models can incorporate high dimensional omic features, facilitating response prediction in the absence of specific, clinically actionable vulnerabilities. However, when evaluated in precision oncology-relevant settings, the performance of existing models falls far below the mark. Motivated by these challenges, we developed ScreenDL, a novel DL-based cancer drug response prediction method designed for eventual use in precision-guided therapy selection.

Methods: ScreenDL accepts drug chemical structures and tumor molecular profiles as input and predicts drug response. ScreenDL was rigorously validated in both cell lines and patient-derived models of cancer, emphasizing the ability to predict differential drug response in never-before-seen cell lines and patients. Uniquely, ScreenDL has the ability to incorporate partial drug screening data, often obtainable in realistic clinical scenarios, through a patient-specific finetuning strategy.

Results: Evaluation of ScreenDL in cancer cell lines shows that ScreenDL outperforms state of the art DL methods in precision oncology-relevant settings, achieving an overall Pearson correlation of 0.55 between observed and predicted response compared to 0.32 for existing methods. Incorporating partial, patient-specific drug screening data further improves overall Pearson correlation to 0.70. Importantly, biomaterial from surgical tumor resection is often available for patient-specific drug screening in practical clinical scenarios, informing an approach to precision oncology wherein deep multiomic tumor characterization and patient-specific drug screening can be integrated to provide reliable precision treatment recommendations.

Conclusion: Here, we demonstrate the utility of incorporating partial, patient-specific drug screening data with deep multiomic tumor characterization for improved cancer drug response prediction. ScreenDL represents an exciting new direction, bringing DL-based cancer drug response prediction models closer to clinical application.

A SURROGATE MODELING FRAMEWORK FOR INTERPRETING DEEP NEURAL NETWORKS IN FUNCTIONAL GENOMICS

Evan Seitz, Peter Koo, Justin Kinney

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

Understanding the *cis*-regulatory grammars that coordinate how proteins interact to regulate transcription is a major goal in genomics. Deep neural networks (DNNs) applied to this task have greatly enhanced our ability to accurately predict experiments in regulatory genomics. Despite their impressive performance compared to traditional methods in computational genomics, it remains difficult to determine how these networks form their decisions. To address this gap, attribution methods are being increasingly used to gain mechanistic insights underlying DNN predictions. Attribution methods probe the trained DNN to assess the importance of each nucleotide in a sequence to produce an attribution map, which has been shown to visualize known functional motifs and their locations. However, current attribution methods are sensitive to the local function properties learned by the DNN, making identification of functional motifs difficult. Due to the high expressivity of DNNs, there is no guarantee this issue can be resolved by altering the DNN to learn smoother functions amenable to attribution maps.

Instead, we surmise that attribution-based explanations can be made more robust by approximating a larger region of function space anchored at a given sequence of interest with an interpretable surrogate model, for which the parameters provide direct interpretations of variant importance similar to attribution maps. Here we introduce this new surrogate modeling approach into genomics, where it has not yet been explored. Our approach, called SQUID for Surrogate QUantitative Interpretability of Deepnets, is a general framework that leverages interpretable surrogates to quantitatively model the sequence-function relationship learned by any black-box genomic model. We demonstrate our framework across several existing DNNs designed to perform a variety of regulatory genomics prediction tasks. For each of these DNNs, we show that SQUID outperforms existing attribution methods in studies spanning ensembles of high-functioning motifs and genomic sequences. From this comparison, we find that SQUID is able to more robustly characterize the direct effect of motifs and their higher-order interactions on predictions, consistently model larger sequence contexts, identify weaker binding sites that enable opportunities for better annotation, and provide better approximations to variant effect predictions. SQUID provides a leap forward in our ability to decipher the quantitative effects of *cis*-regulatory elements throughout the genome.

DISSECTING THE FUNCTIONAL DRIVERS OF COMPLEX PHENOTYPES BY *IN VIVO* PROTEIN-INTERACTION QTL MAPPING (pi-QTL)

Adrian Serohijos^{1,2}

¹University of Montreal, Biochemistry, Montreal, Canada, ²University of Montreal, Robert-Cedergren Center for Bioinformatics and Genomics, Montreal, Canada

A central goal in genetics is to explain how phenotypic traits are determined by genetic variation and has repercussions on almost all aspects of biology and medicine. Genome-Wide Association Studies (GWAS) and expression quantitative trait loci (eQTL) mapping suggest that the association signals tend to be spread across the genome and include many genes without an obvious connection to the phenotype or disease. Finding the causative genetic determinants of complex traits remains a grand challenge. Here, using 354 inbred yeast strains containing ~12,000 SNPs and 62 diverse pairs of protein-protein interactions, we developed protein-interaction quantitative trait loci mapping (or *in vivo* piQTL) that dissects causal and functional drivers of complex phenotypes that include yeast growth under antifungals (5-fluorocystocince and fluconazole), anti-diabetic drug metformin, and antipsychotic compound trifluoperazine. We found that piQTLs are spread throughout the genome, including protein-coding regions, their promoters and 3'UTRs. However, despite being distributed across the genome, the piQTLs cluster in biochemical functional grouping defined by curated PPI interaction networks and genetic interaction network from double-gene knockouts. Most surprisingly, piQTLs also include a significant fraction of inter-genic regions and non-coding RNAs, especially those classified as SUTs (stable undefined transcripts). Altogether, this study demonstrates that the ultimate cellular consequences of the genotype-phenotype relationship (GPR) are reflected in the protein steady-state abundances, post-translational modifications, and subcellular localization of proteins and protein-protein interactions (PPIs). This study also provides a roadmap for a new type of QTL mapping based on PPI networks that reveals the functional and biochemical drivers of complex phenotypes.

MITOCHONDRIAL HAPLOTYPE AND MITO-NUCLEAR MATCHING DRIVE SOMATIC MUTATION AND SELECTION THROUGH AGING

Isabel M Serrano¹, Misa Hirose², Clint Valentine³, Sharie Austin³, Jesse Salk³, Saleh Ibrahim⁴, Peter H Sudmant^{1,5}

¹University of California, Berkeley, Computational Biology Graduate Group, Berkeley, CA, ²University of Lübeck, Lübeck Institute of Experimental Dermatology, Lübeck, Germany, ³TwinStrand Biosciences, Biotechnology Company, Seattle, WA, ⁴Khalifa University, College of Medicine, Abu Dhabi, United Arab Emirates, ⁵University of California, Berkeley, Integrative Biology, Berkeley, CA

Mitochondrial DNA (mtDNA) is a high copy number genome, creating a population of mitochondrial genomes (mt-genomes) within a cell. As organisms age, the mt-genomes in their somatic cells accumulate mutations; yet, we understand little about the evolutionary mechanisms underlying this process. To discern how mitochondrial haplotype (mt-haplotype) impacts the somatic evolution of the mt-genome, we characterized the mitochondrial mutational landscapes across five mitochondrial ancestries (mt-ancestry) and three tissues at two timepoints in the mouse lifespan. We use a panel of conplastic mouse strains, which are identical in their nuclear genomes but differ by their mt-ancestry, and employ ultra-sensitive Duplex Sequencing to generate a map of mutations with an unprecedented level of depth and accuracy. From these maps, we confirm that the increase in mutation frequency with age is a constant molecular phenotype. We identify haplotype-specific mutational profiles across the mt-genome and find that the Light Strand Origin of Replication is a consistent mutational hotspot, exhibiting a higher mutation frequency than the non-coding D-Loop. We highlight mutational spectra differences between rodents and primates, with rodents containing an excess of reactive oxygen species damage (G>T/C>A) mutations. We show that the accumulation of mutations in the mt-genomes is strongly shaped by positive selection. In particular, we discover a remarkable enrichment of somatic reversion mutations (i.e. a preference for alleles that work to realign mitochondrial and nuclear ancestries), suggesting that mito-nuclear matching is an important source of somatic selection in mt-genomes. Together, our findings demonstrate that somatic mutation and selection actively shape the mt-genome throughout aging.

META-ANALYSIS OF EPIGENETIC AGE IN DOGS MODULATED BY BREED LIFESPAN

Aitor Serres-Armero¹, Matteo Pellegrini², Steve Horvath³, Elaine A Ostrander¹

¹National Institutes of Health, National Human Genome Research Institute, Bethesda, MD, ²University of California Los Angeles, Molecular, Cell and Developmental Biology, Los Angeles, CA, ³Altos Labs, Altos Labs, San Diego, CA

The rate at which an individual ages, also known as biological age, differs from chronological age in that biological age depends on lifespan. It is unclear how accurately methylation can recapitulate biological age in species with narrow lifespan differences such as humans. Dog breeds represent an invaluable resource to study biological age differences, as strict date-of-birth records are kept for breed dogs by registering bodies, and dog breeds display large and reproducible lifespan differences. Consequently, biological age can be estimated and contrasted with methylation profiles for any dog of known age and breed, something unfeasible *pre-mortem* in most other species. Previous analyses of DNA methylation in dogs and wolves have successfully identified loci that accurately track chronological age in multiple dog breeds, but as a result these same loci demonstrate incomplete evidence of differing biological ages across breeds. Here we leverage the three largest, publicly available canine DNA methylation datasets to exhaustively study methylation acceleration or deceleration in long- and short-lived breeds.

We adapt the methylation-based age prediction algorithms to accommodate phylogenetic relations between breeds, non-linearities with respect to age, and modulation by breed standard lifespans. We additionally scan each individual methylation site across datasets for signals of biological age while accounting for all the aforementioned factors. In agreement with the component studies of our meta-analysis, we report an overall absence of biological age markers across different methylation experiments, suggesting that biological age has a modest effect on the methylome as compared to chronological age. Replacing lifespan with weight as a modulator for epigenetic age increases the power to detect biological age differences between breeds. This suggests that methylation may be tracking additional covariates such as body mass, which may themselves be related to biological age. Disentangling the effect size and significance of methylation modulation by lifespan and weight in the dog should inform studies of biological age methylation clocks in humans.

DUX4 DOUBLE WHAMMY: THE TRANSCRIPTION FACTOR THAT CAUSES A RARE MUSCULAR DYSTROPHY ALSO KILLS THE PRECURSORS OF THE HUMAN NOSE

Kaoru Inoue¹, Hamed Bostan², MaKenna R Browne¹, Owen F Bevis¹, Carl D Bortner³, Steven A Moore⁴, Aaron A Stence⁵, Negin P Martin⁶, Shih-Heng Chen⁶, Adam B Burkholder², Jian-Liang Li², Natalie D Shaw¹

¹NIEHS/NIH, Clinical Research Branch, Durham, NC, ²NIEHS/NIH, Integrative Bioinformatics, Durham, NC, ³NIEHS/NIH, Signal Transduction Laboratory, Durham, NC, ⁴University of Iowa Carver College of Medicine, Pathology, Iowa City, IA, ⁵University of Iowa, Hospitals and Clinics, Iowa City, IA, ⁶NIEHS/NIH, Viral Vector Core, Durham, NC

Missense mutations in the gene *SMCHD1*, which encodes a master epigenetic repressor, have been implicated in two seemingly unrelated conditions: a severe and striking craniofacial malformation called congenital arhinia (absent nose) and a late-onset, slowly progressive muscular dystrophy called facioscapulohumeral muscular dystrophy type 2 (FSHD2). In FSHD2, loss of *SMCHD1* repressive activity leads to ectopic expression of the double homeobox 4 (DUX4) transcription factor in muscle tissue, where it acts as a potent toxin. The pathophysiology of arhinia is unknown, however, deep phenotyping studies of patients with arhinia coupled with *Smchd1* expression studies in mouse embryos strongly suggest a primary defect in the nasal placode cells, the progenitors of the human nose. Here we show that upon *SMCHD1* ablation, DUX4 becomes de-repressed in H9 human embryonic stem cells (hESC) as they differentiate toward a placode cell fate. Time-course RNAseq and DUX4 ChIP-seq studies show that DUX4 activates a transcriptional program akin to that in muscle cells and triggers placode cell death. In *SMCHD1* KO hESC, DUX4-mediated cell death can be rescued by re-introduction of wildtype *SMCHD1*, DUX4 shRNA knockdown, and small-molecule pharmacological inhibition of DUX4. Arhina and FSHD2 patient-derived induced pluripotent stem cells (iPSC) both express DUX4 when converted to placode cells and demonstrate variable degrees of cell death, akin to *SMCHD1* KO hESC, suggesting the presence of an environmental disease modifier. We show that HSV-1 may be one such modifier as herpesvirus infection induces DUX4 expression in *SMCHD1* KO hESC and arhinia and FSHD2 iPSC. These studies suggest that arhinia, like FSHD2, is due to compromised *SMCHD1* repressive activity in a cell-specific context and provide evidence for an environmental modifier. While placode cell death has been implicated in other congenital malformations due to genetic (e.g., otic placode in Branchio-Oto-Renal (BOR) syndrome due to mutations in *EYA1*) and/or teratogenic mechanisms (e.g., hypoxia, alcohol, hyperthermia), to our knowledge, arhinia represents the first example of a craniofacial malformation caused by epigenetic de-repression of a toxic protein.

TRANSCRIPTOME-WIDE AND PROTEOME-WIDE ASSOCIATION STUDY OF TOURETTE'S SYNDROME

Sudhanshu Shekhar^{1,2}, Peristera Paschou^{1,2}

¹Purdue University, Department of Biological Sciences, West Lafayette, IN,

²Purdue University, Interdisciplinary Life Sciences (PULSe), West Lafayette, IN

Tourette's Syndrome (TS) is a neurodevelopmental disorder that is characterized by motor and phonic tics. A recent genome-wide association study (GWAS) identified the *FLT3* gene as a genome-wide locus significantly linked to TS. However, determining the biological mechanism and pathways of GWAS signals remains difficult, limiting the understanding of TS. Importantly, the current iteration of GWAS does not identify *FLT3* as a significant locus and thus has an ambiguity regarding the role of *FLT3* in TS. To characterize the effect of genetic variation-mediated gene expression in TS and to understand the biological underpinnings of the disease, we perform a global and unbiased transcriptome-wide and proteome-wide association study (TWAS and PWAS respectively) in the largest cohort of TS patient samples of general European ancestry consisting of 6133 TS cases and 13565 healthy controls using FUSION tool. Our TWAS analysis using an individual tissue-based prediction matrix of gene expression validates that the increased expression of *FLT3* in the dorsolateral prefrontal cortex (DLPFC) is associated with TS, as reported in a recent publication. Additionally, the analysis identifies eleven novel genes whose transcript expression is significantly associated with TS. Next, we performed TWAS analysis using a cross-tissue-based prediction matrix of gene expression and identified *FLT3* and *EP300* to be significantly associated with TS. As transcriptome expression sparingly correlates with proteome expression, we performed PWAS to complement TWAS analysis. Our PWAS analysis, based on protein expression from DLPFC tissue, identified three novel genes. For these genes, genetic variant-mediated change in protein expression is significantly associated with TS. In conclusion, results from our TWAS and PWAS analysis allow us to identify novel genes associated with TS and identify biological pathways that can be validated via biological experimentation to strengthen our analysis.

EASTR: IMPROVING RNA-SEQ ALIGNMENT FOR ACCURATE TRANSCRIPTOME ASSEMBLY AND REFERENCE ANNOTATION CURATION

Ida Shinder^{1,2}, Richard Hu^{2,3}, Hyun Joo Ji^{2,3}, Kuan-Hao Chao^{2,3}, Mihaela Pertea^{2,3,4}

¹Johns Hopkins School of Medicine, Cross Disciplinary Graduate Program in Biomedical Sciences, Baltimore, MD, ²Johns Hopkins University, Center for Computational Biology, Baltimore, MD, ³Johns Hopkins University, Department of Computer Science, Baltimore, MD, ⁴Johns Hopkins School of Medicine, Department of Biomedical Engineering, Baltimore, MD

Accurate reference annotation and alignment are essential for reliable RNA sequencing (RNA-seq) analyses. However, splice-aware aligners, such as the widely used STAR and HISAT2, can introduce previously undetected inaccuracies in spliced alignments between repeated sequences, leading to the inclusion of falsely spliced transcripts in reference annotations. To address this issue, we present EASTR (Emending Alignments of Spliced Transcript Reads), a novel tool that improves genome-guided transcriptome assembly and reference annotation curation. EASTR uses a dual approach to analyze the sequence similarity between intron flanking regions and the frequency of flanking sequences in the reference genome to detect and remove falsely spliced alignments and transcripts from alignment and annotation files.

We evaluated the performance of EASTR on diverse species, including humans, maize, and *Arabidopsis thaliana*, and observed a substantial improvement in both precision and sensitivity of genome-guided transcript assembly, with no trade-off between the two. Specifically, we applied StringTie2 to EASTR-filtered alignment files and found that it significantly enhanced transcript assembly precision, particularly for datasets with a high representation of reads from repeat elements. We also observed an overall increase in sensitivity across various species and experimental designs.

Our evaluation of reference annotations using EASTR uncovered a considerable number of questionable introns and transcripts, particularly in gene families featuring duplicated paralogs, complex repetitive structures, and copy number heterogeneity. EASTR identified numerous introns that overlap with tandem repeat regions and indels, and their length matched the periodicity of the tandem repeat, providing support for the potential incorrect inclusion of such introns in gene catalogs. EASTR also detected erroneous splicing between unannotated transposable elements and tandemly duplicated gene copies, leading to frequent mis-assembly of full-length transcripts included in gene catalogs. Our findings emphasize the critical importance of accurate reference annotation and alignment for genome biology studies and highlight the potential of EASTR for improving reliability of downstream RNA-seq analyses.

CHANGES IN CIRCULATING CELL-FREE DNA AS A BIOMARKER OF IMMUNE RESPONSE TO SHORT-DURATION SPACEFLIGHT

Karolina Sienkiewicz¹, Kirill Grigorev^{1,2}, Namita Damle², Deena Najjar², Sebastian Garcia Medina^{1,2}, JangKeun Kim^{1,2}, Jonathan Foox^{1,2}, Eliah G Overbey^{1,2}, Kelly Blease³, Juan Moreno³, Junhua Zhao³, Bryan Lajoie³, Andrew Altomare³, Semyon Kruglyak³, Ari M Melnick^{4,5}, Jaime Mateus⁶, Christopher E Mason^{1,2,7,8}

¹Weill Cornell Medicine (WCM), Institute for Computational Biomedicine, New York, NY, ²WCM, Department of Physiology and Biophysics, New York, NY, ³Element Biosciences, San Diego, CA, ⁴WCM, Meyer Cancer Center, New York, NY, ⁵WCM, Department of Medicine, New York, NY,

⁶SpaceX, Hawthorne, CA, ⁷WCM, The Feil Family Brain and Mind Research Institute, New York, NY, ⁸WCM, WorldQuant Initiative for Quantitative Prediction, New York, NY

The concentration of cell-free DNA (cfDNA) and its molecular profile are emerging biomarkers with great clinical and research potential, which can provide valuable insights into the dynamic response of organisms to environmental or disease-related stress factors. Here, we focus on the systemic immune system response to physiological stress related to microgravity, radiation exposure, and the other unique environmental conditions of short-duration spaceflight.

As a part of the Space Omics and Medical Atlas initiative, we profiled the cfDNA of a cohort of astronauts from SpaceX's Inspiration4 mission, comparing pre-flight baseline, recovery and longitudinal responses. We present a comprehensive cfDNA analysis pipeline which includes (1) a comparison of sample-wise and chromosome-specific fragment size distribution, (2) an assessment of enrichment in cfDNA fragments originating from different tissues based on estimated nucleosomal footprinting and inferred gene expression, as well as (3) cfDNA variant calling and annotation.

Building on the results from the NASA Twins Study, we inspected the previously reported potential biomarkers of interest for long-duration spaceflights, such as the fraction of mitochondrial cfDNA relative to chromosomal cfDNA in plasma. Comparison of the deconvoluted tissue/cell type of origin profile for circulating cfDNA fragments with either reference tissue-specific expression signatures from the Human Proteome Map or individual astronaut expression profiles from peripheral blood (single-cell RNAseq of PBMC) revealed that the most represented sequences are of hematopoietic origin. We noted an increase in the cfDNA originating from immune cells after return to Earth which may reflect a delayed spaceflight-related immunological response. Moreover, our preliminary results highlight the non-invasive monitoring potential of cfDNA may be extended toward circulating epigenomic biomarkers, which could also be extremely valuable in a disease setting where the progression is associated with rapid epigenetic changes.

GENE DESERTS HARBOR NONCODING FUNCTIONS CRITICALLY REQUIRED FOR NORMAL DEVELOPMENT

Neil Slaven*, Yiwen Zhou*, Fabrice Darbellay, Jennifer Akiyama, Ingrid Plazier-Frick, Catherine Novak, Momoe Kato, Axel Visel, Len Pennacchio

Lawrence Berkeley National Lab, Environmental Genomics and Systems Biology, Berkeley, CA

The distribution of genes across the human genome is not random. Regions enriched in protein-coding genes are interspersed with gene-sparse regions. The extreme end of this distribution pattern is represented by gene deserts, vast noncoding regions devoid of genes. The 545 largest deserts, each spanning >640 kbp, collectively make up ~25% of the genome. The location and size of gene deserts is well conserved across mammalian genomes and they contain substantial numbers of non-coding conserved sequences, suggesting that their size and sequence content are subject to selective constraint. Nonetheless, the general functional requirement for gene deserts remains poorly explored due to the challenge of creating large germline deletion knockout mouse models. Here, we present initial findings from the Gene Desert Deletion Project, which aims at the systematic deletion of the largest gene deserts in the mouse genome, followed by functional assessment. To date, we have deleted 10 deserts, together covering 16Mbp (0.6%) of the mouse genome using CRISPR targeting. These deserts are orthologous to deserts in the human genome and collectively harbor over 2,500 evolutionarily conserved regions, with an average of 83 cis-regulatory elements predicted from epigenomic signatures in relevant developmental mouse tissues. Risk SNPs implicated in human traits through genome-wide association studies are found scattered across the human orthologous gene deserts with an average of 122 per desert. Rigorous phenotyping of homozygous knockout mice for individual deserts revealed that 7 of the 10 deserts are required for development or viability. This included 4 cases of pre- or perinatal lethality. Phenotypes for two cases resembled those reported for knockouts of genes neighboring the desert, supporting that the embedded functions include distant-acting regulatory sequences. For example, deletion of a 912kbp gene desert flanked by the Dmrt2 gene caused major skeletal abnormalities resembling those of Dmrt2 gene knockout mice. Taken together, our results demonstrate that the conserved size and sequence content of gene deserts coincides with critical biological functions. Our findings support that gene deserts must be carefully considered as potential reservoirs of deleterious mutations in human sequencing studies. Phase I of the Gene Desert Deletion Project has demonstrated that 7 of the largest mammalian gene deserts are required for organismal development. Continued systematic exploration of deserts will provide a valuable resource for the interpretation of human mutations.

SURVEYING SPECIFIC & SHARED RESPONSES OF HUMAN ISLETS TO T2D-ASSOCIATED STRESSORS

Eishani K Sokolowski^{1,2}, Redwan M Bhuiyan^{1,2}, Romy Kursawe², Michael L Stitzel², Duygu Ucar²

¹University of Connecticut Health Center, Department of Genetics and Developmental Biology, Farmington, CT, ²The Jackson Laboratory For Genomic Medicine, Farmington, CT

Endoplasmic reticulum (ER) and inflammatory stress in islets is linked to Type 2 Diabetes (T2D). We hypothesize that a subset of non-coding T2D-associated genetic variants (T2D SNPs) modulate these pathophysiologic responses in human islets. To test this, we compared chromatin accessibility and gene expression in human islets before and after exposure to ER stress (thapsigargin) or pro-inflammatory cytokines (IL-1 β +IFN- γ) to identify specific and shared stress-responsive cis-regulatory elements (CREs) and their target genes. Approximately 13% of CREs (n=14,966) and 30% of genes (n=5,131) are modulated by ≥ 1 stressor (ER stress-specific: ~41% CREs & ~33% genes, inflammation-specific: ~52% CREs & ~42% genes, shared: ~7% CREs & ~25% genes), implying that changes in accessibility and expression in islets are stress-specific. For example, ER stress-specific genes (ex: *DDIT3* & *ATF4*) are involved in unfolded protein response, and ER stress-specific CREs enrich for 157 transcription factor (TF) binding sites (ex: CHOP & ATF4). Many genes nearest to stress-responsive CREs are also stress-responsive genes (ER stress-specific: n=596, inflammation-specific: n=827, shared: n=91) suggesting a context-specific modulation of gene expression by CREs. Of these CREs, 92 (ER stress: n=53, inflammation: n=39) harbor 114 T2D SNPs, of which 19 T2D SNPs become accessible only upon ER stress. For example, rs6444081 (A>G) overlaps an ER stress-specific CRE and resides downstream of *ETV5* (ER stress-responsive gene). The T2D risk allele (G) is predicted to be bound by ER stress-specific TF ETV1. In a separate cohort, *ETV5* expression is higher in T2D vs non-T2D beta cells, suggesting that *ETV5* is a putative target gene of rs6444081 and that its activity is genetically modulated in islets. Our results nominate T2D SNPs that we propose modulate stress-specific islet responses. Further study of these CREs and their target genes should help to functionally compartmentalize T2D risk variants.

A UNIFIED FRAMEWORK OF REALISTIC IN SILICO DATA GENERATION AND STATISTICAL MODEL INFERENCE FOR SINGLE-CELL AND SPATIAL OMICS

Dongyuan Song¹, Qingyang Wang², Jingyi Li^{1,2,3,4}

¹UCLA, Bioinformatics Interdepartmental Ph.D. Program, Los Angeles, CA, ²UCLA, Department of Statistics, Los Angeles, CA, ³UCLA, Department of Computational Medicine, Los Angeles, CA, ⁴Harvard University, Radcliffe Institute for Advanced Study, Boston, MA

In the single-cell and spatial omics field, computational challenges include method benchmarking, data interpretation, and in silico data generation. To address these challenges, we propose an all-in-one statistical simulator, scDesign3, to generate realistic single-cell and spatial omics data, including various cell states, experimental designs, and feature modalities, by learning interpretable parameters from real datasets. Furthermore, using a unified probabilistic model for single-cell and spatial omics data, scDesign3 can infer biologically meaningful parameters, assess the goodness-of-fit of inferred cell clusters, trajectories, and spatial locations, and generate in silico negative and positive controls for benchmarking computational tools.

T1K: EFFICIENT AND ACCURATE KIR AND HLA GENOTYPING WITH NEXT-GENERATION SEQUENCING DATA

Li Song¹, Gali Bai², X. Shirley Liu³, Bo Li⁴, Heng Li^{5,6}

¹Dartmouth College, Department of Biomedical Data Science, Lebanon, NH, ²University of California, Santa Cruz, Biomolecular Engineering Department, Santa Cruz, CA, ³GV20 Therapeutics, Cambridge, MA,

⁴University of Texas Southwestern Medical Center, Lyda Hill Department of Bioinformatics, Dallas, TX, ⁵Dana-Farber Cancer Institute, Department of Data Science, Boston, MA, ⁶Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Killer immunoglobulin-like receptor (KIR) genes and human leukocyte antigen (HLA) genes play important roles in innate and adaptive immunity. For example, KIR is a family of regulatory receptors presented on NK cells and a subset of T cells. These receptors are inhibited or activated via binding to their cognate ligands, which include the HLA molecules. Both KIR and HLA are highly polymorphic and cannot be genotyped with standard variant calling pipelines. Compared with HLA genes, many KIR genes are similar to each other in sequences and may be absent in the chromosomes. Therefore, while many tools have been developed to genotype HLA genes using common sequencing data, none of them works for KIR genes. Even the specialized KIR genotypers could not resolve all the KIR genes.

Here we present T1K, a novel computational method for the efficient and accurate inference of KIR or HLA alleles from RNA-seq, whole genome sequencing or whole exome sequencing data. T1K jointly considers alleles across all genotyped genes, so it can reliably identify present genes and distinguish homologous genes, including the challenging *KIR2DL5A/KIR2DL5B* genes. In our evaluations, T1K obtains 95% accuracy in genotyping KIR genes. T1K's model also benefits HLA genotyping, where T1K achieves more than 99% accuracy and is the most accurate method in benchmarks. Moreover, T1K can call novel single nucleotide variants and process single-cell data. Applying T1K to tumor single-cell RNA-seq data, we found that *KIR2DL4* expression was enriched in tumor-specific CD8+ T cells.

T1K is free and open source at <https://github.com/mourisl/T1K>, and its versatile framework will contribute to KIR, HLA and other polymorphic gene studies in the future.

ENHANCER REPRESSION IN GENE EXPRESSION FINE-TUNING

Wei Song, Ivan Ovcharenko

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

The spectrum of gene expression in a particular cell type is orchestrated via the complex binding of DNA-binding proteins to enhancers. An interplay between activator and repressor transcription factor binding to enhancers modulates enhancer activity resulting in fine-tuning of target gene expression. We identified a class of active enhancers predominantly populated by the binding sites of repressors, which we named "repressed enhancers". We demonstrate that the activity of repressed enhancers is strongly diminished in comparison to "regular" enhancers, indicative of the intricate modulation of the regulatory activity of these elements.

We built an accurate deep learning model of repressed and regular enhancers active in HepG2 cells. This model provides a detailed breakdown of active sites within the DNA sequence of enhancers and allows quantification of the mutational impact on the binding affinity of transcription factors bound to enhancers. A large fraction (20%; 8,457/42,611) of HepG2 enhancers has been characterized as repressed enhancers according to our model. These repressed enhancers feature a 4.2-fold enrichment in repressor-vs-activator binding sites. An evolutionary sequence analysis demonstrates that while both repressor and activator binding sites are evolving under negative selection, repressed enhancers are actively gaining repressed binding sites in a process of constant evolutionary fine-tuning of their regulatory activity. This results in a significantly greater cell-type-specificity of repressed enhancers than regular enhancers, indicating their role in establishing a refined expression pattern of genes regulated by repressed enhancers.

The target genes of repressed enhancers commonly employ a combination of regular and repressed enhancers and feature distinct cellular activities. These are predominantly developmental genes, which have been well-conserved throughout the evolution of vertebrates. Loss-of-function of such genes is commonly associated with pronounced negative impact on the species phenotype and embryonic viability. Thus, the gradual evolution of their regulatory architecture via modulation of repressed enhancer activity represents an evolutionary path that vertebrate species have been exploring in the adaptation of vital cellular mechanisms.

Overall, this study highlights the complex nature of gene regulation with repressed enhancers fine-tuning cell-type-specific gene expression, which helps expand our understanding of the regulatory elements of the genome and their role in human health and disease.

TRANSCRIPTIONAL ATTENUATION OF CNV AMPLIFICATION AND CONSEQUENCES FOR GENE REGULATORY NETWORKS

Pieter Spealman, Grace Avecilla, Julia Matthews, David Gresham

New York University, Center for Genomics and Systems Biology, New York, NY

Copy number variants (CNVs) are duplications and deletions of genomic regions that are known to contribute to evolutionary adaptation but that can also be deleterious. While the consequences of amplifying individual genes or whole chromosomes have been studied extensively, much less is known about the genetic and functional effects of CNVs. Here, we investigated *Saccharomyces cerevisiae* (yeast) strains with adaptive CNVs generated in the course of long-term evolution experiments.

We fully resolved each CNV structure using hybrid assembly of Illumina short-read and Nanopore long-read sequencing. This allowed us to accurately determine the nature of each CNV and the copy-number of all genes.

Next, we performed RNAseq on both the evolved and ancestral strains and found copy-number strongly correlated with expression, consistent with previous studies of aneuploidy. However, we identified numerous CNV amplified genes with lower expression than expected. Comparing genes with higher expression in the evolved strain, we found CNV amplified genes had significantly lower expression per copy number than their non-CNV counterparts (Mann-Whitney U, p-value < 0.01, per strain median attenuation 13-34% less than expected). This attenuation of over-expression has previously been reported for aneuploid strains, but at post-transcriptional levels of regulation.

One potential source of transcriptional attenuation is dis-regulation generated by altered ratios of transcription factor (TF) supply and demand. We found CNV amplification of TF-encoding genes in multiple strains. Hypothetically, these amplifications would produce excess TF proteins that generate downstream dis-regulation of their targets. However, analysis of these amplified TFs and their targets suggest that this knock-on effect is not occurring, potentially due to compensatory changes in other coregulating TFs. This difference from over-expression studies is probably due to the evolved nature of these strains and the plasticity of gene regulatory networks.

Taken together our results suggest that DNA amplification resulting from CNVs may be a major driver of gene expression at the level of transcription but that additional components such as transcriptional attenuation and the buffering capacity of gene regulatory networks act to mitigate the consequences of these amplifications.

THE EVOLUTIONARY HISTORY OF 17Q21.31 STRUCTURAL HAPLOTYPES IN ANCIENT AND MODERN HUMANS

Samvardhini Sridharan¹, Peter H Sudmant²

¹PhD Candidate, Molecular and Cell Biology and Center for Computational Biology, Berkeley, CA, ²Assistant Professor, Center for Computational Biology and Integrative Biology, Berkeley, CA

Structural variation accounts for a large proportion of genetic variation within human populations and between humans and other primates. One type of structural variation is inversions – segments of the genome that are reversed end-to-end. There are about 150 large inversions segregating in human populations, of which one extraordinary example is the 17q21.31 inversion locus.

The 17q21.31 locus spans over 900 kilobases and comprises ten genes, including *MAPT* and *KANSL1*, which have medical relevance. There are at least seven major inversion haplotypes present in global populations, each present at markedly different frequencies within geographic regions and even within subpopulations. Some of these haplotypes have also been identified as being under selection. Four haplotypes exist in the direct orientation (H1), and three exist in the inverted orientation (H2). Each of these haplotypes maintain a complex duplication architecture in addition to its inversion status. The enormous amount of structural diversity present at this one locus provides a unique system to study the impact of extended linkage disequilibrium and recombination suppression.

Here, we report extensive and surprising patterns of temporal and geographic variation at the 17q21.31 locus. Notably, we find that while the inverted haplotype (H2) is the ancestral haplotype in humans, it is the less common haplotype in both archaic and modern day human samples. We examine human and primate variation in large datasets such as the 1000 Genomes Project, Human Genome Diversity Panel, Simons Genome Diversity Project, UK Biobank, and Stone Age Eurasians. We map the distributions of the 17q21.31 haplotypes across the world, in both modern and ancient populations. Finally, we hypothesize that large and structurally complex regions of the genome may differentially accumulate more deleterious mutations due to recombination suppression. We quantify the relative mutation load in different haplotypes to determine if recombination suppression is influencing deleterious variation at 17q21.31. Together, these results help evaluate the role chromosomal rearrangements play in evolution, diversity, disease, and fitness.

A MACHINE LEARNING APPROACH TO IDENTIFY
FUNCTIONALLY RELEVANT ENDOGENOUS mRNA TARGETS OF
piRNAs IN *C. ELEGANS*.

Margaret R Starostik¹, Charlotte P Choi¹, Rebecca J Tay¹, Lars K Benner¹,
Brooke E Montgomery², Taiowa A Montgomery², Michael C Schatz^{1,3,4},
John K Kim¹

¹Johns Hopkins University, Biology, Baltimore, MD, ²Colorado State University, Biology, Fort Collins, CO, ³Johns Hopkins University, Computer Science, Baltimore, MD, ⁴Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Small RNAs – short, noncoding RNAs – are critical regulators in animal physiology and disease that silence gene expression by complementary base pairing interactions to control mRNA stability/translation and epigenetic modifications. Small RNA-mediated silencing pathways are evolutionarily conserved, and the largest class of small RNAs comprise Piwi-interacting RNAs (piRNAs). Studies in fly established that piRNAs silence transposons and are expressed in a sex-specific manner; piRNAs are essential for genome integrity and germline development. However, most piRNAs in mammals (mouse: 83%; human: 78%) do not map to transposons, a phenomenon also observed in worm. Recent evidence suggests that mammalian and worm piRNAs silence endogenous mRNAs; however, only a few such piRNA targets have been identified. Although understanding how sex-specific piRNA expression regulates germline gene expression is paramount to advancing our knowledge of piRNA-mediated gene regulation, endogenous mRNA targets of piRNAs remain largely unknown due to challenges in profiling piRNA:mRNA interactions.

Using *C. elegans* as a model, we leveraged our recent discovery of sex-specific piRNA biogenesis factors, SNPC-1.2 and SNPC-1.3, and performed small RNA-seq in wild-type and mutant animals for these piRNA biogenesis factors to characterize piRNA expression during germline development. In parallel, we performed mRNA-seq as a strategy for piRNA target analysis by identifying transcripts that were upregulated in the absence of specific subsets of piRNAs. We developed a support vector machine model for genome-wide piRNA target prediction by integrating these datasets with several public small RNA-seq, mRNA-seq, and piRNA target binding data. Three major categories of piRNA targeting features were investigated for algorithm development: (1) position-based features that define the sequence specificity for the piRNA:mRNA interaction, (2) structural features based on secondary alignment, and (3) expression-based features. Comparative analysis indicates that our approach overcomes the limitations of existing rule-driven methods that are based on piRNA:mRNA sequence complementarity and improves computational prediction of functionally relevant endogenous piRNA targets, thereby providing new insights into the biological roles of piRNAs beyond transposon silencing.

A METHOD TO SEQUENCE TRUE FULL-LENGTH CAPPED RNA

Jamie Auxillois^{*1,2}, Arnaud Stigliani^{*1,2}, Albin Sandelin^{1,2}

¹Section for Computational and RNA biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark, ²Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark

RNA-Seq has become a popular tool for the study of gene expression and regulation, due to its ability to provide highly accurate, comprehensive, and quantifiable information. However, most RNA-Seq protocols have several limitations. First, they are limited in their ability to sequence full-length transcripts since they rely on short reads. This can result in the loss of crucial information such as alternative splicing events (which can alter protein structure significantly) as well as isoform identification and quantification. In addition, they do not capture the genes transcription start sites (TSSs) accurately, which is essential when it comes to finding the genes promoters. This is key to better pick out isoforms, shed a light on their function, and understand the dynamics of their expression.

The last two decades have seen the development of new techniques tackling these two limitations. While Oxford Nanopore Technologies sequencing platforms can sequence full-length transcripts, the SLIC-CAGE has settled as a state-of-the art method to locate TSS and quantify their usage. However, no technique has yet managed to combine these two aspects.

Integrating the SLIC-CAGE protocol and the Oxford Nanopore Technologies we introduce a new approach allowing to sequence full-length fragments from the 5' cap to the 3'end. Additionally, this protocol does not rely on selecting poly-Adenylated RNA, making it a prime method to characterize the polyA-tail and study its influence on RNA stability and degradation.

THE SMALL RNA TRANSCRIPTOME AND ITS GENETIC REGULATION ACROSS HUMAN TISSUES

Tim Coorens¹, Petar Stojanov¹, Juan Carlos Fernandez del Castillo¹, Scott Steelman^{1,2}, Sarah Young^{1,3}, Chad Nussbaum^{1,2}, Gad Getz^{1,4,5}, Kristin Ardlie¹, François Aguet⁶

¹Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA,

²Cellarity, Somerville, MA, ³Applied Invention, Cambridge, MA, ⁴MGH, Cancer Center and Dept. of Pathology, Boston, MA, ⁵Harvard Medical School, Boston, MA, ⁶Illumina, Inc., AI Laboratory, San Diego, CA

Population-based analyses mapping associations between genotypes and gene expression have identified expression quantitative trait loci (eQTLs) for almost all genes, enabling the identification of molecular mechanisms underlying genetic associations with complex traits and diseases. However, only a minority of complex trait associations have been linked to QTLs, partly due to limited power to detect regulatory effects in the relevant cell type, context, or molecular phenotype. Notably, standard RNA-sequencing protocols exclude small RNAs and preclude the study of thousands of small noncoding RNAs with essential roles in the post-transcriptional regulation of gene expression.

Here, we present the characterization of small RNAs across 16,814 samples, 47 tissue sites and 978 donors in the GTEx Project, including microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), transfer RNAs, small nuclear RNAs, small nucleolar RNAs, Y RNAs, and others. We used personal transcriptomes and reference databases for alignment and applied stringent quality controls to minimize artifacts related to short sequences (~20-30nt), resulting in the quantification of 41,458 small RNAs. We used supervised classification to identify putative novel RNAs not present in references, and detected 57 novel high-confidence miRNAs. We observed strong tissue specificity for miRNAs, with other small RNAs showing homogenous expression across tissues. Small RNA expression is highly correlated with surrounding mRNAs or lncRNAs, suggesting shared transcriptional regulation. We mapped QTLs in cis and trans, identifying 100s to 1000s of cis-eQTLs for each RNA species. In line with tissue specificity of miRNA expression, miRNA cis-eQTLs also show a higher degree of tissue specificity than mRNA cis-eQTLs. For miRNAs and their predicted target genes, we used colocalization and mediation analyses to study how cis-regulatory effects on miRNA expression propagate to their target genes, in turn modulating their expression, and show how these effects link to complex trait associations.

In summary, we provide the largest resource of small RNA diversity and their eQTLs across human tissues. We demonstrate the importance of characterizing the full spectrum of small RNAs that play critical roles in the regulation of gene expression, including in development and disease.

ABM: BENCHMARKING BIOINFORMATICS TOOLS

Keith Suderman¹, Enis Afgan¹, Nuwan Goonasekera², Michael Schatz¹

¹Johns Hopkins University, Biology, Baltimore, MD, ²University of Melbourne, Bioinformatics, Melbourne, Australia

ABM, the Automated Benchmarking Tool, is an opinionated Python library and command-line tool developed to automate Galaxy benchmarking experiments on cloud providers, such as Amazon Web Services and Google Cloud Platform. It is designed to help researchers better understand how job parameters affect resource requirements and associated costs for bioinformatics tools.

ABM captures benchmark configurations as YAML files, which enable researchers to run repeatable experiments across different cloud providers and Galaxy configurations. The library handles the full lifecycle of an experiment from cluster creation, software installation, server configuration, workflow runs, data collection to resource cleanup.

ABM allows researchers to easily vary the size of inputs, number of CPUs, and amount of memory available, providing valuable insights into the resource requirements of different bioinformatics tools. For example, ABM can help researchers identify which cloud instance type is the most cost-effective for a specific tool by comparing CPU to memory ratios. ABM is highly flexible and can be easily configured to benchmark different bioinformatics tools and cloud providers.

The use of ABM has several benefits for researchers, including making costs more predictable, ensuring efficient use of compute resources, and better managing researcher expectations. ABM can also simplify resource allocation decisions, as it allows researchers to categorize tools based on their resource requirements.

In summary, ABM is a valuable tool for bioinformatics researchers who use cloud resources to run computationally intensive experiments. It provides a flexible and automated approach to benchmarking bioinformatics tools, enabling researchers to gain insights into resource requirements and associated costs. By streamlining the benchmarking process, ABM can help researchers make informed decisions about resource allocation and better manage their cloud spending.

SINGLE-CELL eQTLs MAPPING IN BRAIN CELL TYPES REVEAL CONTEXT SPECIFIC GENETIC REGULATION AND IMPLICATIONS IN AD GENETICS

Na Sun^{1,2}, Yongjin Park^{1,2}, Carles Boix^{1,2}, Lei Hou^{1,2}, Xushen Xiong^{1,2}, Yosuke Tanigawa^{1,2}, Xikun Han^{1,2}, Manolis Kellis^{1,2}

¹Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Lab, Cambridge, MA, ²The Broad Institute of Harvard and MIT, Cambridge, MA

Genetic variants can impact brain gene expression patterns, and contribute to diverse brain neurodegenerative and psychiatric disorders, acting through diverse brain cell types and states. Here, we leverage single-nucleus RNA-seq data from post-mortem human prefrontal cortex samples across 10 major cell types, 57 cell subtypes, 2.3 million nuclei, and 427 individuals to discover 362k fine-mapped cis-acting single-cell expression quantitative trait loci (sc-eQTLs), linked to 9,608 genes (eGenes) in 400k eQTL-eGene pairs. Despite extensive eQTL calling in bulk brain samples in GTEx, we find that ~30% of our eQTL links are newly-detected, and these are enriched in cell-type-specific eQTLs. We also find that chromatin accessibility regions of each brain major cell type can explain cis-QTLs in a cell type-dependent manner. We use Independent component analysis (ICA) to discover 161 co-expressed modules (encompassing 16,409 genes) across combinations of cell types and individuals, and find that 56 of them are significantly enriched in eGenes, suggesting that genetic variants have broad effects on gene expression. We use motifs in eQTL SNPs to predict upstream transcription factors (TFs) for each module, which significantly regulate expression of target genes in the module. We identify trans-acting eQTLs for genes in our modules without enrichment in eGenes if the upstream TFs of these genes have cis-eQTLs. For example, GLI2 has cis-eQTLs in astrocytes and regulates the expression of ASRGL1 by the effect of trans-eQTLs. We investigate condition-specific fine-mapped cis-eQTLs for Alzheimer's disease (AD) and APOE ε4 genotype, and find several genes, including CCR10 and CCRL2, are specific in APOE ε4 carriers in microglia, suggesting the conditional specificity of the effect of genetic variants on gene expression. We perform a transcriptome-wide association study (TWAS) to identify significant cell-type specific expression-AD associations, and find that TWAS genes are significantly enriched in AD-differentially expressed genes (adDEGs) in microglia and oligodendrocytes, providing the potential disease-driving cell types and targets in therapeutics. Overall, our study provides an unprecedented resource of genetic regulation on gene expression in the brain at cell type resolution, uncovers the potential drivers mediating the regulation in a cell type specific manner systematically, and exemplifies the genetic regulatory circuits that may cause brain diseases.

HIGH INTRASPECIES ALLELIC DIVERSITY IN PLANT IMMUNE RECEPTEORS IS ASSOCIATED WITH DISTINCT GENOMIC AND EPIGENOMIC FEATURES

Chandler A Sutherland¹, Daniil M Prigozhin², J. G Monroe³, Ksenia V Krasileva¹

¹University of California, Berkeley, Department of Plant and Microbial Biology, Berkeley, CA, ²Lawrence Berkeley National Laboratory, Molecular Biophysics and Integrated Bioimaging Division, Berkeley, CA,
³University of California, Davis, Department of Plant Sciences, Davis, CA

Plants, lacking the adaptive immune systems of vertebrates, rely exclusively on germline-encoded innate immunity to combat pathogens. Nucleotide-binding, leucine-rich repeat receptors (NLRs) are the intracellular sensors of the plant immune system, detecting pathogen-secreted disease proteins and initiating defense responses. A subset of highly variable NLRs (hvNLRs) show a high degree of intraspecies diversity, while their low variability paralogs (non-hvNLRs) are conserved between ecotypes. At the population level, hvNLRs are hypothesized to act as reservoirs of diversity for future pathogen effectors, while non-hvNLRs may retain successful binding sequences. Our investigation of the genomic features of NLRs in *Arabidopsis* has revealed that hvNLRs show higher expression, less gene body cytosine methylation, and closer proximity to transposable elements than non-hvNLRs. We are expanding this work to include maize NLRs across several ecotypes to investigate the conservation of these trends across species. How non-hv and hvNLRs maintain distinctive genomic features, and how these features may be driving their observed allelic diversity, remain central questions in the evolution of innate immune system receptors. Our findings will serve as a starting point for the investigation of the mechanisms that promote diversity generation in these rapidly evolving genes.

TRANSCRIPTOME-WIDE META-ANALYSIS OF CODON USAGE IN *ESCHERICHIA COLI*

Anima Sutradhar¹, Jonathan Pointon², Christopher Lennon², Giovanni Stracquadanio¹

¹University of Edinburgh, School of Biological Sciences, Edinburgh, United Kingdom, ²FUJIFILM Diosynth Biotechnologies, Belasis Ave, Stockton-on-Tees, Billingham, United Kingdom

The degeneracy of the genetic code allows for a multitude of synonymous codons to be translated into the same amino acid. Driven by various evolutionary forces, synonymous codons display an inherent non-random distribution, called codon usage bias, and can differ at the gene, genome and organism level across species. Such preferential codon usage has been observed in Bacteria and Eukarya, whereby highly expressed genes exhibit stronger selection for certain codons as compared to lowly expressed genes. Knowledge of preferential codon usage across species has also been exploited by the biotechnology industry, where recombinant proteins are back-translated to DNA by selecting codons to maximise transcription and yield. However, obtaining accurate and representative codon bias estimates requires the identification of highly expressed genes and their codon variation across samples and conditions. To do that, we developed Codon Usage Bias from RNA-sequencing (CUBseq), a fully automatic meta-analysis pipeline to build highly expressed gene panels and estimate codon usage biases from RNA sequencing (RNA-Seq) experiments.

Here, we used CUBseq to estimate codon usage bias in *Escherichia coli* using RNA sequencing data from 6,763 samples across 72 strains. We found a set of 115 highly expressed genes in our dataset, with negligible variation across strains, suggesting codon usage to be stable across different strains. Then we compared our codon usage bias estimates to the widely used genome-derived Kazusa and CoCoPUTs codon usage tables, where we found significant variations across several codons, suggesting that the transcriptome plays an important role in influencing codon preference. Overall, CUBseq provides a novel and robust method for transcriptome-based codon usage analysis.

NEOTELOMERES AND TELOMERE-SPANNING CHROMOSOMAL ARM FUSIONS IN CANCER GENOMES REVEALED BY LONG-READ SEQUENCING

Kar-Tong Tan^{1,2,3}, Michael K Slevin^{1,4}, Mitchell L Leibowitz^{1,2,3}, Max Garrity-Janger^{1,2,3}, Heng Li^{5,6}, Matthew Meyerson^{1,2,3,4}

¹Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, ²Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, ³Harvard Medical School, Department of Genetics, Boston, MA,

⁴Dana-Farber Cancer Institute, Center for Cancer Genomics, Boston, MA,

⁵Dana-Farber Cancer Institute, Department of Data Science, Boston, MA,

⁶Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Alterations in the structure and location of telomeres are key events in cancer genome evolution. However, previous genomic approaches, unable to span long telomeric repeat arrays, could not characterize the nature of these alterations. Here, we applied both long-read and short-read genome sequencing to assess telomere repeat-containing structures in cancers and cancer cell lines. Using long-read genome sequences that span telomeric repeat arrays generated by both PacBio HiFi and Nanopore sequencing of cancer cell lines, we defined four types of telomere repeat variations in cancer cells: neotelomeres where telomere addition heals chromosome breaks, chromosomal arm fusions spanning telomere repeats, fusions of neotelomeres, and peri-centromeric fusions with adjoined telomere and centromere repeats. We then analyzed short-read sequences from more than 2600 whole genomes of primary tumors from the Pan-Cancer Analysis of Whole Genomes study to assess the frequency of somatic neotelomere and telomere-spanning fusion alterations across 38 tumor types. These results provide a framework for systematic study of telomeric repeat arrays in cancer genomes, that could serve as a model for understanding the somatic evolution of other repetitive genomic elements.

MODELING GENE EXPRESSION THROUGH PROXIMAL AND DISTAL SEQUENCE ELEMENTS USING DEEP LEARNING

Shushan Toneyan, Peter Koo

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

Sequence-to-function deep neural networks have shown impressive performance at predicting gene expression from DNA sequence inputs. However, their inability to capture long-range interactions has recently come under scrutiny. This is surprising as these models are typically designed to take long DNA segments as input (~200kb–1Mb) and incorporate transformer layers, which should provide a strong inductive bias to learn long-range interactions. This is a major oversight of current expression models given the important role of enhancers and other distal elements in gene expression. Evidently, there is a gap in our expectations of how inductive biases (via modeling choices) should learn long-range interactions and what happens in practice, learning shortcuts that focus only on local sequence signals, which turns out to make decent predictions. To address this gap, here we perform a systematic analysis to compare various modeling and training strategies to reveal the factors that influence the extent to which proximal and distal regulatory elements are taken into account by deep neural network models. Importantly, we also explore the contribution of adding an auxiliary training task of predicting 3D genome contact maps (via Hi-C and micro-C), which should provide a stronger incentive to learn long-range interactions. This work serves to guide the development of sequence-to-gene expression models that overcome the current limitations (i.e., make predictions based on the promoter as well as enhancer sequences) and combine different data modalities (i.e., chromatin architecture and epigenetic information), laying out a path towards a more complete model of gene expression.

AGE-RELATED ACCUMULATION OF *DE NOVO* INDELS IN MITOCHONDRIAL DNA OF MICE AND MACAQUES

Edmundo Torres-Gonzalez^{1,2}, Barbara Arbeithuber^{1,3}, Kateryna D Makova¹

¹Penn State University, Department of Biology, University Park, PA, ²Penn State University, Huck Institute of the Life Sciences, University Park, PA,

³Johannes Kepler University Linz, Experimental Gynaecology, Obstetrics and Gynaecological Endocrinology, Linz, Austria

Mitochondria are becoming the centerpiece of many scientific studies due to their central role in key cellular processes and links to human disease. More than 200 human genetic diseases are caused by mutations in mitochondrial DNA (mtDNA), but how they accumulate with age remains unknown. Because mtDNA is present in hundreds to thousands of copies in each cell, *de novo* mtDNA mutations present themselves as heteroplasmies—with low allele frequency of mutant variants. This creates a challenge in detecting mtDNA mutations with conventional sequencing techniques, which usually have high error rates. Duplex sequencing permits the detection of mutations at very low frequencies with high accuracy. Specifically, by barcoding before sequencing, reads generated from the same genomic fragment can be used to create a consensus sequence, thereby diminishing sequencing and PCR errors. Whereas this technique has been previously used to study point mutations, insertions and deletions (indels) in mtDNA have remained understudied. Here, we studied the age-related accumulation of mtDNA indels by applying highly accurate Duplex sequencing to 221 mouse and 239 macaque somatic and germline samples. We observed an increased indel frequency in younger vs. older animals (significant in all tissues except macaque oocytes). In somatic tissues, the highest fold differences in mutation frequency were seen in mouse liver even though the difference in age groups span just 9 months (compared to 20 years in macaques). Strikingly, in both species, oocytes had lower fold differences in indel frequency than somatic tissues, suggesting the existence of protective mechanisms that allow reproduction later in life. The D-loop region, a non-coding control region in mtDNA known to have a high mutation rate, showed an increased indel frequency compared to the rest of the genome (significant for all tissues in macaques, and in mouse oocytes and brain). Our study of model organisms significantly advances our understanding of mtDNA indel mechanisms and informs human aging and diseases caused by mtDNA indels.

SYSTEMATIC INTERPRETATION OF GENETIC VARIANTS THAT DISRUPT CTCF BINDING SITES

Colby Tubbs¹, Mary Lauren Benton², Evonne McArthur¹, John Capra³, Douglas Ruderfer¹

¹Vanderbilt University, Vanderbilt Genetics Institute, Nashville, TN, ²Baylor University, Department of Computer Science, Waco, TX, ³University of California at San Francisco, Department of Epidemiology and Biostatistics, San Francisco, CA

Background: CCCTC-binding factor (CTCF) regulates 3D genome organization by binding DNA at tens of thousands of sequence motifs throughout the genome.

Genetic variation in these CTCF Binding Sites (CBSs) have clinical impact; their disruption occurs at high frequency in cancers and has been linked to the etiology of rare disease. However, while many variants affecting CBSs exist, only a small number are characterized, and few approaches exist to functionally interpret this class of variation. Here, we hypothesize that modulating a CBS's affinity for CTCF will disrupt its function in the genome and thus will be a target of negative selection.

Methods: To define CBSs, we identified all significant sequence matches to the 19 bp canonical CTCF binding motif (JASPAR, MA0139.1) by scanning the hg38 reference genome. To annotate each CBS with the likelihood that it is bound by CTCF, we integrated known binding sites from the UniBind database with predicted affinity based on the sequence's precision-weight-matrix (PWM) score. For each single nucleotide variant (SNV) overlapping a CBS in gnomAD (v3.1.2), we annotated its impact by taking the difference in likelihood of CTCF binding between the reference and alternate alleles. We call this measure Δ CBS Activity. To assess the utility of our approach in prioritizing CBS variants, we characterized the relationship between Δ CBS Activity, nucleotide conservation between species (GERP), and the mutability adjusted proportion of singletons (MAPS) metric. For additional context, we calculated MAPS for 100,000 gnomAD SNVs annotated as "missense".

Results: We calculated Δ CBS Activity for 2.7 million SNVs (gnomAD) overlapping any CBS ($n=657,593$). Most ($n=2,078,786$) were predicted to have minimal impact (Δ CBS Activity = ~0) or induce loss of activity (Δ CBS Activity > 0), while 618,286 were predicted to strengthen a CBS (Δ CBS Activity < 0).

Positive Δ CBS Activity variants associated with elevated GERP (Spearman, $r=0.05$, $p < 1 \times 10^{-6}$) and MAPS scores (Spearman, $r=0.30$, $p < 1 \times 10^{-6}$). High-impact loss of activity variants (Δ CBS Activity > 0.75, $n=1,659$) are seen more rarely than missense variants (CBS MAPS=0.09, missense MAPS=0.03, MWU, $p < 1 \times 10^{-6}$). Negative Δ CBS Activity variants also associated with elevated GERP (Spearman, $r=-0.02$, $p < 2.3 \times 10^{-27}$) and MAPS (Spearman, $r=-0.09$, $p < 1 \times 10^{-6}$) scores, though the effect size was smaller than for variants with positive effects.

Conclusions: We integrated population genetic data with regulatory genome annotations to develop a novel metric of variant impact (Δ CBS Activity). We detect selective constraint on variation that is predicted to both elevate and reduce the strength of CBSs. This work represents a step toward characterizing the functional impact of an important class of noncoding variation and will enable improved understanding of how CBS variation contributes to diseases.

STRONG CONSERVATION OF PROMOTER-LIKE ELEMENTS, BUT NOT ACTIVE ENHancers, BETWEEN CIRCULATING PORCINE AND HUMAN IMMUNE CELLS

Ryan J Corbett¹, Juber H Uribe¹, Jinyan Teng², Kristen Byrne³, Haibo Liu¹, Houcheng Li⁴, Zhe Zhang², James E Koltes¹, Catherine W Ernst⁵, Crystal L Loving³, Lingzhao Fang⁴, Christopher K Tuggle¹

¹Iowa State University, Animal Science, Ames, IA, ²South China Agricultural University, Guangdong Laboratory of Lingnan Modern Agriculture, Guangzhou, China, ³USDA-ARS-NADC, Food Safety and Enteric Pathogens Research, Ames, IA, ⁴Aarhus University, Center for Quantitative Genetics and Genomics, Aarhus, Denmark, ⁵Michigan State University, Animal Science, East Lansing, MI

Quantitative genetics studies have revealed thousands of loci associated with complex immune system traits in pig, but limited characterization of cis gene regulation has hindered identification of mechanisms directly linking such genetic variants to altered phenotypes. We performed the first annotation of regulatory elements (RE) in nine porcine immune cell populations using mRNA-seq, ChIP-seq (H3K27ac, H3K27me3, H3K4me1, H3K4me3, CTCF) and ATAC-seq assays, and generated RE maps across cell types. We utilized this regulatory data to annotate immune tissue cis-eQTL, and to perform comparative analyses between porcine immune cells and matched human cells. We predicted 15 chromatin states—including active/poised transcription start sites (TSSs), enhancers, repressors, and insulators—and further leveraged CTCF data to predict 2,749 chromatin loops. We predicted 47,674 enhancer-gene interactions through correlative analyses involving 28,142 unique enhancers and 9,845 gene targets. 53,974 active enhancers showed significant cellular specificity, and these enhancers exhibited significant enrichment related to unique cellular functions among predicted gene targets, as well as motif enrichment for cell-specific transcription factors. Clusters of enhancers exhibiting high activation signal were classified as super-enhancers (N=140-843/cell type); these were associated with high target transcript abundance that suggest their putative role as master cis-regulators of cell-specific gene expression. Active regulatory elements including ATAC islands, TSSs, and enhancers showed the strongest enrichment for cis-eQTL in immune tissues, and we identified many such examples overlapping eQTL for immune cell marker genes. We observed strong functional conservation (>50%) of promoter-like elements and comparatively limited conservation (<10%) among active enhancers between porcine and human cells. This project has provided a rich resource of functional RE in porcine immune cells that may prove valuable in future quantitative genetics and translational research studies.

MODELING SINGLE-CELL ACTIVATION STATES ENHANCES POWER TO IDENTIFY EX VIVO STIMULATION RESPONSE eQTLs

Cristian Valencia^{1,2,3}, Aparna Nathan^{1,2,3}, Joyce Kang^{1,2,3}, Laurie Rumker^{1,2,3}, Soumya Raychaudhuri^{1,2,3}

¹Brigham & Women's Hospital, Genetics, Boston, MA, ²Harvard University, Medical school, Boston, MA, ³Broad Institute of MIT and Harvard, -, Cambridge, MA

Gene regulation of immune cells can be dynamic, such the activation of the interferon stimulated genes in response to environmental perturbation like pathogen infection. This dynamic process can be influenced by genetics, environment and their interactions. To understand this, previous studies used ex-vivo stimulation to identify “response” expression quantitative trait loci (eQTL), i.e. eQTLs whose magnitudes change in response to perturbation such as immune activation. Typically, such studies model activation as a discrete cell state. However, ex vivo stimulation conditions do not uniformly activate all perturbed cells. Thus, this model may be underpowered to find response QTLs. We demonstrated that effective modeling of per-cell activation state enhances power to detect response eQTLs. Overall, we used 210 published single-cell peripheral blood mononuclear cell (PBMC) data from Randolph et al. (Science, 2022) and Oelen et al (Nature Comm, 2022), to study the effect of ex-vivo stimulation with 4 pathogens on gene regulation. For each stimulus, we defined the continuous activation state of each cell using a penalized logistic regression. Then, to find response eQTLs, we used a Poisson mixed effects model of gene expression in single cells as a function of genotype and genotype interactions with discrete and continuous activation states, accounting for confounders and batch Nathan et al. (Nature, 2022). By modeling single-cell variation in activation in addition to discrete activation state we had better power to find response eQTL. For example, upon influenza stimulation 166/1854 of eGenes had response eQTL effects. In contrast, an alternative model with discrete activation cell state failed to detect 45% of these response eQTLs like SNCA-rs6828271 and THEMIS2-rs1467464. Furthermore, by comparing activation-dependent effects across cell-types, we detected cell-type-specific response eQTLs like OAS1- rs10774671 in lymphocytes. Our work provides a model for better detection of activation-dependent eQTLs and underscores the importance of modeling cell-level variation. Improving these models will ultimately advance our understanding of the biologic mechanisms that underlie disease heritability

SPATIAL TRANSCRIPTOMICS ANALYSIS REVEALS PATHOLOGY-SPECIFIC CELL AND MOLECULAR CHANGES IN PULMONARY FIBROSIS

Annika Vannan¹, Ruqian Liu², Arianna L Williams¹, Evan D Mee¹, Mei-i Chung¹, Saahithi Mallapragada¹, Jonathan A Kropski³, Davis J McCarthy², Nicholas E Banovich¹

¹Translational Genomics Research Institute, Integrated Cancer Genomics Division, Phoenix, AZ, ²St. Vincent's Institute of Medical Research, Bioinformatics and Cellular Genomics, Fitzroy, Australia, ³Vanderbilt University Medical Center, Department of Medicine, Nashville, TN

Pulmonary fibrosis (PF) is a chronic, progressive condition that characterizes the end stage of many interstitial lung diseases (ILDs). Studies using single-cell RNA sequencing (scRNA-seq) have allowed researchers to interrogate the complex cellular heterogeneity of the lung, lending insight into cell-specific molecular mechanisms of PF. However, scRNA-seq is limited by a lack of spatial context and fails to capture the regional heterogeneity that is characteristic of progressive PF. The advent and expansion of commercially available platforms to profile gene expression *in situ* (collectively termed spatial transcriptomics) are enabling researchers to probe the complex patterns of gene expression and cell-cell communication within local microenvironments and understand how spatial niches vary with disease progression. Using the Vizgen MERFISH platform, an imaging based spatial transcriptomics platform providing subcellular resolution, we characterized the expression of 140 genes across 21 samples from 13 donors, including 6 unaffected and 15 ILD samples. The ILD samples included donor-matched pairs of less and more fibrotic regions of tissue, allowing us to investigate the dynamics of disease progression. We identified differences in cell type composition and gene expression between unaffected and disease samples. Importantly, these data provide a uniquely accurate representation of cell composition in ILDs to date, with substantial recovery of cell types (e.g. capillary and other endothelial cells) that are typically depleted during scRNA-seq sample preparation. In addition, we found patterns of cellular and molecular dysregulation within clinician-annotated histopathological features such as interstitial fibrosis, granulomas, and honeycombing. In addition to this targeted strategy, we conducted a transcript-based, cell-agnostic analysis to identify distinct spatial niches across our sample set. We observed several niches that were specific to ILD, including some associated with greater disease-driven remodeling within and between samples. Together, these results advance our understanding of the molecular programs regulating PF *in situ*.

THE DISTINCT GENETIC DETERMINANTS OF REPRODUCTIVE HORMONES AND INFERTILITY

Samvida S Venkatesh^{*1}, Laura B L Wittemans^{*1}, Benjamin M Jacobs², Jessica F Campos de Jesus³, Minna Karjalainen⁴, Anu Pasanen⁴, Ahmed Elhakeem⁵, Deborah A Lawlor⁶, Nicholas J Timpson⁵, Triin Laisk³, Hannele Laivuori⁴, David van Heel², Cecilia M Lindgren¹

¹Big Data Institute, University of Oxford, Oxford, United Kingdom, ²Blizard Institute, Queen Mary Institute of London, London, United Kingdom, ³Estonian Genome Center, University of Tartu, Tartu, Estonia, ⁴Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland, ⁵Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom

*Equal contributions

Reproductive hormones regulate physiological processes such as sexual development & the reproductive cycle, and are implicated in a range of disorders, including infertility. However, the pathways driving hormonal imbalances & reproductive disorders remain poorly understood. Only a few genetics-based studies to understand the molecular aetiology of hormonal & non-hormonal contributions to infertility have been reported.

To elucidate the genetic factors driving reproductive hormone levels & infertility, we conducted large-scale genome-wide association study (GWAS) meta-analyses involving participants from 5 biobanks. We assessed > 9 million common genetic variants for sex-specific and sex-combined associations with testosterone (N=352,170), oestradiol (70,224), follicle stimulating hormone (FSH) (31,053), luteinizing hormone (LH) (27,263), & progesterone (8,468) levels; we also performed GWAS meta-analyses for female infertility of all causes (26,781 cases/487,561 controls), anovulatory infertility (5,764/491,845), and male infertility (3,151/358,780).

We identified genome-wide significant SNPs ($P<5E-8$) associated with testosterone (71 independent SNPs), FSH (6), LH (3), oestradiol (1), & progesterone (1) levels in women. These include SNPs in loci for FSHB (FSH subunit beta), SHBG (sex-hormone binding globulin), and CYP19A1 & CYP3A7 (involved in steroid hormone synthesis & metabolism). Between 30-90% of genome-wide significant SNPs identified in sex-specific analyses did not reach significance in the sex-combined strata ($P>5E-8$), and up to half of all SNPs show heterogeneous effects between female & male strata ($P<0.05$), indicating substantial sex-specific genetic architecture of hormone levels. We did not observe significant genetic correlations between any hormone and infertility traits (all $P>0.05$), and none of the SNPs associated with female infertility (4) or male infertility (2) overlap with reproductive hormone loci. In this large-scale GWAS meta-analysis, we highlight the distinct genetic drivers of reproductive hormones & infertility. To provide further insights into their biology, we plan to prioritise relevant genes & cell types using publicly available resources and a bespoke single-cell gene expression atlas of the ovary.

300 BILLION ASSOCIATIONS - GENETIC ARCHITECTURE OF 2,071 PHENOTYPES IN 658,582 INDIVIDUALS OF DIVERSE ANCESTRY IN THE VA MILLION VETERAN PROGRAM

Anurag Verma¹, Jennifer E Huffman², Alex Rodriguez³, Yuk-Lam Ho²,
Mitchell Conery⁶, Molei Liu⁴, Benjamin Voight¹, Tianxi Cai¹, Ravi K Madduri⁶,
Scott M Damrauer¹, Katherine P Liao²

¹CMC VA Medical Center, Department of Medicine, Philadelphia, PA, ²VA Boston Healthcare System, MAVERIC, Boston, MA, ³Argonne National Laboratory, Data Science and Learning, Lemont, IL, ⁴Columbia University's Mailman School of Public Health, Department of Biostatistics, New York, NY,
⁵Harvard Medical School, Department of Biomedical Informatics, Boston, TN,
⁶University of Pennsylvania, Department of Genetics, Philadelphia, PA

Genome-wide association studies (GWAS) have accelerated our understanding of the phenotypic and biological consequences of genetic variation across the genome. However, existing GWASs have small numbers of individuals of non-European ancestry, and a significant knowledge gap exists in genetic diversity. The Department of Veteran Affairs (VA) Million Veteran Program (MVP) is a cohort of over 900,000 Veterans of diverse ancestry with genetic information linked to electronic health records (EHR) and questionnaire data. As part of a larger effort to detect genetic associations across 2,071 phenotypes (1,629 diagnoses, 213 laboratory measurements, 211 survey questions, 18 anthropometric traits) in a subset of 658,582 MVP participants of African (AFR, n=123,328), Admixed American (AMR, n=61,111), East Asian (EAS, n=7,076), and European (EUR, n=457,720) ancestry. This large-scale analysis was enabled by the US Department of Energy's (DOE) Summit supercomputer, where analyses were optimized and deployed. In total, we conducted 5,801 GWASs and cross-ancestry meta-analysis. We report 10,319 independent loci (p -value $< 4.6 \times 10^{-11}$) identified in meta-analysis across the 1,758 phenotypes. Then, we performed fine-mapping on 1,207 phenotypes where at least one significant locus was detected in a trans-ancestral meta-analysis using in-sample linkage disequilibrium (LD) reference panels and the Sum of Single Effects (SuSiE) framework. By pooling the results across ancestries in a trans-ancestral approach, we were able to narrow down the field of putative causal variants and identify signals that were truly ancestry-specific. Overall, we found 55,820 trait pair associations, with the majority of signals being in the European ancestry. However, we also report that 10,314 (~18%) of trait pair associations were only identified in non-European ancestry. We also identified evidence of heterogeneity where 1,207 traits had significantly heterogeneous associations with at least one SNP across ancestries. Hyperlipidemia had the most significant differences in AFR vs. EUR and were also present in AMR, EAS vs. EUR. Our findings represent a significant advance in the genetic architecture of complex human traits in non-European populations. This study significantly expands the corpus of fine-mapped genotype-phenotype associations in diverse individuals available to the global research community.

GENETIC VARIATION SHAPING THE TRANSCRIPTOMIC IMMUNE RESPONSE TO *YERSINIA PESTIS*

Tauras P Vilgalys, Mari Shiratori, Anne Dumaine, Luis B Barreiro

University of Chicago, Genetic Medicine, Chicago, IL

Yersinia pestis is the causative agent of plague and responsible for the greatest single mortality event in history: the first outbreak of the Second Plague Pandemic, now commonly referred to as the Black Death, which killed 30-50% of the population in Afro-Eurasia.

To identify variants which affect the response to *Y. pestis*, we used single-cell RNA sequencing of peripheral blood mononuclear cells (PBMCs) for 90 individuals of African- and European-American ancestry infected with live, fully virulent *Y. pestis*. We assign cells into major PBMC cell types (monocytes, NK cells, B cells, CD4+ and CD8+ T cells), and find each cell type mounts a robust response to *Y. pestis* infection (1,881-8,672 genes, 5% FDR). Combined with genotypes for these individuals, we mapped expression quantitative trait loci (eQTL) and detect 2,105 genes with at least one eQTL in one or more cell types, including 402 variants that only affect gene expression levels in *Y. pestis* infected cells (i.e., immune response eQTL).

Immune response eQTL were up to 4x as likely to also affect the expression response to *Listeria monocytogenes* ($\log_2(\text{OR}) = 1.6$; $p = 1.6 \times 10^{-3}$), *Salmonella typhimurium* ($\log_2(\text{OR}) = 0.89$; $p = 0.012$), and influenza A virus ($\log_2(\text{OR}) = 2$; $p = 1.1 \times 10^{-8}$). Furthermore, immune response eQTL also overlap signatures of recent positive selection and genetic variants associated with the risk of autoimmune and inflammatory diseases.

Together, our results suggest that genetic variation that affects the immune response to *Y. pestis* also impacts the response to other infectious agents and the risk of developing inflammatory and autoimmune diseases. This is consistent with our recent results suggesting a trade-off between variants that were protective during the Black Death and variants that are today associated with increased risk for immune-related disorders.

GENETIC IMPACTS ON DNA METHYLATION HELP DISSECT THE INTERPLAY BETWEEN GENETICS, EPIGENETICS AND DISEASE.

Sergio Villicaña¹, Juan Castillo-Fernandez¹, Eilis Hannon², Colette Christiansen¹, Pei-Chien Tsai¹, Jane Maddock³, Diana Kuh³, Matthew Suderman⁴, Christine Power⁵, Caroline Relton^{4,5}, George Ploubidis⁶, Andrew Wong³, Rebecca Hardy⁷, Alissa Goodman⁶, Ken K Ong⁸, Jordana T Bell¹

¹KCL, Department of Twin Research, London, United Kingdom, ²UOE, Medical School, Exeter, United Kingdom, ³UCL, MRC Unit for Lifelong Health and Ageing, London, United Kingdom, ⁴UOB, MRC Integrative Epidemiology Unit, UOB, United Kingdom, ⁵UCL, Population, Policy and Practice, London, United Kingdom, ⁶UCL, Centre for Longitudinal Studies, London, United Kingdom, ⁷UCL, Social Research Institute, London, United Kingdom, ⁸UC, MRC Epidemiology Unit and Department of Paediatrics, Cambridge, United Kingdom

Background: DNA methylation (DNAm) levels at a proportion of sites across the genome are influenced by genetic variants, or methylation quantitative trait loci (meQTLs). The purpose of this work was to identify meQTLs through genome-wide association analyses and to explore the relationship between DNA methylation, gene expression and human phenotypes.

Methods: Blood DNA samples were obtained from 2,358 participants from three UK-based population cohorts including TwinsUK, the MRC National Survey for Health and Development, and the National Child Development Study. Genome-wide association analyses compared genotypes and DNAm levels at 724,499 sites. Local (DNAm sites and SNPs within 1 Mb) and long-range (all other) associations were considered. Summary-based Mendelian Randomization (SMR) and heterogeneity independent instruments (HEIDI) tests were used to find associations between DNAm and 56 phenotypes. Further SMR and HEIDI analyses tested associations between DNAm and blood gene expression levels for 19,250 genes using published expression QTLs. Cell type-specific meQTL effects were carried out for CD4⁺ T cells and monocytes. Tissue specificity of blood meQTLs was explored in adipose and skin tissue datasets in subsets of 390–542 TwinsUK samples.

Results: Whole blood DNA methylation levels at 34.1% of the DNAm sites genome-wide were affected by meQTLs, and 98% of meQTLs had local effects. Cell type-specific meQTLs were found for a minority of DNAm sites with meQTLs in whole blood. Transcription factor binding sites, enhancers, and other regions showed enrichment for meQTL SNPs. SMR and HEIDI analyses identified 1,521 co-localisation events between 1,326 CpGs and 35 phenotypes, suggesting potential pleiotropic effects. We found that meQTL SNPs also colocalize with putative causal variants for gene expression of 44.2% of tested genes. The results provide evidence in support of specific examples of hypothesized underlying mechanisms of genetic regulation on DNAm.

Conclusion: We observed genetic influences on blood DNA methylation levels at a substantial proportion of DNAm sites profiled by the EPIC array. The results contribute to a better understanding of the mechanisms underlying DNA methylation variability.

DISSECTING THE INTERPLAY BETWEEN AGEING, SEX AND BODY MASS INDEX ON A MOLECULAR LEVEL.

T D Michaletou¹, MG Hong², J Fernandez-Tajes³, S Sharma⁴, C A Brorsson⁵, R W Koivula⁶, J Adamski⁴, S Brunak⁷, P W Franks⁶, E T Dermitzakis⁸, E R Person⁹, J M Schwenk¹⁰, M Walker¹, A A Brown⁹, A Viñuela¹

¹Newcastle University, Newcastle, United Kingdom, ²National Bioinformatics Infrastructure, Stockholm, Sweden, ³University of Oxford, Oxford, United Kingdom, ⁴Helmholtz Zentrum, München, Germany, ⁵Technical University, Copenhagen, Denmark, ⁶Lund University, Lund, Sweden, ⁷Copenhagen University, Copenhagen, Denmark, ⁸University of Geneva, Geneva, Switzerland, ⁹University of Dundee, Dundee, United Kingdom, ¹⁰KTH, Stockholm, Sweden

Ageing is a complex process, entangled with a variety of age-related traits. Here, we aim to separate age-related molecular changes from the influence of sex and BMI on gene expression, proteins and metabolites abundances. Utilizing data from the DIRECT consortium from 3,027 participants, we identified 10,643 (66%), 10,163 (63%) and 8,256 (51%) genes differentially expressed with age, sex and BMI respectively (FDR<0.05). Protein associations were 316 (85%), 260 (70%) and 271 (73%). More than 20% of genes and 45% of measured proteins were independently associated with all three factors such as ITGB1, and EGFR. Of the 286 proteins for which gene expression was measured, 164 were significantly associated to age with 81 (49.4%) showing opposite direction of effects. For example, *CD85* expression increased with age while the protein abundance decreased with age. When studying the overlap of associations between biological factors, for instance, sex-related proteins were mostly involved in T-cell activation responses. Moreover, BMI-related proteins included *CDH1*, where mutations in the corresponding gene have been correlated with multiple types of cancer. Using interaction models, we identified 850 genes, 103 proteins and 48 metabolites with significant sex-dependent changes with age. For example, abundance of Sclerostin increased with age in men, but decreased in women, while levels of L-carnitine showed the inverse effect. BMI-dependent changes with age were detected for only for *GAS6* levels that decreased with age in individuals with high BMI, while 87 proteins changed their abundances with BMI in a sex-dependent manner. All in all, we observed a majority of tested molecular phenotypes to be independently associated with age, sex and BMI. However, we also identified phenotype-dependent changes in molecular abundances.

SCMETABRAIN: FEDERATED SINGLE-CELL CONSORTIUM FOR CELL-TYPE SPECIFIC eQTL ANALYSIS OF NEUROLOGICAL DISEASE VARIANTS

Martijn Voscheloo^{1,2}, Roy Oelen^{1,2}, Drew R Neavin³, Robert Warmerdam^{1,2}, Urmo Võsa^{1,4}, Maryna Korshevniuk^{1,2}, Dan Kaptijn^{1,2}, Monique van der Wijst^{1,2}, Marc Jan Bonder^{1,5}, scEQTLGen Consortium¹, Tõnu Esko⁴, Julien Bryois⁶, Ellen A Tsai⁷, Heiko Runz⁷, Lude Franke^{1,2}, Harm-Jan Westra^{1,2}

¹University Medical Center Groningen, University of Groningen, Department of Genetics, Groningen, Netherlands, ²Oncode Investigator, Utrecht, Netherlands, ³Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute for Medical Research, Darlinghurst, NSW, Australia, ⁴Estonian Genome Centre, University of Tartu, Institute of Genomics, Tartu, Estonia, ⁵Deutsches Krebsforschungszentrum (DKFZ), Division of Computational Genomics and Systems Genetics, Heidelberg, Germany, ⁶F. Hoffmann-La Roche Ltd., Neuroscience and Rare Diseases, Basel, Switzerland, ⁷Biogen Inc., Translational Biology, Research & Development, Cambridge, MA

Background/Objectives: Recently we have completed large scale eQTL meta-analyses in blood (eQTLGen; Võsa *et al.* 2021) and brain (MetaBrain; de Klein *et al.* 2021), providing insight into the downstream effects of disease-associated genetic risk factors. Although having substantial sample sizes, these studies lack the cell type context that single-cell RNA sequencing (scRNA-seq) can provide. To pinpoint these cellular contexts and better interpret neurological diseases, we have initiated the scMetaBrain study, which aims to enable a federated scRNA-seq eQTL analysis in human brain.

Methods: To harmonize and compare with our single cell eQTL efforts in blood, we have adapted our recently developed pipeline (scEQTLGen, van der Wijst *et al.* 2020) to brain. This includes robust containerized pipelines that perform quality control, demultiplexing, cell type classification, and eQTL analysis per cell-type and dataset. By applying a meta-analysis of the summary statistics, we enable easy inclusion of new datasets without the need of sharing person-identifiable data.

Results: As proof of concept, we have applied our pipeline to automatically analyze 35 samples from Mathys *et al.* 2019. We report good replication statistics compared to the largest brain single-cell eQTL study (Bryois *et al.* 2022).

Conclusion: Here we introduce scMetaBrain, in which we aim to setup a single-cell brain consortium for the identification of downstream consequences of trait-related genetic variants in specific brain cell types. We envision that this consortium will enable a unique opportunity to disentangle tissue and cell type specific regulatory effects in the brain.

Grant References: NWO VICI 09150182010019, Oncode Senior Investigator

THE POOR PORTABILITY OF POLYGENIC SCORES IS ONLY PARTIALLY ATTRIBUTABLE TO GENETIC ANCESTRY

Joyce Y Wang¹, Michael Zietz², Jason Mares³, Paul J Rathouz^{4,5}, Vagheesh M Narasimhan^{1,5}, Molly F Przeworski^{6,7}, Arbel Harpak^{1,4}

¹The University of Texas at Austin, Department of Integrative Biology, Austin, TX, ²Columbia University Irving Medical Center, Department of Biomedical Informatics, New York City, NY, ³Columbia University, Institute of Social and Economic Policy, New York City, NY, ⁴Dell Medical School, Department of Population Health, Austin, TX, ⁵The University of Texas at Austin, Department of Statistics and Data Sciences, Austin, TX, ⁶Columbia University, Department of Biological Sciences, New York City, NY, ⁷Columbia University, Department of Systems Biology, New York City, NY

A major obstacle hindering the broad adoption of polygenic scores (PGS) is their lack of “portability” to people that differ—in genetic ancestry, environmental exposures or other characteristics—from the GWAS samples in which genetic effects were estimated. First principle considerations suggest some of the portability problem arises from differences in genetic ancestry; how much, however, remains unknown. In this work, we tackle two related questions.

First, we derive an estimator of individual PGS prediction error that is based on local genetic similarity to the GWAS sample, i.e., similarity in the genomic regions used in the construction of the PGS. Using the UK Biobank data, we show that our estimates better predict PGS prediction error than those based on global ancestry, and far better than those based on self-identified race. One implication is that measures of local genetic similarity should better predict the clinical applicability of a given PGS to a target individual than currently used designators.

Second, while previous studies have shown decreased prediction accuracy in genetic ancestry groups other than that of the GWAS sample, such groupings confound differences in allele frequencies and linkage disequilibrium patterns with many other factors, including distinct environmental exposures. To gain more resolution, we measure PGS prediction accuracy as a function of continuous genetic distance from the GWAS sample to a target genome. We estimate that genetic dissimilarity accounts for 3-60% of the variance in prediction accuracy (depending on the trait, genetic distance metric, and measure of prediction accuracy). In some cases, crude measures of socio-economic status account for comparable variance in prediction accuracy. These findings imply, in contrast to previous reports, that factors other than genetic ancestry play an important role in the portability problem.

Together, our analyses help quantify the factors that impede the portability of PGS to individuals underrepresented in a GWAS sample, and inform the applicability of PGS in the clinic and beyond.

THE ROLE OF RACE AND GENETIC ANCESTRY IN KRAS MUTATION AND SUBTYPES IN NON-SMALL CELL LUNG CANCER

Xinan Wang¹, Kangcheng Hou², Rounak Dey³, Biagio Ricciuti⁴, Xihong Lin³, Bruce E Johnson⁴, David C Christiani¹

¹Harvard University, Environmental Health, Boston, MA, ²University of California, Bioinformatics, Los Angeles, CA, ³Harvard University, Biostatistics, Boston, MA, ⁴Dana-Farber Cancer Institute, Thoracic Oncology, Boston, MA

Background: KRAS mutation is one of the most common oncogene aberrations in patients with non-small cell lung cancer (NSCLC). However, the racial and genetic ancestral effects and their interplay on complex KRAS mutation and subtypes remain largely unknown.

Methods: A total of 3918 NSCLC patients from the Chinese OrigMed (OM) cohort (n=2039) and Boston Lung Cancer Survival (BLCS) cohort (n=1879) were included. Baseline covariates including age at diagnosis, sex, clinical stage, self-reported race/ethnicity, histological subtypes and smoking status were well-annotated. KRAS mutation and subtypes were determined by targeted NGS panels in each cohort. KRAS mutation outcomes including binary KRAS mutant vs. wild type, multinomial KRAS subtypes, and KRAS transversion and transition events. Global and local genetic ancestry component were estimated from paired germline data for patients in BLCS. Multivariable logistic and multinomial logistic regressions were utilized to assess the association between self-identified race, genetic ancestry and KRAS outcomes.

Result: In the entire cohort, 804 (20.5%) patients harbored KRAS mutation. 92.4% (1736/1879) of the BLCS patients were genetically admixed with more than one genetic ancestry component. Self-identified Asian is significantly associated with lower risk of developing KRAS mutation (OR=0.44, 95%CI:0.22-0.81, P=0.01), specifically fewer transversion events (OR=0.24, 95% CI:0.10-0.62, P=0.003) and KRAS^{G12C} (OR=0.17, 95% CI:0.08-4.78, P=0.02) compared to White, and this effect was further modified by sex. A 10% increase in Asian genetic ancestry is associated with a lower risk of developing KRAS mutation (OR=0.94, 95% CI: 0.89-0.98, P=0.05) while a 10% increase in European ancestry is associated with a higher risk of KRAS mutation (OR=1.05, 95% CI:1.00-1.09, P=0.03), specifically more transition events (10% increase OR =1.10, 95% CI:1.01-1.21, P=0.03) and KRAS^{G12D} (10% increase OR=1.18, 95% CI:1.04-1.34, P=0.01) in the adjusted analyses. Local genetic ancestry of genes enriched in receptor tyrosine kinase and growth factor signaling pathways, including KDR ($P=6*10^{-5}$) and KIT ($P=1*10^{-4}$), may be associated with KRAS mutation.

Conclusion: Race and genetic ancestry are associated with KRAS mutation and subtypes in NSCLC and may complement each other in our understanding of lung cancer molecular disparities, leading to more accurate intervention and clinical decisions.

COSPAR IDENTIFIES EARLY CELL FATE BIASES FROM SINGLE CELL TRANSCRIPTOMIC AND LINEAGE INFORMATION

Shou-Wen Wang^{1,4}, Michael J Herriges², Kilian Hurley³, Darrell N Kotton², Allon M Klein¹

¹Department of Systems Biology, Blavatnik Institute, Harvard Medical School, Boston, MA, ²Center for Regenerative Medicine, Boston University, Boston, MA, ³Department of Medicine, Royal College of Surgeons in Ireland, Education and Research Centre, Beaumont Hospital, Dublin, Ireland, ⁴School of Life Sciences, Westlake University, HangZhou, China

A goal of single cell genome-wide profiling is to reconstruct dynamic transitions during cell differentiation, disease onset, and drug response. Single cell assays have recently been integrated with lineage tracing, a set of methods that identify cells of common ancestry to establish bona fide dynamic relationships between cell states. These integrated methods have revealed unappreciated cell dynamics, but their analysis faces recurrent challenges arising from noisy, dispersed lineage data. Here, we develop coherent, sparse optimization (CoSpar) as a robust computational approach to infer cell dynamics from single-cell transcriptomics integrated with lineage tracing. Built on assumptions of coherence and sparsity of transition maps, CoSpar is robust to severe down-sampling and dispersion of lineage data, which enables simpler experimental designs and requires less calibration. In datasets representing hematopoiesis, reprogramming, and directed differentiation, CoSpar identifies early fate biases not previously detected, predicting transcription factors and receptors implicated in fate choice. Documentation and detailed examples for common experimental designs are available at <https://cospar.readthedocs.io/>.

IMPACT OF INDIVIDUAL LEVEL UNCERTAINTY OF LUNG CANCER POLYGENIC RISK SCORE ON RISK STRATIFICATION AND PREDICTION

Xinan Wang^{*1}, Ziwei Zhang^{*2}, Tony Chen³, Yi Ding⁴, Xihong Lin³, David C Christiani¹

¹Harvard University, Environmental Health, Boston, MA, ²Dana-Farber Cancer Institute, Medical Oncology, Boston, MA, ³Harvard University, Biostatistics, Boston, CA, ⁴University of California, Bioinformatics, Los Angeles, MA

Background: Although polygenic risk score has emerged as a promising tool for predicting cancer risk, the accuracy of lung cancer PRS at the individual level and its impact on subsequent clinical applications remains largely unexplored.

Method: Lung cancer PRSs and confidence/credible interval (CI) were constructed for each individual using - 1) the weighed sum of 16 GWAS-derived significant SNP loci and CI through the bootstrapping method (PRS-16-CV), and 2) LDpred2 and CI through posteriors sampling (PRS-Bayes), among 17166 cases and 12894 controls with European ancestry from the International Lung Cancer Consortium. Individuals were classified into different genetic risk subgroups based on the relationship between their own PRS mean/CI and the population level threshold. Multivariable logistic analyses were conducted to estimate the relative risk of lung cancer for different PRS risk subgroups. Risk prediction models were constructed using both PRS risk subgroups and non-genetic risk factors.

Result: Considerable variances in PRS point estimates at the individual level were observed for both methods, with an average standard deviation (s.d.) of 0.12 (95%CI:0.09-0.15) for PRS-16-CV and a much larger s.d. of 0.88 (95%CI:0.68-1.11) for PRS-Bayes. With a confidence/credible level $\rho=95\%$, using PRS-16-CV, only 25% of individuals with PRS point estimates in the lowest decile of PRS and 16.8% in the highest decile have their entire 95% CI fully contained in the lowest and highest decile, respectively; while PRS-Bayes was unable to find any eligible individuals. Only 19% of the individuals were concordantly identified as high genetic risk ($>90^{\text{th}}$ percentile) using the two PRS estimators. An increased relative risk of lung cancer comparing high genetic risk to low genetic risk ($0^{\text{th}}-10^{\text{th}}$ percentile) was observed when taking the CI into account ($OR=2.73$, 95%CI:2.12-3.50, $P=4.13 \times 10^{-15}$) compared to using PRS-16-CV mean ($OR=2.23$, 95%CI:1.99-2.49, $P=5.70 \times 10^{-46}$). Improved risk prediction performance was consistently observed in individuals identified by PRS-16-CV CI and the best performance was achieved by including age, sex and smoking pack-years (AUC:0.73, 95%CI:0.72-0.74).

Conclusion: Individual level uncertainty of lung cancer PRS greatly impacted PRS-based ranking, risk stratification and prediction in populations with European ancestry, highlighting the importance of taking it into considerations when evaluating the practical utility of PRS.

CALYPSO: LONGITUDINAL GENOMIC DIAGNOSTIC CARE WITH INNOVATIVE WEB TOOLS

Alistair Ward^{1,2}, Isabelle Cooperstein¹, Tony Di Sera¹, Stephanie Georges¹, Anders Pitman¹, Marti Tristani-Firouzi¹, Gabor Marth^{1,2}

¹University of Utah, Human Genetics, Salt Lake City, UT, ²Frameshift Labs, Cambridge, MA

Patients presenting with complex phenotypes represent some of the most challenging diagnostic cases. To compound the complexity, patient phenotypes can evolve over time. To achieve a genetic diagnosis, the combined effort of large diagnostic teams from multiple medical disciplines are often required. The treating physician provides a detailed understanding of the patient's clinical presentation and family history; the medical geneticist brings a wide and deep knowledge of genetic diseases bioinformaticians provide computational analysis skills and an ability to adjudicate the role of individual variants; and diagnostic pathologists synthesize all information to draw conclusions on a variant's clinical significance. Currently available software tools cater primarily to those with computational expertise rather than clinical expertise and focus on diagnostic analysis at a single point in time. This results in A) treating physicians, or genetic counselors being left out of the diagnostic process, and; B) a failure to follow a patient over a potentially lengthy diagnostic process, where patient phenotypes and/or variant interpretations can change, potentially impacting the diagnosis.

Calypso is a software system being developed to address these concerns and comprises:

1. A computational backend to perform comprehensive annotation of incoming genomes and construct a large filtered set of variants for querying. The backend pipeline is being expanded to enable re-annotation and re-analysis of all variants on a periodic or user triggered basis,
2. Web tools including the iobio suite to ensure each case is summarized with quality statistics, clinical information and other analytics. The user-friendly suite of web tools ensures that all team members are able to interact with, and perform diagnostic tasks commensurate with their individual expertise,
3. Dashboards summarize individual cases with visual elements including a watchlist of candidate diagnostic variants, a visual timeline of the patient case, including when re-annotation / re-analysis tasks, or changes to phenotypes occurred along with their effects. Cohort level dashboards allow examination of trends across large numbers of cases and samples,
4. A communication infrastructure allows close collaboration between all diagnostic team members. Steering conversations away from email and attaching them to their object of discussion (e.g. a variant, a chart describing a failing quality metric etc.) makes the diagnostic process more efficient, searchable, and user-friendly.

Initial versions of **Calypso** have been deployed at the University of Utah, where it is used to aid team-based diagnostic analysis as part of the rapid NICU sequencing program and at the UDN network level where it is accessible to all UDN members.

TOP2B-INHIBITING BREAST CANCER DRUGS INCLUDING ANTHRACYCLINES AFFECT CARDIOMYOCYTE HEALTH THROUGH A SHARED GENE EXPRESSION RESPONSE SIGNATURE

Elizabeth R Matthews¹, Omar D Johnson², Kandace J Horn³, Jose A Gutierrez¹, Simon Powell⁴, Michelle C Ward¹

¹University of Texas Medical Branch, Department of Biochemistry and Molecular Biology, Galveston, TX, ²University of Texas Medical Branch, Biochemistry, Cellular and Molecular Biology Graduate Program, Galveston, TX, ³University of Texas Medical Branch, John Sealy School of Medicine, Galveston, TX, ⁴University of Texas Medical Branch, Neuroscience Graduate Program, Galveston, TX

Breast cancer drugs, such as the anthracycline (AC) Doxorubicin (DOX), cause off-target effects on the heart in some, but not all women. There are tens of AC and non-AC drugs approved for breast cancer treatment therefore predicting adverse drug reactions in each individual could inform on optimal patient treatment. The DNA topology regulator, Top2B, is a target of DOX and mediates DOX-induced cardiotoxicity. To gain insight into breast cancer drug-specific effects on the heart across individuals, we developed an iPSC-derived cardiomyocyte (iPSC-CM) model from six genotyped females. We exposed iPSC-CMs to three AC-based Top2B inhibitors (Top2Bi): DOX, Daunarubicin and Epirubicin; a non-AC-based Top2Bi: Mitoxantrone; an unrelated monoclonal antibody against HER2: Trastuzumab (TRA), and a vehicle control. Experimentally-derived 50% lethal doses of the Top2Bi drugs are 1-2.5 uM, which is within the range observed in cancer patients. TRA does not affect viability. We chose a sub-lethal dose (0.5 uM) to measure primary responses to drug treatment at 24 hours. We observe a significant increase in cell stress marker release across drugs, and effects on cardiomyocyte contractility in Top2Bi-treated cells. Global gene expression data shows that drug treatment accounts for most variation between samples followed by the individual from which samples were derived. Joint analysis across all five drugs reveals two predominant gene expression signatures: Top2Bi response genes (5,611), and non-response genes (9,211). Top2Bi response genes are enriched in cell cycle, DNA replication, and DNA damage pathways. To dissect drug-specific effects, we performed pairwise comparisons between each drug and vehicle treatment. Of the 7,192 genes that respond to DOX, 4.5% are specific to DOX treatment, while 50% are shared across AC drugs. 377 published DOX response eQTLs are equally distributed amongst DOX-specific and AC-shared response genes suggesting some genetic effects on gene expression may be drug specific. Our data thus demonstrate that Top2Bi affect cardiomyocyte health and induce a shared gene expression response signature. However, hundreds of genes respond in a drug-specific manner and warrant further investigation for genetic effects across individuals.

TRANSFORMING GENOMICS RESEARCH THROUGH COMMUNITY ENGAGEMENT AND RETURN OF RESULTS: A CASE STUDY FROM FRENCH POLYNESIA

Kaja A Wasik¹, Sarah LeBaron von Baeyer¹, Keolu Fox ², Tristan Pascart ³, Tehani Mairai¹, Vehia Wheeler⁴

¹Variant Bio, Genomic Discovery, Seattle, WA, ²University of California San Diego, Indigenous Futures Lab, La Jolla, CA, ³Lille Catholic University, Rheumatology, Lille, France, ⁴Australia National University, Land and Ocean Management, Canberra, Australia

Many communities around the world are wary of genomics research, and for good reasons. All too often, researchers have failed to address local priorities and power inequities when designing and executing these studies. What is more, findings are rarely communicated to participating communities. Here, we highlight an approach that seeks to address colonial and ongoing legacies of extractive research by performing in-depth community engagement, respecting local cultural protocols, and sharing meaningful benefits as well as population-level genomic and health results first and foremost with participating communities. The study was conducted in French Polynesia where ~1100 individuals participated in a sample and data collection campaign spanning genomics, transcriptomics, metabolomics, and extensive biomedical health evaluation. In the course of the study we made multiple findings with significant implications for public health in French Polynesia. For instance we discovered that more than a quarter of the adult population of French Polynesia is estimated to have gout (compared to 0.9% in metropolitan France and 3.9% in the US). We will discuss how appropriate community engagement enabled us to conduct the study and gather data, and how these results are currently shaping the discourse around treatment of gout and public health across the Polynesian archipelago.

RURAL AND URBAN LIFESTYLES ARE ASSOCIATED WITH DIFFERENTIAL IMMUNE GENE REGULATION IN TURKANA

Marina M Watowich¹, Kristina M Garske^{2,3,4}, Varada Abhyankar⁴, Echwa John², Michael Gurven⁵, John Kahumbu², Joseph Kamau^{6,7}, Dino J Martins^{2,3,8}, Charles Miano², Benjamin Muhoya^{2,3}, Julie Peng^{3,4}, Jenny Tung^{9,10,11}, Julien F Ayroles^{3,4}, Amanda J Lea^{1,10}

¹Vanderbilt University, Department of Biological Sciences, Nashville, TN,

²Turkana Health and Genomics Project, Mpala Research Centre, Nanyuki, Kenya,

³Princeton University, Department of Ecology and Evolutionary Biology, Princeton, NJ, ⁴Princeton University, Lewis Sigler Institute for Integrative Genomics, Princeton, NJ, ⁵University of California Santa Barbara, Department of Anthropology, Santa Barbara, CA, ⁶National Museums of Kenya, Institute of Primate Research, Nairobi, Kenya, ⁷University of Nairobi, Department of Biochemistry, Nairobi, Kenya, ⁸Stony Brook University, Turkana Basin Institute, Stony Brook, NY, ⁹Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, ¹⁰Canadian Institute for Advanced Research, Child and Brain Development, Toronto, Canada, ¹¹Duke University, Department of Evolutionary Anthropology, Durham, NC

The recent spread of industrialization and urbanization has produced global changes in human diet, activity levels, and environmental exposures, and is associated with substantial increases in rates of non-communicable diseases (NCDs; e.g., cardiometabolic diseases). Increased NCD risk among urban populations is thought to arise from a mismatch between current and ancestral environments. If so, phenotypes that are beneficial in one environment may be disadvantageous in another, yet this hypothesis has rarely been tested at the molecular level. We investigated environmental mismatch at multiple genomic levels among the Turkana of northwest Kenya, a group undergoing a rapid transition from traditional pastoralist practices to living in urban, market-integrated areas. Specifically, we investigated the effect of lifestyle (pastoralist versus urban) on genome-wide gene expression (n=452 individuals, 8438 genes) and DNA methylation (n=320 individuals and 598712 CpG sites). We found 1155 genes and 168 CpG sites significantly associated with lifestyle (FDR<10%). Genes upregulated in urban individuals were primarily enriched for immune pathways and also overlapped with known NCD genes. Further, lifestyle-associated CpG sites were more likely to fall near genes differentially expressed with lifestyle than expected by chance, suggesting that environmentally-induced epigenetic changes may regulate gene transcription. Additionally, we have previously found that Turkana living in urban environments have biomarker profiles indicative of worse cardiometabolic health, and here we find that a small subset of differentially expressed genes significantly mediate these lifestyle-biomarker links. Together, our results provide new insight into environmentally-dependent gene regulation in a unique population, with implications for understanding the molecular underpinnings of cardiometabolic disease.

A COMPUTATIONAL TOOLKIT TO INTEGRATE MULTI-OMICS TIME-SERIES DATA ACROSS SPECIES IN BRAIN DEVELOPMENT

Beatrice Borsari^{1,2}, Eve S Wattenberg*^{1,2}, Ke Xu^{*1,2,3}, Xuezhu Yu^{*1,2}, Mor Frank^{1,2}, Susanna Liu^{1,2}, Mark Gerstein^{1,2}

¹Yale University, Program in Computational Biology and Bioinformatics, New Haven, CT, ²Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT, ³Yale University, Department of Biostatistics, New Haven, CT

Steady-state molecular studies provide a static description of molecular processes in cell differentiation or organismal development. Longitudinal studies can instead offer a deeper understanding of epigenetic and transcriptional events and how these processes evolve over time. Here, we integrate multi-omics data (ATAC-seq, DNase-seq, ChIP-seq, RNA-seq) from different consortia (Roadmap, ENCODE, PsychENCODE) to study brain development at equivalent post-conception (p.c.) dates in human and mouse (7-20 p.c. weeks, and 10.5-16.5 p.c. days, respectively). Genes with key roles in human brain development are associated with enhancers showing many different dynamic patterns, while genes performing non-brain functions are more often associated with just one type of enhancers (either up- or down-regulated). This suggests that proper expression of key human brain genes may require tight, multi-gate epigenetic control which is not seen in mouse. To further understand differences in brain development between human and mouse, we trained a random-forest classifier that predicts the species-specificity of enhancer-based time-series data. Overall, we found both activation and repression of enhancers happen earlier in mice than humans. This is in line with previous studies comparing molecular processes in these species. The 10-12 p.c. weeks window is key to distinguishing enhancer activation patterns between the two species. At this time, 26% of mouse enhancers have already undergone major activation and 45% have undergone major repression, while human enhancers are mostly activated or repressed between 12-13 p.c. weeks. While most of our analyses use bulk sequencing data, we find that roughly 30% of our dynamic brain enhancers are not detected by single cell ATAC-seq data from fetal brains. This suggests bulk assays detect a larger amount of epigenetic changes than single-cell data. Therefore, we are developing a multivariate Hidden Markov Model to estimate brain cell composition at each time-point by integrating bulk and single-cell epigenomics data. Altogether, this work aims at developing a computational toolkit to best harmonize and analyze time-series data generated across species, consortia and technologies. These analyses will contribute to new predictive modeling approaches highlighting changes in epigenetic patterns that have evolved between species.

*Authors contributed equally

METHYLOME, TRANSCRIPTOME AND ALTERNATIVE SPLICING PROFILING OF NEURONS, ASTROCYTES, AND MICROGLIA

Xiaoran Wei¹, Michelle L Olsen²

¹Virginia Tech, Virginia-Maryland College of Veterinary Medicine, Biomedical and Veterinary Sciences Program, Blacksburg, VA, ²Virginia Tech, School of Neuroscience, Blacksburg, VA

Methylation, both 5-methylcytosine (m5c) and 5-hydroxymethylcytosine (hm5c), represent two abundant and important epigenetic regulators in brain development, learning and memory and brain disease. Yet, how these epigenetic marks vary across different CNS cell types has not been evaluated. Using a magnetic cell separation technique, we isolated neurons, astrocytes, and microglia from postnatal day 28 WT male mice cortex. RNA and high molecular weight DNA were extracted from each cell population. Nanopore sequencing, which enables long read sequencing without the need for PCR amplification or bisulfite sequencing, was used to quantitatively evaluate m5c and hm5c across the entire genome in neurons, astrocytes and microglia. RNA sequencing was used to evaluate the expression level of all coding genes. Our analysis revealed marked differences in the neuron, astrocyte and microglia methylome. Notably, we identified microglia demonstrated higher levels of m5c relative to both astrocytes and neurons while astrocytes demonstrated higher levels of hm5c. Further, we found that m5c levels in the promoter region negatively correlated with gene expression levels in previously identified and newly described cell type-specific markers for all three cell populations while hm5c levels in the promoter region positively correlated with gene expression levels for cell markers in both astrocyte and neurons. Ongoing work is aimed at exploring the correlation between cell type specific transcription factor expression, differential exon usage and m5c and hm5c methylation status.

BAYESIAN CAUSAL INFERENCE OF GENE REGULATORY NETWORKS FROM CRISPR PERTURBATIONS IN CD4+ T CELLS

Joshua S Weinstock^{*1,2}, Maya Arce^{*3}, Jacob W Freimer^{1,3}, Mineto Ota^{1,3}, Alexander Marson#³, Alexis Battle#^{2,4}, Jonathan K Pritchard#^{1,5}

¹Stanford University, Genetics, Stanford, CA, ²Johns Hopkins University, Biomedical Engineering, Baltimore, MD, ³Gladstone-UCSF, Institute of Genomic Immunology, San Francisco, CA, ⁴Johns Hopkins University, Malone Center for Engineering in Healthcare, Baltimore, MD, ⁵Stanford University, Biology, Stanford, CA

*,# denote equal contribution

The effects of genetic variation on complex traits manifest predominately through the contribution of regulatory variation. Despite the importance of trans-regulatory variation, mapping trans-regulators based on natural genetic variation, including eQTL mapping, has been challenging due to small effects and limited sample size. Experimental perturbation approaches, which are unconstrained by natural selection, offer a complementary approach to mapping trans-regulators. We used CRISPR knockouts of 84 genes in primary CD4+ T cells to experimentally perturb an immune cell gene network. Our knockout gene set is comprised of three groups: 1) 24 upstream regulators of IL2RA; 2) 30 inborn error of immunity (IEI) disease transcription factors (TFs); 3) 30 background TFs that are matched in constraint and expression level to the IEI TFs, but without a known immune disease association. We then developed a novel Bayesian structure learning method to estimate the gene regulatory network, which in contrast to many causal inference approaches, is not constrained to directed acyclic graphs, enabling us to capture cyclic effects such feedback loops. We systematically characterized the differences between the IEI and background TFs, finding that the gene groups were highly interconnected, but that IEI TFs were much more likely to regulate immune cell specific machinery and autoimmune GWAS loci. We observed 211 directed edges among the 84 genes, which were validated using orthogonal data modalities, and could not be detected in existing CD4+ trans-eQTL data, demonstrating the value of experimental perturbations for mapping trans-regulators of highly constrained genes. We observed that three of the background TFs (BPTF, YBX1, and DR1) had more downstream effects than any of the IEI TFs and we characterize candidate novel gene members of canonical immune signaling pathways. We found that KMT2A had downstream effects very similar to genes in the JAK-STAT module, and that ZBTB14 is predicted to regulate the Rel sub-unit of NF- κ b. We observed that SNPs linked to our network using the Activity-By-Contact method were strongly enriched for heritability of autoimmune traits and used our network to characterize the regulation of autoimmune GWAS loci. We contribute to study design and methodology of arrayed CRISPR screens by demonstrating the value of including control TFs and using bespoke structure learning methods for network estimation. Our approach can provide a roadmap for mapping causal regulatory networks in other cell types and contexts.

PREDICTION OF PRIME EDITING INSERTION EFFICIENCIES
USING SEQUENCE FEATURES AND DNA REPAIR DETERMINANTS

Jonas Koeppe¹, Juliane Weller^{*1}, Elin Madli Peets^{*1}, Ananth Pallaseni¹,
Ivan Kuzmin², Uku Raudvere², Hedi Peterson², Fabio Liberante¹, Leopold
Parts^{1,2}

¹Wellcome Sanger Institute, Human Genetics Programme, Hinxton, United Kingdom, ²University of Tartu, Department of Computer Science, Tartu, Estonia

* authors contributed equally

Short sequences can be precisely written into a selected genomic target using prime editing without creating double-strand breaks. This facilitates correcting pathogenic deletions, controlling gene expression, modifying proteins, and many other exciting applications. However, it remains unclear what types of sequences prime editors can efficiently insert, and how to choose optimal reagents for a desired outcome.

To characterize features that influence insertion efficiency, we designed a library of 3,604 sequences up to 69 nt in length and measured their insertion frequency into four genomic sites in three human cell lines, using different prime editor systems. We discover that insertion sequence length, nucleotide composition and secondary structure all affect insertion rates, and that mismatch repair proficiency is a strong determinant for the shortest insertions. We also discover that 3' flap nucleases TREX1 and TREX2 suppress the insertion of longer sequences. Combining the sequence and repair features into a machine learning model, we can predict 70% of the repeatable variation in insertion frequency for new sequences. The tools we provide facilitate optimal design choices for inserting short sequences into genomes.

UNCOVERING GENE FUSIONS WITH 3D GENOMICS: FROM CLINICAL VALIDATION TO ACTIONABLE INSIGHTS FOR UNDIAGNOSABLE SOLID TUMORS

Allyson Whittaker¹, Kristin Sikkink¹, Anthony Schmitt¹, Kristyn Galbraith², Michelle Perez-Arreola¹, Misha Movahed-Ezazi², George Jour², Matija Snuderl²

¹Arima Genomics, Research, Carlsbad, CA, ²NYU School of Medicine, Pathology, New York, NY

Identifying gene fusions in tumor biopsies is critical for understanding disease etiology, however, clinical NGS panels often fail to yield clear genetic drivers. One challenge is that RNA-seq does not perform well in FFPE tissue blocks due to RNA degradation, low transcript abundance, and/or RNA panel design. To overcome this, we developed a novel DNA-based partner-agnostic approach for identifying fusions from FFPE tumors using 3D genomics based on Arima-HiC technology, in some cases with target enrichment (Capture-HiC), and NGS. Using this approach, we have profiled 108 FFPE tumors across tumor types. We first performed clinical validation of the Capture-HiC approach by re-analyzing FFPE tumors comprising 33 actionable gene fusions detected by the RNA-based NYU FUSION SEQer CLIA assay. We observed a 100% concordance (33/33) between Capture-HiC and RNA panels. We then analyzed 63 FFPE tumors using genome-wide HiC, including 36 CNS tumors, 10 gynecological sarcomas, and 12 solid hematological tumors (lymphoma / plasmacytoma). These tumors had no genetic drivers from prior CLIA-validated DNA and RNA panels. Amongst these, HiC analysis identified previously undetected fusions in 71% (45/63) of tumors. To attribute clinical significance to the fusions detected, we compared the genes implicated in our fusion calls with NCCN and WHO guidelines, and OncoKB, and assigned which tumors had a therapeutic level biomarker (e.g. PD-L1, NTRK, RAD51), or a diagnostic / prognostic biomarker (e.g. MYBL1 in glioma). We found 39.7% (25/63) of tumors had fusions involving a therapeutic level biomarker and a further 12.7% (8/63) had fusions involving a diagnostic or prognostic biomarker, indicating an overall diagnostic yield of 52.4%. The remaining 19% (12/63) had fusions of potential clinical significance, according to OncoKB. To highlight examples, we identified a novel PD-L1 rearrangement in a pediatric glioma that was not detected by DNA or RNA panels. Our finding was confirmed by PD-L1 IHC, and the patient was put on pembrolizumab off-label after tumor recurrence and has exhibited a complete response. We also identified MYBL1 fusions in two glioma cases that were previously missed by RNA panels. In one case, our MYBL1 fusion spared the patient unnecessary chemo post resection. Together, our findings demonstrate clinical validation, and highlights the utility for 3D genome profiling to increase diagnostic yield by finding clinically actionable fusions as a reflex test and as a frontline test in tumors without available NGS fusion assays (e.g. solid hematological tumors).

SPATIAL REGRESSION MODELS FOR THE ANALYSIS OF CHROMATIN CONFORMATION DATA

Wilfred Wong^{1,2}, Julian Pulecio³, Renhe Luo³, Nan Zhang³, Jielin Yan³, Effie Apostolou⁴, Danwei Huangfu³, Christina Leslie^{1,2}

¹Memorial Sloan Kettering Cancer Center, Computational and Systems Biology, New York, NY, ²Tri-Institutional Training Program, Computational Biology and Medicine, New York, NY, ³Sloan Kettering Institute, Developmental Biology, New York, NY, ⁴Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY

Chromatin conformation assays have been used to capture how the genome reorganizes across development, during disease, and in response to experimental perturbations. This data is frequently visualized as a contact matrix which contains information on how often two genomic loci undergo proximity ligation. On this contact matrix, the genome is partitioned into discrete bins, and these bins are used to identify loops and interactions, topologically associating domains, and A/B compartments for further analysis. Current treatments of the contact matrix, e.g., as an image or a multivariate time series, have partly obscured the fact that the counts are a realization of a stochastic spatial process and do not explicitly exploit the spatial structure of the data during their analysis. We propose to apply a class of spatial regression models to study the spatial variation that exists on the contact matrix to understand genome organization and reorganization. To that end, we will demonstrate how this approach can be used for estimating underlying spatial effects in the genome and use it as a basis for analyzing diverse chromatin conformation assays, such as HiC, HiCAR, and HiChIP, and identify similar chromatin reorganization events at developmentally important loci, such as at GATA6 and PDX1.

STRUCTURE-AWARE ANNOTATION OF SOLENOID PROTEIN DOMAINS

Boyan Xu¹, Daven Lim¹, Christopher J Tralie³, Alois Cerbu¹, Daniil Prigozhin^{1,2}, Ksenia Krasileva¹

¹UC Berkeley, Plant and Microbial Biology, Berkeley, CA, ²Ursinus College, Mathematics And Computer Science, Collegeville, PA, ³Lawrence Berkeley National Laboratory, Molecular Biophysics and Integrated Bioimaging, Berkeley, CA

Protein domain annotation is typically done by predictive models such as HMMs trained on sequence motifs. However, sequence-based annotation methods are prone to error, particularly in accurately calling domain boundaries and motifs within them. These methods are fundamentally limited by a lack of structural information accessible to the model. With the advent of deep learning-based protein structure prediction, we aim to leverage the geometry of protein structures to assist in domain annotation. In particular, we developed a novel dimensionality reduction method to annotate repeat units of the Leucine Rich Repeat (LRR) domain. The method is able to correct mistakes made by existing Machine Learning-based annotation tools. We also utilize topological methods to precisely determine start and end boundaries for the LRR solenoid region which sequenced-based methods cannot resolve. The methods are validated on 178 experimentally-derived LRR structures from Protein Data Bank, and applied to predicted structures of LRR-containing intracellular innate immune proteins in the model plant *Arabidopsis thaliana*.

STRUCTURAL VARIANTS DRIVE CONTEXT DEPENDENT ONCOGENE ACTIVATION IN CANCER

Zhichao Xu¹, Dong-Sung Lee⁴, Sahaana Chandran¹, Victoria T Le¹, Rosalind Bump¹, Jean Yasis¹, Sofia Dallarda¹, Samantha Marcotte¹, Benjamin Clock¹, Nicholas Haghani¹, Chae Yun Cho¹, Kadir Akdemir², Selene Tyndale³, P. Andrew Futreal², Graham McVicker³, Geoffrey M Wahl¹, Jesse R Dixon¹

¹Salk Institute for Biological Studies, Gene Expression Laboratory, La Jolla, CA, ²UT MD Anderson Cancer Center, Department of Genomic Medicine, Houston, TX, ³Salk Institute for Biological Studies, Integrative Biology Laboratory, La Jolla, CA, ⁴University of Seoul, Department of Life Sciences, Seoul, South Korea

Higher order chromatin structure plays a critical role in the regulation of genes by distal regulatory sequences. Structural variants that alter 3D genome organization can lead to enhancer-promoter rewiring and human disease, particularly in the context of cancer. However, it remains unclear how widespread structural variants that alter 3D genome structure are in cancer genomes and what genes are affected by such events. Furthermore, recent studies have shown that only in a small minority of structural variants are associated with altered gene expression. Therefore, it is unclear whether or how an individual structural variant may contribute to oncogene activation. To address these questions, we have analyzed Hi-C data and structural variants from 92 cancer cell lines and patient samples representing diverse cancer types. We identified loci affected by recurrent alterations to 3D genome structure, including loci containing oncogenes such as MYC, TERT, and CCND1. We also find evidence that these loci are frequently affected by structural variants predicted to alter 3D genome structure from whole genome sequencing datasets. Using CRISPR/Cas9 genome engineering to generate de novo structural variants, we show that “Activity-by-Contact” models predict the likelihood of oncogene activation in the context of structural variants. However, such Activity-by-Contact models are only predictive of specific subsets of genes in the genome, suggesting that different classes of genes engage in distinct modes of regulation by distal regulatory elements. These results indicate that structural variants that alter 3D genome organization are widespread in cancer genomes and begin to illustrate predictive rules for the interpretation of the consequences of non-coding structural variants on oncogene activation.

MAPPING THE *CIS*- AND *TRANS*-REGULATORY LANDSCAPE OF THE IMMUNE CHECKPOINT PD-L1 WITH PAIRED GENETIC SCREENS

Xinhe Xue^{1,2}, Zoran Gajic^{1,2}, Christina Caragine^{1,2}, Mateusz Legut^{1,2}, Conor Walker^{1,3}, James Kim^{1,2}, Hans-Hermann Wessels^{1,2}, Congyi Lu^{1,2}, Gamze Gursoy^{1,3}, Neville E Sanjana^{1,2}

¹New York Genome Center, New York, NY, ²New York University, Department of Biology, New York, NY, ³Columbia University, Department of Systems Biology, New York, NY

Functional genomics approaches such as pooled CRISPR screens or massively-parallel reporter assays typically focus on the discovery of either *cis*-regulatory elements (CREs) or *trans*-acting factors that control gene expression. However, gene regulation requires both *cis* and *trans* elements. Here, we develop a scalable, integrated approach that pairs discovery of *cis*- and *trans*-acting elements to experimentally elucidate the key noncoding regulatory elements and transcription factors that function at a given locus. We focus on regulation of the immune checkpoint PD-L1, whose expression on tumor cells serves as a clinically-actionable (and FDA-approved) diagnostic for checkpoint inhibitor therapy across multiple cancer types. In human pancreatic adenocarcinoma (PDAC) cells, we tiled the *PD-L1* locus (25,537 CRISPR perturbations spanning ~1 Mb) in basal and interferon-stimulated states and identified multiple enhancer-like CREs and insulator-like CREs. After quantifying the impact of CRE mutagenesis on PD-L1 protein expression, we co-cultured CRE-edited mesothelin-positive PDAC cells with primary T cells engineered with a chimeric antigen receptor (CAR-T) targeting mesothelin. CRE disruption not only impacts PD-L1 expression level in tumor cells but also alters cytokine secretion and cytotoxic activity of CAR-T cells, reinforcing the functional relevance of somatic noncoding mutations at this locus. Using 3D genome mapping, we show that a subset of these CREs physically contact the PD-L1 promoter, identifying a clear mechanistic basis for distal regulatory activity.

To further elucidate CRE mechanisms, we performed a pooled CRISPR screen targeting all ~2,000 transcriptional regulators in the human genome with 10 guide RNAs each. We identify both established PD-L1 regulators (*e.g.* SP1, STAT1 and IRF2) and several novel ones (*e.g.* SRF, BPTF and KMT2D). By integrating the two CRISPR screens, we detail complete regulatory circuits: For example, we find that the transcription factor SRF partners with a chromatin remodeler BPTF to regulate PD-L1 at a CRE located 4 kb upstream of PD-L1 promoter. Of clinical relevance, we find that multiple top-ranked PD-L1 *trans*-regulators, including SMAD4 and FOXA3, are recurrently mutated in PDAC patients ($n = 241$ tumors), altering PD-L1 expression *in vivo* and overall survival of patients treated with immune checkpoint therapies ($n = 437$ patients).

PREDICTING THE STRUCTURE AND FUNCTION OF ALTERNATIVE PROTEINS

Feriel Yala, Sébastien Leblanc, Xavier Roucou

University of Sherbrooke, Department of Biochemistry and Functional Genomics, Sherbrooke, Canada

According to conventional annotation rules, mRNAs contain a single protein-coding sequence (CDS) or functional open reading frame (ORF), usually the longest, and non-coding RNAs do not encode proteins. This canonical view of the coding genome which annotates a restricted set of proteins (reference proteins) has been challenged by proteogenomic approaches that have highlighted functional ORFs in “non-coding” regions of the transcriptome (UTRs of mRNA and non-coding RNAs) or overlapping canonical CDSs in a different reading frame. We developed a new annotation, termed OpenProt, that takes this complexity into account to explore the human genome more deeply. OpenProt was used to re-analyze large public datasets of mass spectrometry and ribosome profiling and found evidence for the expression of 77748 alternative proteins.

To accelerate research on their functions, we predicted the 3D structures of 557,568 alternative proteins annotated by OpenProt. We used machine learning tools, including AlphaFold and OmegaFold, to predict the structure of alternative proteins. AlphaFold relies on multiple sequence alignments and was used for proteins with sufficient evolutionary information (31% of alternative proteins), while OmegaFold was used for alternative proteins with no known homologs.

The results obtained from AlphaFold showed that 8.25% of the alternative proteins have an average pLDDT score lower than 50, suggesting the presence of a large number of intrinsically disordered regions (IDRs). To further assess whether alternative proteins are enriched in IDRs we used fIDPnn, a computational tool for the prediction of IDRs in protein sequences. Our analysis showed that IDRs are found in 16.68% of reference proteins and 23.24% of alternative proteins. The reference and alternative proteome contain 4.30% and 16.01% of fully disordered proteins (without secondary structure), respectively.

Access to a large number of predicted structures enables various in-silico studies useful to determine the function of alternative proteins. For example the prediction of protein-protein interactions by using the docking and high-throughput screening tools or by fingerprint technique using the MaSIF tool. Partner detection can then be validated by molecular dynamics. In addition, for each IDR prediction fIDPnn also indicates the probability of high-level functions such as protein binding, DNA binding or RNA binding. These will guide laboratory work for more precise functional characterization of individual candidates.

In-silico structure analysis of alternative proteins predicted by OpenProt has the potential to shed light on their function and accelerate research in proteomics.

EPIPHANY: PREDICTING THE HI-C CONTACT MAP FROM 1D EPIGENOMIC DATA

Rui Yang^{*1}, Arnav Das^{*2}, Vianne R Gao¹, Alireza Karbalayghareh¹, William S Noble³, Jeffrey A Bilmes², Christina S Leslie¹

¹Memorial Sloan Kettering Cancer Center, Computational & Systems Biology Program, New York City, NY, ²University of Washington, Department of Electrical and Computer Engineering, Seattle, WA,

³University of Washington, Department of Genome Sciences, Seattle, WA

Recent deep learning models that predict the Hi-C contact map from DNA sequence achieve promising accuracy, however, they fall short in terms of generalizing to new cell types and capturing cell-type-specific differences.

We propose Epiphany, a neural network that aims to build a bridge between cell-type specific epigenomic features and the chromatin 3D structure. Epiphany is initially trained with five epigenomic tracks that are already available in hundreds of cell types and tissues: DNase I hypersensitive sites and ChIP-seq for CTCF, H3K27ac, H3K27me3, and H3K4me3. The network employs 1D convolutional layers to learn local representations from the input tracks, bidirectional long short-term memory (Bi-LSTM) layers to capture long term dependencies along the epigenome, as well as a generative adversarial network (GAN) architecture to encourage contact map realism. To optimize the usability of predicted contact matrices, we trained and evaluated models using multiple normalization and matrix balancing techniques including KR, ICE, and HiC-DC+ Z-score and observed-over-expected count ratio. Epiphany is trained with a combination of MSE and adversarial (a GAN) loss to enhance its ability to produce realistic Hi-C contact maps for downstream analysis.

Epiphany shows robust performance and generalization to held-out chromosomes within and across cell types and species, and its predicted contact matrices yield accurate TAD and significant interaction calls. An ablation analysis with various subsets of epigenomic marks further revealed the relationship between 1D epigenomic marks and the 3D structures.

At inference time, Epiphany can be used to study the contribution of specific epigenomic peaks to 3D architecture and to predict the structural changes caused by perturbations of epigenomic signals. In the future, we hope that Epiphany can contribute to the study of epigenomic signals and genome 3D structure, and the enhancer-promoter (E-P) and promoter-promoter (P-P) interactions.

*Authors contributed equally.

HIGH LEVEL OF STRUCTURAL COMPLEXITY OF CANINE OLFACTORY RECEPTOR GENE FAMILIES REVEALED BY GENOME ASSEMBLIES OF SIX DOG BREEDS

Feyza Yilmaz, Kwondo Kim, Pille Hallast, Wonyeong Kang, Qihui Zhu,
Charles Lee

The Jackson Laboratory, Genomic Medicine, Farmington, CT

The exceptional sensitivity of the canine olfactory system is demonstrated by the usage of sniffer dogs in both military and non-military contexts to identify a variety of odors. Dogs' ability to smell is hypothesized to be influenced by genetic diversity in the olfactory receptor (OR) genes. With a mean of one single nucleotide polymorphism every 577 nucleotides, the OR genes are known to be highly polymorphic; nevertheless, it is unclear how structural variations (SVs) affect the degree of olfactory ability. We created highly contiguous chromosome-length de novo genome assemblies (mean QV=42.7) from six dog breeds with different sniffer capacities using data from PacBio CLR long-read sequencing and Bionano optical mapping. In order to identify and fully resolve complex structural variations of these gene loci that can affect sniffer dogs' ability to smell in different circumstances for specialized detections, we investigated SVs overlapping OR genes on 25 chromosomes. Protein-coding OR genes in our dog assemblies ranged in copy number from 598 to 668. On 21/25 of the chromosomes that were analyzed, allelic variants were discovered, the majority of which overlapped protein-coding OR gene copies. Furthermore, OR gene families on six chromosomes displayed complex variation, including tandem repeat expansions, duplications, deletions, and inversions. Strikingly, six chromosomes contain OR gene copies which are associated with cOR6C, which was shown to have an expansion to eight distant loci in the dog genomes, suggesting that variants in these loci might have an impact the effectiveness of odor recognition as well as a role for particular alleles in odor detection.

PROTEOMIC ANALYSES REVEAL MECHANISTIC LINKS
BETWEEN CLONAL HEMATOPOIESIS OF INDETERMINATE
POTENTIAL AND CORONARY ARTERY DISEASE

Zhi Yu^{1,2}, Bing Yu³, Amélie Vromman⁴, Ngoc Quynh H Nguyen³, Alexander G Bick^{1,5}, Benjamin L Ebert^{1,6,7}, Rajat M Gupta^{1,4}, Peter Libby⁴, Robert E Gerszten^{1,7,8}, Pradeep Natarajan^{1,2,7}

¹Broad Institute of MIT and Harvard, Program of Medical and Population Genetics and Cardiovascular Disease Initiative, Cambridge, MA, ²Massachusetts General Hospital, Cardiovascular Research Center and Center for Genomic Medicine, Boston, MA, ³Brigham and Women's Hospital, Division of Cardiovascular Medicine, Boston, MA, ⁴The University of Texas Health Science Center at Houston, Department of Epidemiology, Human Genetics, and Environmental Sciences, Houston, TX, ⁵Vanderbilt University Medical Center, Department of Medicine, Nashville, TN, ⁶Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, ⁷Harvard Medical School, Department of Medicine, Boston, MA, ⁸Beth Israel Deaconess Medical Center, Division of Cardiovascular Medicine, Boston, MA

Clonal hematopoiesis of indeterminate potential (CHIP) is a phenomenon where hematopoietic stem cells acquire leukemogenic mutations without blood cancer. CHIP is a causal risk factor for coronary artery disease (CAD). The plasma proteome may provide novel mechanistic insights into the links between CHIP and CAD. We studied the impact of CHIP on the plasma proteome and CAD risk in five multi-ancestry Trans-Omics for Precision Medicine (TOPMed) cohorts: ARIC (N=8,194), CHS (N=1,689), JHS (N=2,058), and MESA (N=976) for SomaScan-based discovery and WHI (N=1,166) for Olink-based replication. CHIP was identified from whole genome sequences of blood DNA and modeled both as a composite and, for the most common drivers (*DNMT3A*, *TET2*, and *ASXL1*), separately. Levels of serum proteins (~1,300 for SomaScan and ~550 for Olink) were log-transformed and adjusted for study-specific covariates. The association between CHIP and protein levels was estimated within each study and meta-analyzed. Functional validation was performed by comparing single-cell gene expression between *Tet2*-/- and wild-type mice. Pathway analysis and examination for shared and non-shared proteomic associations between CHIP and CAD were conducted among proteins nominally associated with CHIP. Across all studies (mean age: 66.1 y; 43.8% male; 37.2% non-European ancestry), 766 (6.0%) individuals were identified with CHIP. We discovered 42 significant CHIP-protein pairs (FDR<0.05), with pappalysin-1 and carbonic anhydrase 1 being the most significantly associated proteins. *TET2* had the largest number of, and strongest, associations among examined driver genes. Twelve CHIP-protein pairs were replicated cross-platform at the nominal threshold. Among top *TET2*-associated proteins, we observed sex-specific differential expression of genes encoding six proteins in monocytes or B-cells between *Tet2*-/- and wild-type mice. Pathway analysis implicated driver-gene-specific inflammation and signaling pathways. Eighty-four proteins (including PCSK9 and interleukin-1 receptor antagonist) were significantly associated with both CHIP and CAD at the nominal threshold, suggesting potential shared mechanisms. CHIP, particularly *TET2*, is broadly associated with the plasma proteome, as validated by evidence in *Tet2*-/- mice. Further studies are needed to understand the biological mechanisms linking CHIP to CAD.

ACCURATE QUANTIFICATION OF MULTI-MAPPING READS IN SINGLE-CELL RNA-SEQ WITH STARSOLO

Dinar Yunusov¹, Nathan Castro-Pacheco^{1,2}, Alexander Dobin¹

¹Cold Spring Harbor Laboratory, Cancer Center, Cold Spring Harbor, NY,

²Scale Biosciences, Inc., Boston, MA

Accurate gene expression quantification is crucial for interpretation of single-cell RNA sequencing (scRNA-seq) data. Multi-gene reads - reads that map to multiple genes and can arise from closely related paralogs, gene fusions, alternative splicing, and read-through transcription events - are especially challenging to map and quantify. While accounting for up to 40% of mapped reads, they are usually discarded, which can result in a loss of important biological information unique to individual cells or cell types. To address this issue, various methods have been developed, ranging from filtering based on mapping quality to more sophisticated algorithms that use statistical models and reference databases to resolve ambiguities.

Of popular tools for scRNA-seq mapping and quantification, CellRanger, a proprietary tool from 10X Genomics company, excludes multi-gene reads from analysis completely, while STARSolo, Alevin-fry and Kallisto/bustools can parse multi-gene reads with Expectation-Maximization Maximum Likelihood (EM-ML) algorithms. However, these algorithms do not consider biases introduced during scRNA-seq library preparation, such as uneven distribution of RNA-seq reads across the gene length.

We enhanced the functionality of STARSolo with a novel EM-ML algorithm that accounts for idiosyncrasies of certain scRNA-seq protocols, such as 3' or 5' end cloning biases in the 10X Genomics Chromium 3' and 5' library construction kits. For each cell, reads are divided into non-overlapping bundles based on the transcripts to which they map. Since reads within each bundle map to an exclusive set of transcripts, the EM-ML transcript quantification step is performed independently and in parallel for each bundle. Transcript quantifications are then aggregated at the gene level. Flexible barcode/UMI processing scheme allows this algorithm to be used with datasets produced with other established and emergent technologies.

We show that inclusion of multi-gene reads significantly impacts cell clustering, differential expression analysis, and identification of marker genes in the identified clusters. Better resolution of overlapping genes with STARSolo results in assignment of reads to well-annotated, rather than to provisionally annotated genes. In turn, information recovered from multi-gene reads results in STARSolo strongly outperforming the standard EM-ML, particularly when analyzing small cell populations. Finally, we demonstrate that STARSolo robustly quantifies gene expression even in the presence of large numbers of reads from pre-mRNA sequences, successfully overcoming this common obstacle in scRNA-seq analysis.

UNDERSTANDING CAUDAL DEVELOPMENTAL ABNORMALITIES USING SINGLE-NUCLEUS MULTI-OMICS DATA FROM WILD TYPE AND DANFORTH'S SHORT TAIL MOUSE E9.5 TAILBUDS

Cynthia K Zajac¹, Ricardo D Albanus^{1,4}, Nandini Manickam^{1,2}, Erika Curka^{1,2}, Catherine Keegan^{2,3}, Stephen C Parker^{1,2,5}

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

³University of Michigan, Pediatrics, Ann Arbor, MI, ⁴Washington University, NeuroGenomics and Informatics Center, St. Louis, MO, ⁵University of Michigan, Biostatistics, Ann Arbor, MI

Background: Caudal embryo development is important for multiple organ systems arising from the three primary cell layers. Embryonic malformations associated with caudal birth defects affect 1 in 10,000 human live births, but their mechanisms are largely unknown. The Danforth's short tail (Sd) mouse exhibits cessation of the vertebral column at the lumbar level and malformations of urogenital and gastrointestinal organs, providing a tractable model for the etiology of human caudal birth defects. The Sd mutation is caused by an endogenous retroviral (ERV) insertion upstream of the *Ptf1a* gene in the promoter region of the lncRNA gene *Gm13344*. This region is orthologous to a human pancreatic developmental enhancer, which provides additional motivation for using the Sd mouse to study caudal dysgenesis. **Methods:** To better understand the cell-specific regulatory landscape of the Sd mouse, we generated single-nucleus multi-omics (ATAC+RNA) data from control (WT) and Sd E9.5 tailbuds. We used a custom pipeline for droplet QC before clustering based on the RNA modality. Then, we performed single nuclei 5' RNA sequencing (sn-5') using the same sample types and applying similar quality control. We additionally performed SCAFE (Single-Cell Analysis of Five-prime Ends) on this dataset to visualize gene expression and identify transcription start sites and transcribed cis-regulatory elements. **Results:** For our dual-modality data set, we obtained 3,544 (WT) and 3,391 (Sd) high-quality nuclei. We generated 11 clusters at the RNA level that grouped together cells expressing important gene markers for embryonic development such as Sonic Hedgehog (*Shh*). For our sn-5' dataset, we obtained 5,804 (WT) and 9,768 (Sd) high-quality nuclei for joint clustering analyses. We found 18 clusters representing different cell types, some of which have differential chromatin accessibility and are linked to caudal embryonic developmental gene markers including *Ptf1a*, *Shh*, and *Gm13344*. In addition, we validated these findings using the SCAFE results from our sn-5' dataset, where we observed co-expression of *Gm13344* and *Ptf1a* solely in a restricted subset of clusters in the Sd sample. **Future direction:** Preliminary results provide an insight to cluster characterization. We identified a subset of cells that may represent the early cellular events that are causal to the Sd caudal phenotype. The next steps in this project involve further integration of the 5' RNA and chromatin accessibility modalities to identify transcription factors perturbed as a result of the Sd mutation. We expect these integrative analyses will improve our understanding of the causal cell types and regulatory networks that influence the Sd phenotype and, consequently, caudal developmental disorders in humans.

DENISOVAN AND NEANDERTAL GENE VARIANTS INFLUENCE BRAIN MORPHOLOGY IN PRESENT-DAY PEOPLE

Hugo Zeberg^{1,2}

¹Karolinska Institutet, Dept. of Physiology and Pharmacology, Stockholm, Sweden, ²Max Planck Institute for Evolutionary Anthropology, Dept. of Evolutionary Genetics, Leipzig, Germany

While there are few records of Denisovan remains, most anatomical traits of archaic humans would likely fall within the range of variation of present-day humans. No single human living today, however, would have all anatomical traits of a Neandertal or a Denisovan. Since there are no archaic soft tissue remains we cannot study the brain morphology of archaic humans directly, although differences in braincase anatomy are likely to be reflected in brain anatomy. The gene-flow that took place from archaic to modern humans provides us with an opportunity to test if there is any effect of some archaic gene variants on the morphology of people living today. Here we use two brain imaging dataset comprising 33,224 genotyped British individuals (Elliott et al., 2018) and 7,058 genotyped Chinese individuals (Yu et al., 2022). We investigated 62 brain volume phenotypes, 62 brain area phenotypes and 135 white matter fiber tract phenotypes. We tested if any alleles associated with any of the phenotypes co-segregated among Brits or Chinese people with alleles shared among Neandertals or Denisovans while not being seen among 108 Yoruban individuals. Correcting for the number of independent archaic-like alleles tested and the number of phenotypes, we find two introgressed Denisovan haplotypes with associations in the Chinese dataset ($p < 8.0\text{e-}8$) and five introgressed Neandertal haplotypes with associations in the British dataset ($p < 2.2\text{e-}8$). Two of the Neanderthal haplotypes were associated with changes in brain volume in specific regions, while four Neanderthal haplotypes affected white matter fiber tract phenotypes. A Neandertal haplotype in the *PRDM5* locus increased volume of the middle temporal gyrus and a Neandertal haplotype in the *DAAM1* locus was associated with increased volume of the parietal lobe and a decreased volume of the occipital lobe. The two Denisovan haplotypes in the Chinese dataset both affected white matter fiber tract phenotypes; a haplotype in *NPAS2* locus influencing the external capsule and a haplotype in the *RBM17* locus influencing the posterior limb of the internal capsule. The data presented here shed light on the genetic basis of differences in brain morphology between archaic and modern humans.

References

Yu, C., et al. (2022) Trans-ancestral genome-wide association studies of brain imaging phenotypes, *Research Square* (preprint version 1). Data from <http://chimgen.tmu.edu.cn/> [Aug 16, 2022]

Elliott, L.T., et al. (2018) Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 562, 210–216. Data from <https://open.win.ox.ac.uk/ukbiobank/big40/> [Aug 16, 2022]

GENETIC PERTURBATION OF PU.1 BINDING IN MICROGLIA AND DISEASE RISK ASSOCIATION

Falak Sher*, Lu Zeng*, Hans-Ulrich Klein, Julie J McInvale, Philip L De Jager

Columbia University Irving Medical Center, Department of Neurology and the Taub Institute for Research on Alzheimer's disease and the Aging brain, New York, NY

Microglia are the resident myeloid cells of the central nervous system (CNS), which play critical roles in neuroinflammation and the etiology of Alzheimer's disease (AD), yet our understanding of how the genetic regulatory landscape controls microglial function is limited, hampering therapeutic development.

In this study, we profile the binding of myeloid master regulator PU.1 from the dorsolateral prefrontal cortex of 167 individuals of advanced age using ChIP-Seq. 54,699 PU.1 peaks were identified genome-wide using MACS2, and they show the expected enrichment in chromosomal segments annotated as enhancers in microglia. We then related these data to single nucleus RNA-Seq data from the same individuals, and, to examine the functional consequences of PU.1 binding on AD and AD endophenotypes such as the accumulation of amyloid- β and tau proteinopathy. We also used genome-wide association study (GWAS) results for AD (N=1,126,563) along with genetic instruments for inferring the transcription of 9,495 PU.1 peaks in microglia to perform Epigenome-Wide Association Study (EWAS) of AD.

Our results show that differential PU.1 binding may drive the differentiation of 6 out of 15 microglial cell subpopulations, especially subpopulations which are homeostatic-like and reactive-like, suggesting that this transcription factor may have a more targeted role beyond the one involved in myeloid differentiation in general. In addition, we found amyloid- β proteinopathy and cognitive decline but not tau proteinopathy are broadly associated with differential PU.1 binding, signaling that PU.1 has a preferential role in amyloid- β pathology and may not have a strong role in the separate role of microglia in tau pathology. Integrative analysis using the EWAS strategy identified 106 peaks associated with AD susceptibility in 55 loci. Integrating PLAC-seq data from a previous study, we detect and quantify these PU.1 peaks anchored at genomic regions bound by promoters or enhancers of 19 genes, including 15 known AD genes and 4 new candidate risk genes for AD: DCTPP1, ARMC5, SCARF1 and CIRBP. Collectively, these results demonstrate a role for PU.1 in a subset of the many, heterogeneous roles that microglia play in AD, suggesting that it will be particularly important in the early events (amyloid- β deposition) leading to AD, and it directs our attention to a subset of AD susceptibility that may be exerting their role through an effect on differential PU.1 binding to cause accumulation of amyloid proteinopathy.

MODEL-BASED CHARACTERIZATION OF THE EQUILIBRIUM DYNAMICS OF TRANSCRIPTION INITIATION AND PROMOTER-PROXIMAL PAUSING IN HUMAN CELLS

Yixin Zhao¹, Lingjie Liu², Adam Siepel^{1,2}

¹Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, ²Stony Brook University, Graduate Program in Genetics, Stony Brook, NY

In metazoans, both transcription initiation and the escape of RNA polymerase (RNAP) from promoter-proximal pausing are key rate-limiting steps in gene expression. These processes play out at physically proximal sites on the DNA template and appear to influence one another through steric interactions, leading to a complex dynamic equilibrium in RNAP occupancy of the ~100 bp immediately downstream of the transcription start site. In this article, we examine the dynamics of these processes using a combination of statistical modeling, simulation, and analysis of real nascent RNA sequencing data. We develop a simple probabilistic model that jointly describes the kinetics of transcription initiation, pause escape, and elongation, and the generation of nascent RNA sequencing read counts under steady-state conditions. We then extend this initial model to allow for variability across cells in promoter-proximal pause site locations and steric hindrance of transcription initiation from paused RNAPs. In an extensive series of simulations over a broad range of parameters, we show that this model enables accurate estimation of initiation and pause-escape rates even in the presence of collisions between RNAPs and variable elongation rates. Furthermore, we show by simulation and analysis of data for human cell lines that pause-escape is often more strongly rate-limiting than conventional “pausing indices” would suggest, that occupancy of the pause site is elevated at many genes, and that steric hindrance of initiation can lead to a pronounced reduction in apparent initiation rates. Our modeling framework is generally applicable for all types of nascent RNA sequencing data and can be applied to a variety of inference problems. Our software for estimation and simulation is publicly available.

MULTIOMIC ANALYSIS REVEALS CELLULAR AND EPIGENETIC PLASTICITY IN INTESTINAL POUCHES OF ULCERATIVE COLITIS PATIENTS

Yu Zhao^{*3}, Ran Zhou^{*1}, Bingqing Xie¹, Cambrian Y Liu¹, Martin Kalski¹, Candace M Cham¹, Jason Koval¹, Christopher R Weber¹, Jingwen Xu¹, David T Rubin¹, Mitch Sogin⁴, Sean Crosson⁵, Jun Huang³, Aretha Fiebig⁵, Sushila Dalal¹, Eugene B Chang¹, Anindita Basu¹, Sebastian Pott²

¹University of Chicago, Department of Medicine, Division of the Biological Sciences, Chicago, IL, ²University of Chicago, Division of the Biological Sciences, Chicago, IL, ³University of Chicago, Pritzker School of Molecular Engineering, Chicago, IL, ⁴Marine Biological Laboratory, Woods Hole, MA, ⁵Michigan State University, East Lansing, MI

Total proctocolectomy with ileal pouch anal anastomosis (IPAA) is the standard treatment for severe treatment resistant ulcerative colitis (UC). Despite the significant improvements in patient outcomes, up to 50% of patients suffer from UC-like inflammation of the pouch (pouchitis). Here, we performed single-cell accessible chromatin (scATAC-seq) and single-cell RNA (scRNA-seq) profiling of paired biopsy samples from the ileal pouch and ileal segment above the pouch (prepouch) of UC-IPAA patients without clinical symptoms, to assess their cellular composition, and their transcriptomic and epigenetic profiles. We distinguished 47 cell types across epithelial, immune and mesenchymal lineages, and further identified cell-type-specific gene expression, cis-regulatory elements, their target genes and putative regulatory factors. Remarkably, we identified a mature enterocyte population that is specific to pouch, exhibiting substantial expression of colonic marker genes like CEACAM5 and corresponding chromatin accessibility signatures, while lacking most of the typical ileal features. Through examining splicing kinetics in epithelial lineage together with chromatin accessibility, we identified two main differentiation trajectories for intestinal stem cells in the pouch, and identified transcriptional and epigenetic regulators that underlie these two paths (HNF4G and CDX2, respectively). Together, we present a comprehensive transcriptomic and epigenetic atlas of cell populations in the pouch, and provide a roadmap for understanding the underlying molecular mechanisms of pouchitis.

GENOME IN A BOTTLE BENCHMARKS IN THE ERA OF COMPLETE HUMAN GENOMES

Nathan D Olson¹, Justin Wagner¹, Nathan Dwarshuis¹, Adam English², Fritz J Sedlazeck², Justin M Zook¹, Genome in a Bottle Consortium¹

¹National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD, ²Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

Recent advances in genome sequencing, assembly, and polishing have made high-quality, nearly-complete diploid genome assemblies possible. The Genome in a Bottle Consortium (GIAB) is using these diploid assemblies to generate benchmark sets for evaluating accuracy of challenging small variants and structural variants. Previous versions of the GIAB benchmark sets relied primarily on variant calls from mapped reads. These mapping-based benchmark sets excluded regions with large variants and complex variants, excluding ~8% of the GRCh38 reference genome plus 7% of sequence missing from the reference. Initial exploratory work in assembly-based variant calling provided promising results, first with the development of a benchmark for the difficult, highly variable but medically relevant MHC region, then with a benchmark set targeting 273 medically relevant genes not sufficiently covered by our existing mapping-based benchmark sets. Building off this initial work we developed a framework for generating and evaluating assembly-based benchmarks sets. We form these benchmarks using the assembly-based variant caller dipcall, followed by excluding regions with potential assembly errors, alignment errors, and variant types problematic for current benchmarking tools. This framework enables the automated generation of small and structural variant draft benchmark sets for high-quality diploid assemblies aligned to multiple references (GRCh37, GRCh38, and T2T-CHM13). We are currently working with the GIAB community to evaluate draft assembly-based benchmarks for tandem repeats and for X and Y chromosomes, the latter derived from complete chromosome assemblies. As we include increasingly complex variants in the benchmark, we are also developing new tools to compare these variants, e.g., enabling robust benchmarking in tandem repeats. To systematize evaluation of the benchmarks, we are piloting a new active evaluation approach to target curations. These new curated benchmarks will be valuable as the community moves towards variant calling in the most challenging regions, spurring development of increasingly accurate sequencing technologies and bioinformatics methods.

NOTES

NOTES

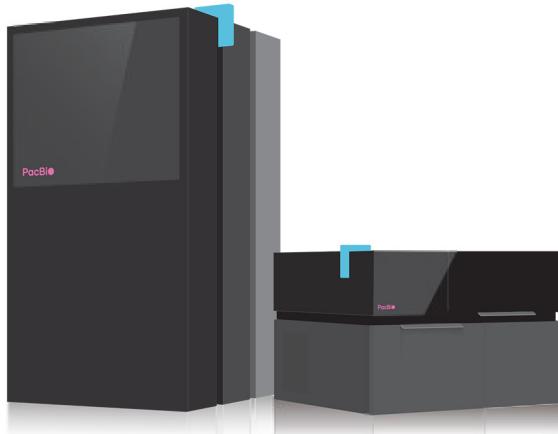
NOTES

NOTES

NOTES

NOTES

Say hello to Revio + Onso



Get ready to make discoveries that will change the world

Scan the QR code and explore the unparalleled accuracy of our new long-read and short-read sequencing systems.



Participant List

Mr. Temidayo Adeluwa
University of Chicago
temi@uchicago.edu

Dr. Enis Afgan
Johns Hopkins University
afgane@gmail.com

Dr. Saumya Agrawal
RIKEN
saumya.agrawal@riken.jp

Richard Agren
Karolinska Institutet
richard.agren@ki.se

Francois Aguet
Illumina, Inc.
faguet@illumina.com

Robin Aguilar
University of Washington
eaguil@uw.edu

Mr. Matthew Aguirre
Stanford University
magu@stanford.edu

Ms. Rhea Ahluwalia
University of Toronto
rhea.ahluwalia@mail.utoronto.ca

Mr. Omar Ahmed
Johns Hopkins University
omaryfekry@gmail.com

Dr. Keiko Akagi
The University of Texas MD Anderson
Cancer Center
kakagi@mdanderson.org

Yo Akiyama
The Broad Institute of MIT and Harvard
yakiyama@broadinstitute.org

Dr. Nirmala Akula
NIH, NIMH
akulan@mail.nih.gov

Dr. Frank Albert
University of Minnesota
falbert@umn.edu

Mr. Jamie Allen
EMBL-EBI
jma@ebi.ac.uk

Nathan Anderson
University of Wisconsin Madison
nathan.wally.anderson@gmail.com

Ms. Mathilde Andre
University of Tartu
mathilde.andre.student@gmail.com

Jeremy Arbesfeld
The Ohio State University
arbesfeld.1@osu.edu

Dr. Kristin Ardlie
Broad Institute of Harvard and MIT
kardlie@broadinstitute.org

Dr. Peter Arndt
Max Planck Institute for Molecular Genetics
arndt@molgen.mpg.de

Peter Audano
The Jackson Laboratory for Genomic
Medicine & UCHC
peter.audano@jax.org

Ms. Chiara Auwerx
University of Lausanne
chiara.auwerx@unil.ch

Randi Avery
University of Minnesota
randia@umn.edu

Dr. Doris Bachtrog
University of California, Berkeley
dbachtrog@berkeley.edu

Dr. Nick Banovich
Translational Genomics Research Institute
nbanovich@tgen.org

Prof. Elizabeth Bartom
Northwestern University at Chicago
ebartom@northwestern.edu

Dr. Alexis Battle
Johns Hopkins University
ajbattle@jhu.edu

Dr. Philippe Batut
Princeton University
pbatut@princeton.edu

Dr. Christine Beck
The Jackson Laboratory for Genomic
Medicine & UCHC
christine.beck@jax.org

Dr. Kynon Benjamin
Lieber Institute for Brain Development
jbenja13@jh.edu

Prof. Minou Bina
Purdue University
bina@purdue.edu

Gage Black
University of Utah
gage.black@utah.edu

Prof. Ran Blekhman
University of Chicago
blekhman@uchicago.edu

George Boateng-Sarfo
North Dakota State University
george.boateng@ndsu.edu

Dr. Stephanie Bohaczuk
University of Washington
bohaczuk@uw.edu

Dr. Małgorzata Borczyk
Maj Institute of Pharmacology PAS
malgorzata.m.borczyk@gmail.com

Dr. Christelle Borel
University of Geneva
christelle.borel@unige.ch

Ms. Caroline Bridge
Ontario Institute for Cancer Research
cbridge@oicr.on.ca

Dr. Andrew Brown
University of Dundee
a.z.t.brown@dundee.ac.uk

Dr. Reuben Buckley
National Human Genome Research
Institute
reuben.buckley@nih.gov

Ms. Cassandra Buzby
New York University
cb4097@nyu.edu

Dr. Francisco Callejas Hernandez
New York University
fcallejas@nyu.edu

Ms. Maria Elena Campoy Garcia
Universitat Autònoma de Barcelona
elena.campoy@uab.cat

Dr. Samuele Cancellieri
University of Trento
samuelec113@gmail.com

Dr. Han Cao
Dimension Genomics Inc
han@dimensiongen.com

Rodrigo Castro
The Jackson Laboratory
rodrigo.castro@jax.org

Dr. Taylor Cavazos
Exai Bio
taylorc@exai.bio

Dr. Monika Cecheva
UC Santa Cruz Genomics Institute
mcechova@ucsc.edu

Dr. Kevin Celestrin
Arima Genomics
kevin@arimagenomics.com

Dr. Weizhong Chang
Leidos Biomedical Research
weizhong.chang@nih.gov

Katherine Chao
Broad Institute of MIT and Harvard
kchao@broadinstitute.org

Mr. Yuhe Cheng
UCSD
yuc211@eng.ucsd.edu

Dr. Sagar Chhangawala
eGenesis
sagar.chhangawala@egenesisbio.com

Prof. Hyun Hoon Chung
Seoul National University Hospital
chhkmj1@snu.ac.kr

Ms. Erin Kim Chung
EMBL Heidelberg
erin.chung@embl.de

Dan Ciotlos
University of Michigan
dciotlos@umich.edu

Mr. Francesco Cisternino
Human Technopole
francesco.cisternino@fht.org

Dr. Melina Claussnitzer
Broad Institute, MGH / Harvard Medical School
melina@broadinstitute.org

Dr. Dorothy Clyde
Springer Nature
d.clyde@nature.com

Kelly Cochran
Stanford University
kcochran@stanford.edu

Dr. Charles Cole
Juniper Genomics
charles@junipergenomics.com

Mr. Jared Cole
University of Texas at Austin
jmcole@utexas.edu

Dr. Natacha Comandante-Lou
Columbia University Irving Medical Center
nc3018@cumc.columbia.edu

Mr. Mark Consugar
Cambridge Epigenetix
mark.consugar@cegx.co.uk

Dr. Laura Cook
Lawrence Berkeley National Laboratory
lecook@lbl.gov

Isabelle Cooperstein
University of Utah School of Medicine
isabelle.cooperstein@genetics.utah.edu

Dr. Tim Coorens
Broad Institute of MIT and Harvard
tcoorens@broadinstitute.org

Prof. Montserrat Corominas
Universitat de Barcelona
mcorominas@ub.edu

Ms. Christina Costa
New York University
cec701@nyu.edu

Dr. Robert Crawford
Cambridge Epigenetix
robert.crawford@cegx.co.uk

Dr. Paidi Creed
Cambridge Epigenetix
paidi.creed@cegx.co.uk

Dr. Alessandro Crnjar
Cold Spring Harbor Laboratory
crnjar@cshl.edu

Michael Cuoco
University of California, San Diego
mcuoco@ucsd.edu

Dr. Hannah Currant
University of Copenhagen
hannah.currant@cpr.ku.dk

Prof. Christina Curtis
Stanford University
cncurtis@stanford.edu

Ms. Helena Beatriz da Conceicao
University of Sao Paulo/Hospital Sírio Libanes
hconceicao@mochsl.org.br

Mr. Yifan Dai
University of Copenhagen
yifan.dai@bio.ku.dk

Prof. Mehdi Damaghi
Stony Brook University
mehdi.damaghi@stonybrookmedicine.edu

Mr. Maitreya Das
The Pennsylvania State University
mud367@psu.edu

Mr. Arun Das
Johns Hopkins University
arun.das@jhu.edu

Dr. Kushan Dasgupta
UCLA
kushan@g.ucla.edu

Monica Dave
UConn Health Center
modave@uchc.edu

Mr. David Davtian
University of Dundee
2400580@dundee.ac.uk

Dr. Carl de Boer
University Of British Columbia
carl.deboer@ubc.ca

Dr. Michiel de Hoon
RIKEN
michiel.dehoon@riken.jp

Dr. Jennifer DeLeon
Genome Research, Assistant Editor
deleon@cshl.edu

Dr. Ahmet Denli
Genome Research, Associate Editor
denli@cshl.edu

Dr. Alisha Dhiman
Purdue University
dhimana@purdue.edu

Sarah Djeddi
Boston Childrens Hospital
sarah.djeddi@childrens.harvard.edu

Tristram Dodge
Stanford University
tododge@stanford.edu

Dr. Julia Domingo
New York Genome Center
jdomingo@nygenome.org

Dr. Elisa Donnard
Broad Institute
edonnard@broadinstitute.org

Mr. Theo Dupuis
University of Dundee, School of Medicine
2395453@dundee.ac.uk

Dr. Nathan Dwarshuis
National Institute of Standards and
Technology
njd2@nist.gov

Dr. Scott Edmunds
BGI Hong Kong
scott@gigasciencejournal.com

Ms. Melise Edwards
University of Massachusetts Amherst
medwards@umass.edu

Dr. Evan Eichler
University of Washington & HHMI
eee@gs.washington.edu

Dr. Hjorleifur Einarsson
University of Copenhagen
hjorleifur.einarsson@bio.ku.dk

Dr. Jesse Engreitz
Stanford University
engreitz@stanford.edu

Basak Eraslan
Genentech
eraslab1@gene.com

Dr. Jose Espejo Valle-Inclan
EMBL-EBI
jespejo@ebi.ac.uk

Xiao Fan
University of Florida
xiaofan@ufl.edu

Dr. Kyle Farh
Illumina
kfarh@illumina.com

Christiana Fauci
Duke University
cf189@duke.edu

Ms. Lynn Fellman
Fellman Studio Inc.
lynn@fellmanstudio.com

Dr. Adam Felsenfeld
National Institutes of Health
adam_felsenfeld@nih.gov

Mr. Jiawu Feng
IPK Gatersleben
fengj@ipk-gatersleben.de

Dr. Yayan Feng
Cleveland clinic
fengy2@ccf.org

Mr. Juan Carlos Fernandez del Castillo
Broad Institute of MIT and Harvard
jfernand@broadinstitute.org

Ardian Ferraj
The Jackson Laboratory for Genomic
Medicine & UCHC
ardian.ferraj@jax.org

Dr. Douglas Fowler
University of Washington
dfowler@uw.edu

Eden Francoeur
The Jackson Laboratory for Genomic
Medicine & UCHC
eden.francoeur@jax.org

Dr. Jonathan Frazer
Centre for Genomic Regulation
jonathan.frazer@crg.eu

Max Frenkel
University of Wisconsin-Madison
mfrenkel@wisc.edu

Ms. Amelie Fritz
Technical University Denmark
ameli@dtu.dk

Mr. Peter Fromen
Cambridge Epigentix
peter.fromen@cegx.co.uk

Dr. Naoko Fujito
National Institute of Genetics
nfujito@nig.ac.jp

Prof. Julien Gagneur
Technical University of Munich
gagneur@in.tum.de

Dr. Pedro Galante
Hospital Sirio-Libanes
pgalante@mochsl.org.br

Dr. Irene Gallego Romero
University of Melbourne
irene.gallego@unimelb.edu.au

Dr. Hong Gao
Illumina, Inc.
hgao@illumina.com

Manik Garg
AstraZeneca UK Ltd
manik.garg1@astrazeneca.com

Dr. Erik Garrison
University of Tennessee Health Science
Center
erik.garrison@gmail.com

Dr. Kristina Garske
Princeton University
kg8086@princeton.edu

Ms. Rylee Genner
NIH
gennerrm@nih.gov

Prof. Ilias Georgakopoulos-Soares
Penn State College of Medicine
georgakopoulos.soares@gmail.com

Dr. Yoav Gilad
University of Chicago
gilad@uchicago.edu

Dr. Miri Gitik
NIH/NIMH
miri.gitik@nih.gov

Dr. Nicholas Gladman
USDA-ARS
gladman@cshl.edu

Dr. Michael Goldberg
University of Utah
megoldberg2@gmail.com

Alejandro Gomez
University of North Carolina - Chapel Hill
alegomez@ad.unc.edu

Ms. Ruth Gomez Graciani
Universitat Autonoma de Barcelona
g.g.ruth020695@gmail.com

Saideep Gona
University of Chicago
sgona@uchicago.edu

Dr. Angela Goncalves
German Cancer Center (DKFZ)
a.goncalves@dkfz.de

Ms. Sayeh Gorjifard
University of Washington
sgorji@uw.edu

Dr. Sager Gosai
The Broad Institute
sgosai@broadinstitute.org

Samantha Graham
University of Minnesota
graha880@umn.edu

Brenton Graveley
UConn Health Center
graveley@uchc.edu

Marta Greedy Escudero
Copenhagen University
marta.escudero@bio.ku.dk

Dr. Jonathan Griffiths
Altos Labs
jgriffiths@altoslabs.com

Dr. Jeremy Grushcow
Juniper Genomics
jeremy@junipergenomics.com

Dr. Andreas Gschwind
Stanford University
andreas.gschwind@stanford.edu

Mr. Xavi Guitart
University of Washington
guitarfx@uw.edu

Dr. Rodrigo Gularte Merida
MSKCC
gularter@mskcc.org

Xinyi Guo
New York Genome Center
xg740@nyu.edu

Dr. Kerstin Haase
Charite Universitätsmedizin Berlin
kerstin.haase@charite.de

Mr. Yasuhiko Haga
The university of Tokyo
7364133817@edu.k.u-tokyo.ac.jp

Dr. Bo Han
Intellia Therapeutics
bo.han@intelliatx.com

Mr. Nick Harding
Biomodal
nick.harding@biomodal.com

Dr. Arbel Harpak
The University of Texas at Austin
arbelharpak@utexas.edu

Dr. Ricardo Harrington
Massachusetts General Hospital
rharripon@mgh.harvard.edu

Samuel Hart
University of Washington
sfhart33@gmail.com

Prof. Shinichi Hashimoto
Kanazawa University
hashimot@wakayama-med.ac.jp

Dr. Gareth Hawkes
University of Exeter
g.hawkes2@exeter.ac.uk

Mr. Chenjun He
Stony Brook University
chenjun.he@stonybrook.edu

Laurel Hiatt
University of Utah
laurel.hiatt@hsc.utah.edu

Remy Hilsabeck
Frederick National Laboratory for Cancer Research
leidosbiomedtravel@nih.gov

Dr. Benjamin Hitz
Stanford University
hitz@stanford.edu

Dr. Hopi Hoekstra
Harvard University
hoekstra@oeb.harvard.edu

Dr. James Holt
Pacific Biosciences
mholt@pacificbiosciences.com

Ms. Mian Horvath
UConn Health Center
horvath@uchc.edu

Dr. Yuan Hou
Lerner Research Institute, Cleveland clinic
houy2@ccf.org

Kathleen Houlahan
Stanford University
khoulaha@stanford.edu

Dr. PingHsun Hsieh
University of Minnesota
hsiehph@umn.edu

Dr. Zheng Hu
Chinese Academy of Sciences
zheng.hu@siat.ac.cn

Dr. Mengqi Huang
University of Pittsburgh
huangmenqi@gmail.com

Prof. Kuan-lin Huang
Icahn School of Medicine at Mount Sinai
Kuan-lin.Huang@mssm.edu

Dr. Yi Huang
Broad Institute of MIT and Harvard
huangy@broadinstitute.org

Billy Huggins
UConn Health
bhuggins@uchc.edu

Mr. King Hung
Stanford University
khung@stanford.edu

Dr. Carolyn Hutter
NIH/NHGRO
carolyn.hutter@nih.gov

Dr. Oleg Iartchouk
Saliogen
o.iartchouk@saliogen.com

Dr. Kazuki Ichikawa
The University of Tokyo
ichikawa@edu.k.u-tokyo.ac.jp

Hae Kyung Im
University of Chicago
haky@uchicago.edu

Dr. Joo Hyun Im
Vertex Pharmaceuticals
joohyun_im@vrtx.com

Dr. Kaoru Inoue
NIEHS
kaoru.inoue@nih.gov

Dr. Richard Isaac
Harvard Medical School
richard_isaac@hms.harvard.edu

Dr. Sadahiro Iwabuchi
Wakayama Medical University
iwabuchi@wakayama-med.ac.jp

Dr. Aishwarya Iyer
Juniper Genomics
aishwarya@junipergenomics.com

Dr. Erich Jarvis
Rockefeller University/HHMI
ejarvis@rockefeller.edu

Mr. Ned Jastromb
Nabsys
jastromb@nabsys.com

Dr. Matthew Jensen
Yale University
matthew.jensen@yale.edu

Dr. Eek-hoon Jho
University of Seoul
ej70@uos.ac.kr

Dr. Xin Jin
BGI Research
jinjin@genomics.cn

Granton Jindal
University of California San Diego
gjindal@ucsd.edu

Ms. Swati Jivanji
Livestock Improvement
swati.jivanji@lic.co.nz

Dr. Masahiro Kanai
Broad Institute of MIT and Harvard
mkanai@broadinstitute.org

Dr. Alireza Karbalayghareh
Memorial Sloan Kettering Cancer Center
karbalayghareh@gmail.com

Dr. Konrad Karczewski
Massachusetts General Hospital
konradk@broadinstitute.org

Dr. Meltem Ece Kars
Icahn School of Medicine at Mount Sinai
meltemece.kars@mssm.edu

Dr. Yukie Kashima
The University of Tokyo
y_kashima@edu.k.u-tokyo.ac.jp

Eugene Katsevich
University of Pennsylvania
ekatsevi@wharton.upenn.edu

Ms. Katarzyna Kedzierska
University of Oxford
kasia@well.ox.ac.uk

Dr. Rebecca Keener
Johns Hopkins University
rkeener@jhmi.edu

Dr. Eimear Kenny
Icahn School of Medicine at Mount Sinai
eimear.kenny@mssm.edu

Prof. Ekta Khurana
Weill Cornell Medicine, New York City
ekk2003@med.cornell.edu

Mr. Gianluca Kikidis
Lieber Institute for Brain Development
gianluca.kikidis@libd.org

Min Cheol Kim
University of California, San Francisco
mincheol.kim@ucsf.edu

Dr. Kwondo Kim
The Jackson Laboratory
kwondo.kim@jax.org

Prof. Tae Kim
The University of Texas at Dallas
genome@utdallas.edu

Dr. Jessica Kissinger
University of Georgia
jkissing@uga.edu

Mr. Jonas Koeppel
Welcome Sanger Institute
jk24@sanger.ac.uk

Dr. Mikhail Kolmogorov
National Institutes of Health
benjamin.douek@nih.gov

Dr. Peter Koo
Cold Spring Harbor Laboratory
koo@cshl.edu

Dr. Eli Kopel
Dexoligo by Dexcel Pharma
eli.kopel@dexoligo.com

Dr. Aleksandra Kornienko
GMI
kornienkoalexandra@gmail.com

Michael Kosicki
Lawrence Berkeley Lab
mkosicki@lbl.gov

Dr. Paul Kotturi
Oxford Nanopore Technologies
paul.kotturi@nanoporetech.com

Dr. Kanako Koyanagi
Hokkaido University
kkoyanag@ist.hokudai.ac.jp

Prof. Ksenia Krasileva
UC Berkeley
kseniak@berkeley.edu

Dr. Judith Kribelbauer
EPFL
Judith.kribelbauer@epfl.ch

Natalie Kucher
Johns Hopkins University
nkucher3@jhu.edu

Lukas Kuderna
Illumina Inc.
lkuderna@illumina.com

Mr. Soumya Kundu
Stanford University
soumyak@stanford.edu

Jason Kunisaki
University of Utah
jhk275@gmail.com

Alexander Kwakye
Stony Brook University
alexander.kwakye@stonybrook.edu

Mr. Kip Lacy
The Rockefeller University
klacy@rockefeller.edu

Dr. Avantika Lal
Genentech
avantikalal02@gmail.com

Dr. Dan Landau
Weill Cornell Medicine
dlandau@nygenome.org

Prof. Tuuli Lappalainen
New York Genome Center & SciLifeLab
tlappalainen@nygenome.org

Dr. Delphine Lariviere
The Pennsylvania State University
lariviere.delphine@gmail.com

Ceejay Lee
Harvard University
ceejaylee@g.harvard.edu

Dr. Soo Ching Lee
NIH/NIAID
sooching.lee@nih.gov

Mr. Sool Lee
University of North Carolina - Chapel Hill
sool_lee@unc.edu

Ms. Young-Lim Lee
University of Liege
younglim.lee.kim@gmail.com

Dr. Kjong Lehmann
University Hospital Cologne
kjlehmann@ukaachen.de

Dr. Ben Lehner
Centre for Genomic Regulation
ben.lehner@crg.es

Dr. Costin Leu
Cleveland Clinic
leuc@ccf.org

Mr. Taibo Li
Johns Hopkins School of Medicine
taiboli@jhu.edu

Dr. Ruijuan Li
inari Agriculture
rli@inari.com

Ms. Ang Li
University of Queensland
ang.li@uq.edu.au

Jiangtao Li
Virginia Tech
jtli@vt.edu

Mr. Jeremy Li
Gencove, Inc.
jeremy.li@gencove.com

Stacy Li
University of California, Berkeley
stacy-l@berkeley.edu

Dr. Siran Li
Cold Spring Harbor Laboratory
siranli@cshl.edu

Dr. Qingnan Liang
UT MD Anderson Cancer Center
qliang3@mdanderson.org

Ms. Shiqi Lin
Beijing Institute of Genomics,CAS
linshiqi2019m@big.ac.cn

Mr. Wenhe Lin
University of Chicago
wenhelin@uchicago.edu

Meng Lin
University of Colorado Anschutz Medical
Campus
meng.lin@cuanschutz.edu

Dr. Xihong Lin
Harvard T.H. Chan School of Public Health
xlin@hsph.harvard.edu

Ms. Eulalie Liorzou
Institut Pasteur
eulalie.liorzou@pasteur.fr

Ms. Lingjie Liu
Cold Spring Harbor Laboratory
liliu@cshl.edu

Andi Liu
UTHealth Houston
andi.liu@uth.tmc.edu

Prof. Boxiang Liu
National University of Singapore
boxiangliu@nus.edu.sg

Dr. George Liu
USDA/NEA
george.liu@usda.gov

Wuzhen Liu
Hong Kong Baptist University
19482000@life.hkbu.edu.hk

Prof. Jianjun Liu
Genome Institute of Singapore
liuj3@gis.a-star.edu.sg

Mr. Wendao Liu
The University of Texas MD Anderson
Cancer Center
wendao.liu@uth.tmc.edu

Dr. Ayelen Lizarraga
University of Chicago
lizarraga@uchicago.edu

Ms. Palmira Llorens Giralt
University of Barcelona
plllorens@ub.edu

Runyang Lou
UC Berkeley
nicolas931010@berkeley.edu

Prof. Craig Lowe
Duke University
craig.lowe@duke.edu

Mr. Martin Loza
The University of Tokyo
mloza@g.ecc.u-tokyo.ac.jp

Shuangjia Lu
Yale University
shuangjia.lu@yale.edu

Prof. Francesca Luca
Wayne State University
fluca@wayne.edu

Laura Luebbert
California Institute of Technology
lauraluebbert@caltech.edu

Suchita Lulla
University of Utah
suchita.lulla@genetics.utah.edu

Anina Lund
Icahn School of Medicine at Mount Sinai
anina.lund@icahn.mssm.edu

Ms. Yanting Luo
Duke University
yanting.luo@duke.edu

Chongyuan Luo
University of California Los Angeles
cluo@mednet.ucla.edu

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Dr. Sorina Maciuca
Genomics England
sorina.maciuca@genomicsengland.co.uk

Kateryna Makova
Penn State University
kdm16@gmail.com

Dr. Joel Malek
Weill Cornell Medicine - Qatar
jom2042@qatar-med.cornell.edu

Mr. Riley Mangan
Duke University
riley.mangan@duke.edu

Dr. Andrew Marderstein
Stanford University
amarder@stanford.edu

Michael Margolis
University of California Los Angeles
mpmargolis@mednet.ucla.edu

Dr. John Marioni
European Bioinformatics Institute/EMBL
marioni@ebi.ac.uk

Prof. Gabor Marth
University of Utah
gabor.marth@gmail.com

Dr. Carlos Marti-Gomez
Cold Spring Harbor Laboratory
martigo@cshl.edu

Dr. Daphne Martschenko
Stanford University
daphnem@stanford.edu

Dr. Diyendo Massilani
Yale School of Medicine
diyendo_massilani@eva.mpg.de

Dr. Iain Mathieson
University of Pennsylvania
mathi@penmedicine.upenn.edu

Jamie Matthews
UCLA
jamieem@g.ucla.edu

Dr. Shakhawan Mawlood
University of Strathclyde
shakhawan.mawlood@gmail.com

Dr. David McCandlish
Cold Spring Harbor Laboratory
mccandlish@cshl.edu

Prof. W. Richard McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Ms. Brianah McCoy
Arizona State University
bmccoy6@asu.edu

Dr. Francis McMahon
NIH/NIMH
mcmahonf@mail.nih.gov

Dr. Marta Mele
Barcelona Supercomputing Center
marta.mele.meseguer@gmail.com

Dr. Stephen Meyn
University of Wisconsin - Madison
stephen.meyn@wisc.edu

Dr. Dan Mishmar
Ben-Gurion University of the Negev
dmishmar@bgu.ac.il

Sneha Mitra
Memorial Sloan Kettering Cancer Center
mitras2@mskcc.org

Mr. Ziyi Mo
Cold Spring Harbor Laboratory
mo@cshl.edu

Dr. Mike Morgan
University of Aberdeen
michael.morgan@abdn.ac.uk

Dr. John Morris
New York Genome Center
jmorris@nygenome.org

Dr. Stephanie Morris
NIH/NHGRI
morriess2@mail.nih.gov

Prof. Ali Mortazavi
University of California Irvine
ali.mortazavi@uci.edu

Dr. Stephen Mosher
Johns Hopkins University
stephen.mosher@jhu.edu

Kristy Mualim
Stanford University
kmualim@stanford.edu

Shandukani Mulaudzi
Harvard Medical School
smulaudzi@g.harvard.edu

Ms. Maddie Murphy
Broad Institute of MIT and Harvard
mamurphy@broadinstitute.org

Dr. Francesc Muyas Remolar
EMBL-EBI
fmuyas@ebi.ac.uk

Anne Nakamoto
UC BERKELEY
annen@berkeley.edu

Dr. RK Narayanan
Cold Spring Harbor Laboratory
narayan@cshl.edu

Mx. Lou Nassar
UCSC Genome Browser
lrnassar@ucsc.edu

Dr. Heini Natri
Translational Genomics Research Institute
hnatri@tgen.org

Dr. Anton Nekrutenko
Penn State University
anton@nekrut.org

Prof. David Nelson
Baylor College of Medicine
nelson@bcm.edu

Mr. Daniel Nesbitt
University of Wisconsin - Madison
djnesbitt@wisc.edu

Mr. Bohan Ni
Johns Hopkins University
bni1@jhu.edu

Natalie Niepeth
Columbia University
natalieniepeth@gmail.com

Festus Nyasimi
The University of Chicago
fnyasimi@uchicago.edu

Dr. Taras Oleksyk
Oakland University
oleksyk@oakland.edu

Dr. Meritxell Oliva
AbbVie
meri.oliva@abbvie.com

Ms. Winona Oliveros Diez
Barcelona Supercomputing Center
winona.oliveros@bsc.es

Dr. Hanna Ollila
University of Helsinki
hanna.m.ollila@helsinki.fi

Mr. Nathaniel Omans
Weill Cornell Medical College
omansn@gmail.com

Dr. Ron Ophir
Agricultural Research Organization
ron@volcani.agri.gov.il

Dr. Sara Oppenheim
American Museum of Natural History
soppenheim@amnh.org

Mr. Omead Ostadan
Cellanome
oostadan@hotmail.com

Dr. Meng Ouyang
Cold Spring Harbor Laboratory
ouyang@cshl.edu

Prof. Svante Paabo
Max Planck Institute for evolutionary
Anthropology
paabo@eva.mpg.de

Dr. Petra Palenikova
Broad Institute
ppalenik@broadinstitute.org

Mr. Ananth Pallaseni
Wellcome Sanger Institute
ap32@sanger.ac.uk

Dr. Arijit Panda
Mayo Clinic
panda.arijit@mayo.edu

Dr. Baoxu Pang
Leiden University Medical Center
b.pang@lumc.nl

Stella Park
New York Genome Center
spark@nygenome.org

Dr. Jiyeon Park
Catholic University
parkji7@gmail.com

Ms. Phoebe Parrish
Fred Hutchinson Cancer Center
pparrish@uw.edu

Mr. Aman Patel
Stanford University
patelas@stanford.edu

Mr. Christopher Pathoulas
UConn Health Center
pathoulas@uchc.edu

Mr. Deshan Perera
University of Calgary
duwagedahampriyabala@ucalgary.ca

Dr. Silvia Perez-Lluch
Centre de Regulacio Genomica
silvia.perez@crg.cat

Dr. Luca Pinello
MGH/Harvard/Broad
lpinello@mgh.harvard.edu

Sebastian Pott
University of Chicago
spott@uchicago.edu

John Prufeta
Medical Excellence Capital, LLC
john.prufeta@medexcelcap.com

Dr. Aaron Quinlan
University of Utah
aquinlan@genetics.utah.edu

Ms. Carolyn Quinlan
University of Toronto
carolyn@junipergenomics.com

Anil Raj
Calico Life Sciences LLC
anil@calicolabs.com

Chandana Rajesh
Cold Spring Harbor Laboratory
rajesh@cshl.edu

Mr. Jose Miguel Ramirez
Barcelona Supercomputing Center
jose.ramirez1@bsc.es

Darius Ramkhalawan
University of Florida
ramkhalawand@ufl.edu

Dr. David Rand
Brown University
david_rand@brown.edu

Ali Ranjbaran
Wayne State University School of Medicine
ali.ranjbaran@med.wayne.edu

Dr. Aline Real
New York Genome Center
areal@nygenome.org

Elisabeth Rebboah
University of California, Irvine
erebboah@uci.edu

Mx. Fairlie Reese
University of California, Irvine
freese@uci.edu

Dr. Steven Reilly
Yale Univeristy
steven.reilly@yale.edu

Jenna Ridge
University Of Missouri-Columbia
jenna.kalleberg@mail.missouri.edu

Ms. Aida Ripoll Cladellas
Barcelona Supercomputing Center
aidaripollcladellas@gmail.com

Kaeli Rizzo
Cold Spring Harbor Laboratory
rizzo@cshl.edu

Elizabeth Roberts
Broad Institute of MIT and Harvard
eroberts@broadinstitute.org

Dr. Joana Rocha
University of California, Berkeley
joana_rocha@berkeley.edu

Juan Rodriguez-Flores
Regeneron Genetics Center
juan.rodriguezflores@regeneron.com

Dr. Mostafa Ronaghi
Cellanome
ronaghi@gmail.com

Federico Rosconi
Boston College
federico.rosconi@bc.edu

Jonathan Rosenski
The Hebrew University of Jerusalem
jonathan.rosenski@mail.huji.ac.il

Marjorie Roskes
Weill Cornell Medicine
mcl4001@med.cornell.edu

Dr. Jeffrey Roskes
Johns Hopkins University
hroskes@jhu.edu

Ms. Fabiana Rossi
Lieber Institute for Brain Development
fabiana.rossi@libd.org

Dr. Daphna Rothschild
Stanford
daphna@stanford.edu

Dr. Maxime Rotival
Institut Pasteur
rotival.maxime@gmail.com

Ms. Laurie Rumker
Harvard University
laurie_rumker@hms.harvard.edu

Ms. Erica Rutherford
Stanford University
emruther@stanford.edu

Dr. Henry Sadowski
Bionano Genomics
hsadowski@bionanogenomics.com

Mr. Mitchell Sanchez Rosado
University of Puerto Rico Medical Sciences
Campus
mitchell.sanchez@upr.edu

Dr. Manbir Sandhu
DNAexus
msandhu@dhanexus.com

Thomas Sasani
University of Utah
thomas.a.sasani@gmail.com

Dr. Carolin Sauer
EMBL-EBI
csauer@ebi.ac.uk

Dr. Michael Schatz
Johns Hopkins University
mschatz@cs.jhu.edu

Dr. Megan Schertzer
New York Genome Center
mschertzer@nygenome.org

Dr. Joshua Schiffman
New York Genome Center
jschiffman@nygenome.org

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Dr. Jacob Schreiber
Stanford University
jmschreiber91@gmail.com

Prof. David Schwartz
University of Wisconsin-Madison
dcschwartz@wisc.edu

Dr. Erich Schwarz
Cornell University
ems394@cornell.edu

Casey Sederman
University of Utah
crs7240@gmail.com

Evan Seitz
Cold Spring Harbor Laboratory
seitz@cshl.edu

Adrian Serohijos
University of Montreal
adrian.serohijos@umontreal.ca

Ms. Isabel Serrano
University of California, Berkeley
isabel_serrano@berkeley.edu

Dr. Aitor Serres Armero
National Institutes of Health
aitor.serresarmero@nih.gov

Dr. Natalie Shaw
NIEHS/NIH
natalie.shaw@nih.gov

Mr. Sudhanshu Shekhar
Purdue University
shekhas@purdue.edu

Ruhollah Shemirani
Icahn School of Medicine at Mount Sinai
ruhollah.shemirani@mssm.edu

Ida Shinder
Johns Hopkins
ishinde1@jhmi.edu

David Siegel
Exai Bio
davids@exai.bio

Ms. Karolina Sienkiewicz
Weill Cornell Medicine
kms4004@med.cornell.edu

Dr. Corinne Simonti
AAAS
csimonti@aaas.org

Dr. Layla Siraj
Harvard
sirajl@broadinstitute.org

Dr. Neil Slaven
Lawrence Berkeley National Lab
nslaven@lbl.gov

Dr. Ashley Smart
Massachusetts Institute of Technology
asmart@mit.edu

Eishani Sokolowski
University of Connecticut Health Center
esokolowski@uchc.edu

Mr. Dongyuan Song
University of California, Los Angeles
dongyuansong@ucla.edu

Wei Song
NCBI
songw2@nih.gov

Dr. Li Song
Geisel School of Medicine at Dartmouth
College
li.song@dartmouth.edu

Susie Song
Broad Institute of MIT and Harvard
ssong@broadinstitute.org

Dr. Pieter Spealman
New York University
ps163@nyu.edu

Dr. Kylee Spencer
American Society of Human Genetics
kspencer@ashg.org

Samvardhini Sridharan
University of California, Berkeley
sridharan@berkeley.edu

Ms. Margaret Starostik
Johns Hopkins University
mstaros1@jhu.edu

Mr. Alexander Starr
Stanford University
astarr97@stanford.edu

Dr. Arnaud Stigliani
Copenhagen University
astigliani31@gmail.com

Petar Stojanoiv
Broad Institute
petar@broadinstitute.org

Caleb Stull
University of Missouri-Columbia
cs744@missouri.edu

Chang Su
Yale University
c.su@yale.edu

Mr. Keith Suderman
Johns Hopkins University
suderman@jhu.edu

Na Sun
Massachusetts Institute of Technology
nasun@mit.edu

Ms. Adithi Sundaresh
University of Helsinki
adithi.sundaresh@helsinki.fi

Prof. Shamil Sunyaev
Brigham & Women's Hospital, Harvard
Medical School
ssunyaev@hms.harvard.edu

Dr. Hillary Sussman
Genome Research, Executive Editor
hsussman@cshl.edu

CHANDLER SUTHERLAND
UC BERKELEY
chandlersutherland@berkeley.edu

Ms. Anima Sutradhar
University of Edinburgh
s2129952@ed.ac.uk

Ms. Lexi Sweet
The Broad Institute
sweet@broadinstitute.org

Ms. Elizabeth Szabo
UConn Health
szabo@uchc.edu

Dr. Lara Szewczak
Cell
lszewczak@cell.com

Dr. Stanley Tahara
USC Keck School of Medicine
stahara@usc.edu

Dr. Michael Talkowski
Massachusetts General Hospital
mtalkowski@mgh.harvard.edu

Mr. Kar-Tong Tan
Broad Institute of MIT and Harvard
ktan@broadinstitute.org

Dr. Ryan Tewhey
Jackson Laboratory
ryan.tewhey@jax.org

James Thomas
NIH
thomasjw4@mail.nih.gov

Dr. Geng Tian
Geneis(Beijing)Co.,Ltd
tiang@geneis.cn

Ms. Yijie Tian
Stony Brook University
yijie.tian@stonybrook.edu

Ms. Shushan Toneyan
Cold Spring Harbor Laboratory
toneyan@cshl.edu

Edmundo Torres-Gonzalez
Penn State University
ejt89@psu.edu

Prof. Richard Trembath
King's College London
richard.trembath@kcl.ac.uk

Dr. Michelle Trenkmann
Springer Nature
michelle.trenkmann@nature.com

Colby Tubbs
Vanderbilt University
colby.a.tubbs@vanderbilt.edu

Christopher Tuggle
Iowa State University
cktuttle@iastate.edu

Mr. Cristian Valencia
Brigham & Women's Hospital
cvalencia4@bwh.harvard.edu

Mr. Peter Van Loo
University of Texas MD Anderson Cancer Center
PVanLoo@mdanderson.org

Dr. Annika Vannan
Translational Genomics Research Institute
avannan@tgen.org

Dr. Lars Velten
Centre for Genomic Regulation (CRG)
lars.velten@crg.eu

Ms. Samvida Venkatesh
University of Oxford
samvida@well.ox.ac.uk

Christa Ventresca
University of Michigan
christav@umich.edu

Mr. Xabier Vergara
Netherlands Cancer Institute
x.vergara.ucin@nki.nl

Prof. Anurag Verma
University of Pennsylvania
anurag.verma@pennmedicine.upenn.edu

Dr. Tauras Vilgalys
University of Chicago
vilgalys@uchicago.edu

Mr. Sergio Villicana Munoz
King's College London
sergio.villicana_munoz@kcl.ac.uk

Dr. Ana Vinuela
Newcastle University
ana.vinuela@newcastle.ac.uk

Dr. Maria Viskadourou
Johns Hopkins University ,SOM
mviskad1@jh.edu

Mr. Martijn Vochteloo
University Medical Center Groningen
m.vochteloo@umcg.nl

Mr. John Walsh
Pacific Biosciences
jwalsh@pacificbiosciences.com

Dr. Ledong Wan
Cold Spring Harbor Laboratory
wan@cshl.edu

Ms. Xinan Wang
Harvard University
xinanwang@hsph.harvard.edu

Shou-Wen Wang
Westlake University
wangshouwen@westlake.edu.cn

Ms. Joyce Wang
The University of Texas at Austin
joyce.wang@utexas.edu

Dr. Gao Wang
Columbia University
wang.gao@columbia.edu

Dr. Jinhua Wang
University of Minnesota
wangjh@umn.edu

Xutao Wang
Boston University
xutaow@bu.edu

Dr. Michelle Ward
University of Texas Medical Branch
miward@utmb.edu

Dr. Alistair Ward
University of Utah
AlistairNWard@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Lab/USDA-ARS
ware@cshl.edu

Dr. Kaja Wasik
Variant Bio
kaja@variantbio.com

Dr. Marina Watowich
Vanderbilt University
marina.watowich@vanderbilt.edu

Eve Wattenberg
Yale University
eve.wattenberg@yale.edu

Dr. Robbee Wedow
Purdue University / Indiana University
School of Medicine
rwedow@purdue.edu

Prof. Michael Weedon
University of Exeter
m.n.weedon@exeter.ac.uk

Xiaoran Wei
Virginia Tech
xrwei@vt.edu

Dr. Josh Weinstock
Johns Hopkins University
jweins17@jhu.edu

Ms. Julianne Weller
Wellcome Sanger Institute
jw38@sanger.ac.uk

Dr. Sarah Wheelan
NIH/NHGRI
sarah.wheelan@nih.gov

Dr. Allyson Whittaker
Arima Genomics
allyson@arimagenomics.com

Dr. Jan Witkowski
Cold Spring Harbor Laboratory
witkowsk@cshl.edu

Mr. Wilfred Wong
Memorial Sloan Kettering Cancer Center
wiw4002@med.cornell.edu

Dr. Andrew Wood
University of Exeter
a.r.wood@exeter.ac.uk

Dr. Zhichao Xu
Salk Institute for Biological Studies
zxu@salk.edu

Mr. Boyan Xu
UC Berkeley
boxu@berkeley.edu

Ms. Xinhe Xue
New York University/New York Genome
center
xx803@nyu.edu

Feriel Yala
University of Sherbrooke
yalf2601@usherbrooke.ca

Rui Yang
MSKCC
ruy4001@med.cornell.edu

Prof. Huanming Yang
BGI-China
yanghm@genomics.cn

Dr. Feyza Yilmaz
The Jackson Laboratory
feyza.yilmaz@jax.org

Mr. Ken Youens-Clark
DNAnexus
kyclark@dnanexus.com

Greg Young
PacBio
gyoung@pacb.com

Dr. Zhi Yu
Broad Institute of MIT and Harvard
zyu@broadinstitute.org

Dr. Dinar Yunusov
Cold Spring Harbor Laboratory
dyunusov@cshl.edu

Dr. Laura Zahn
Cell Press
lzahn@cell.com

Ms. Cynthia Zajac
University of Michigan
herrerab@umich.edu

Dr. Hugo Zeberg
Karolinska Institute
hugo.zeberg@ki.se

Dr. Lu Zeng
Columbia University Irving Medical Center
lz2838@cumc.columbia.edu

Dr. Yixin Zhao
Cold Spring Harbor Laboratory
yizhao@cshl.edu

Yu Zhao
University of Chicago
yuzhao1@uchicago.edu

Mr. Yichao(Charles) Zhou
UChicago
yichaozhou@uchicago.edu

Dr. Xiang Zhou
University of Michigan
xzhousph@umich.edu

Justin Zook
National Institute of Standards and
Technology
justin.zook@nist.gov

Dr. James Zou
Stanford University
jamesyzou@gmail.com

VISITOR INFORMATION

EMERGENCY (to dial outside line, press 3+1+number)

CSHL Security	516-367-8870 (x8870 from house phone)
CSHL Emergency	516-367-5555 (x5555 from house phone)
Local Police / Fire	911
Poison Control	(3) 911

CSHL SightMD Center for Health and Wellness Dolan Hall, East Wing, Room 111 cshlwellness@northwell.edu	516-422-4422 x4422 from house phone
Emergency Room Huntington Hospital 270 Park Avenue, Huntington	631-351-2000
Dentists Dr. William Berg Dr. Robert Zeman	631-271-2310 631-271-8090
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	631-549-9400

GENERAL INFORMATION

Meetings & Courses Main Office

Hours during meetings: M-F 9am – 9pm, Sat 8:30am – 1pm

After hours – See information on front desk counter

For assistance, call Security at 516-367-8870

(x8870 from house phone)

Dining, Bar

Blackford Dining Hall (main level):

Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00

Blackford Bar (lower level): 5:00 p.m. until late

House Phones

Grace Auditorium, upper / lower level; Cabin Complex;
Blackford Hall; Dolan Hall, foyer

Books, Gifts, Snacks, Clothing

CSHL Bookstore and Gift Shop

516-367-8837 (hours posted on door)

Grace Auditorium, lower level.

Computers, E-mail, Internet access

Grace Auditorium

Upper level: E-mail and printing in the business center area

WiFi Access: GUEST (no password)

Announcements, Message Board Mail, ATM, Travel info

Grace Auditorium, lower level

Russell Fitness Center

Dolan Hall, east wing, lower level

PIN#: (On your registration envelope)**Laundry Machines**

Dolan Hall, lower level

Photocopiers, Journals, Periodicals, Books*CSHL Main Library*

Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri

10:00 a.m. – 6:00 p.m. Saturday

Use PIN# (On your registration envelope) to enter Library
after hours.

See Library staff for photocopier code.

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m.

Two tennis courts open daily.

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 CSHL Shuttle OR Uber/Lyft/Taxi to Syosset Train Station

Long Island Railroad to Penn Station

Train ride about one hour.

TRANSPORTATION**Limo, Taxi**

Syosset Limousine	516-364-9681
Executive Limo Service	516-826-8172
Limos Long Island	516-400-3364
Syosset Taxi	516-921-2141
Orange & White Taxi	631-271-3600
Uber / Lyft	

Trains

Long Island Rail Road	718-217-LIRR (5477)
Amtrak	800-872-7245
MetroNorth	877-690-5114
New Jersey Transit	973-275-5555

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.

CODE OF CONDUCT FOR ALL PARTICIPANTS IN CSHL MEETINGS

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. To this end, the Laboratory aims to maintain a safe and respectful environment that is free from harassment and discrimination for all attendees of our meetings and courses as well as associated support staff, in accordance with federal, state and local laws.

By registering for and attending a CSHL meeting, either in person or virtually, participants agree to:

1. Treat fellow meeting participants and CSHL staff with respect, civility and fairness, without bias based on sex, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, nationality or national origin, citizenship status, disability status, veteran status, marital or partnership status, age, 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 if attending in person, as you would at your home institution.
3. Abide by the CSHL Meeting Alcohol Policy if attending in person.

Similarly, meeting participants agree to refrain from:

1. Harassment and discrimination, either in person or online, in violation of Laboratory policy based on sex, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, nationality or national origin, citizenship status, disability status, veteran status, marital or partnership status, age, genetic information, or any other criteria prohibited under applicable federal, state or local law.
2. Sexual harassment or misconduct.
3. Disrespectful, uncivil and/or unprofessional interpersonal behavior, either in person or online, that interferes with the working and learning environment.
4. Misappropriation of Laboratory property or excessive personal use of resources, if attending in person.

DEFINITIONS AND EXAMPLES

Uncivil/disrespectful behavior is not limited to but may take the following forms:

- Shouting, personal attacks or insults, throwing objects, and/or sustained disruption of talks or other meeting-related events

Harassment/discrimination is not limited to but may take the following forms:

- Threatening, stalking, bullying, demeaning, coercive, or hostile acts that may have real or implied threats of physical, professional, or financial harm
- Signs, graphics, photographs, videos, gestures, jokes, pranks, epithets, slurs, or stereotypes that comment on a person's sex, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, nationality or national origin, citizenship status, disability status, veteran status, marital or partnership status, age, genetic information, or physical appearance

Sexual misconduct is not limited to but may take the following forms:

- Unwelcome and uninvited attention, physical contact, or inappropriate touching
- Groping or sexual assault
- Use of sexual imagery, objects, gestures, or jokes in public spaces or presentations
- Any other verbal or physical contact of a sexual nature when such conduct creates a hostile environment, prevents an individual from fulfilling their professional responsibilities at the meeting, or is made a condition of employment or compensation either implicitly or explicitly

REPORTING BREACHES OR VIOLATIONS

Cold Spring Harbor Laboratory aims to maintain in-person and virtual conference environments that accord with the principles and expectations outlined in this 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

[Reports may be submitted](#) by those who experience harassment or discrimination as well as by those who witness violations of the behavior laid out in this Code.



The Laboratory will act 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.

Since many CSHL meetings and courses are funded by NIH grants, you may also contact the [Health & Human Services Office for Civil Rights \(OCR\)](#).

See [this page](#) for information on filing a civil rights complaint with the OCR; filing a complaint with CSHL is not required before filing a complaint with OCR, and seeking assistance from CSHL in no way prohibits filing complaints with OCR. You [may also notify NIH directly](#) about sexual harassment, discrimination, and other forms of inappropriate conduct at NIH-supported events.

CSHL Campus Map



